Lecture Notes in Computer Science 3546

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
New York University, NY, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Takeo Kanade Anil Jain
Nalini K. Ratha (Eds.)

Audio- and Video-Based Biometric Person Authentication

5th International Conference, AVBPA 2005
Hilton Rye Town, NY, USA, July 20-22, 2005
Proceedings

Springer

Volume Editors

Takeo Kanade
Carnegie Mellon University
Robotics Institute
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
E-mail: kanade@cs.cmu.edu

Anil Jain
Michigan State University
Dept. of Computer Science and Engineering
3115 Engineering Building, East Lansing, MI 48824-1226, USA
E-mail: jain@cse.msu.edu

Nalini K. Ratha
IBM Thomas J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10598, USA
E-mail: ratha@us.ibm.com

Library of Congress Control Number: 2005928845

CR Subject Classification (1998): I.5, I.4, I.3, K.6.5, K.4.4, C.2.0

ISSN 0302-9743
ISBN-10 3-540-27887-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27887-0 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2005
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11527923 06/3142 5 4 3 2 1 0

Preface

We are delighted to present to you this volume containing the research papers presented at the 5th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA 2005), Hilton Rye Town, NY, July 20–22, 2005. As society becomes more security conscious, so the interest has heightened in biometrics. The large number of papers submitted to AVBPA 2005 reflects the intensity of the research in this important growing field.

Papers in AVBPA 2005 discuss all aspects of biometrics, including iris, fingerprint, face, palm print, gait, speaker and signature recognition. Iris recognition has drawn considerable interest from researchers worldwide. Also observed is a significant increase in papers on multi-modal and fusion techniques as well as security enhancements to biometrics. The strong industrial flavor of AVBPA 2005 is evident in several presentations on smart cards, wireless devices, architectures and implementation factors – the consequence of the recent development and deployment of commercial biometric systems.

Thanks to the dedicated efforts of previous organizers and participants, the AVBPA conference series has positioned itself as the premier biometrics research conference. The hard work of the AVBPA 2005 Program Committee – one of the best, with leaders in all the biometrics areas – and reviewers made it possible to continue that tradition. With their high-quality and timely feedback, we were able to put together an excellent technical program. In addition, we are grateful to have had many industrial sponsorships – a key factor in the success of the conference.

We would also like to thank the International Association for Pattern Recognition (IAPR) for its continued support of AVBPA. Previous AVBPA conferences were held in Crans-Montana, Switzerland (1997), Washington, DC, USA (1999), Halmstad, Sweden (2001), and Surrey, UK (2003).

AVBPA continues to offer a valuable snapshot of research in biometrics from leading institutions around the world. We hope that the papers contained in this volume and this conference will inspire new research in this important area.

May 2005

Takeo Kanade,
Anil Jain,
Nalini Ratha

Executive Committee

General Chair
Dr. Takeo Kanade, CMU
Program Co-chair
Dr. Anil Jain, MSU
Program Co-chair
Dr. Nalini K. Ratha, IBM Research

Program Committee

Shigeru Akamatsu, Japan
Vijayakumar Bhagavatula, USA
Bir Bhanu, USA
Josef Bigun, Sweden
Horst Bunke, Switzerland
Joe Campbell, USA
Rama Chellappa, USA
John Daugman, UK
Bernadette Dorizzi, France
Pat Flynn, USA
Kazuhiro Fukui, Japan
Sadaoki Furui, Japan
Paul Griffin, USA
Jaihie Kim, Korea
Josef Kittler, UK
Shihong Lao, Japan
Seing-whan Lee, Korea
Stan Li, China
Davide Maltoni, Italy
Mark Nixon, UK
Larry O'Gorman, USA
Javier Ortega, Spain
Jonathon Philips, USA
Arun Ross, USA
Sudeep Sarkar, USA
Xiaoou Tang, Hong Kong, China
Tieniu Tian, China
Kar-Ann Toh, Singapore
Pim Tuyls, Netherlands
Thomas Vetter, Germany
Jim Wayman, USA
David Zhang, Hong Kong, China

Local Arrangements and Finance Chair
Atul Chhabra, USA

Publicity Chairs
Yunhong Wang, China
Davide Maltoni, Italy
Arun Ross, USA

Industry Liaison Chair
Jim Wayman, SJSU

Web Support Coordinator
Umut Uludag, MSU

Additional Reviewers
Haizhou Ai, Tsinghua University, China
Mike Beattie, CMU, USA
Boris Blank, Identix, USA
Ruud Bolle, IBM Research, USA
Theodore Camus, Sarnoff Corporation, USA
Upendra Chaudhari, IBM Research, USA
Yi Chen, MSU, USA
Hong Chen, MSU, USA
Xu Chenghua, NLPR, CAS, China
Wei Fan, NLPR, China
Craig Fancourt, Sarnoff Corporation, USA
Yanlin Guo, Sarnoff Corporation, USA
Andre Gustavo Adami, OGI, USA
Pablo Hennings, CMU, USA
Vincent Hsu, Identix, USA
Jens Hube, Identix, USA
Bon-Woo Hwang, CMU, USA
Cui Jiali, NLPR, CAS, China
Li Jiangwei, NLPR, China
Steve Krawczyk, MSU, USA
Q.S. Liu, City University of Hong Kong, Hong Kong, China
Marcus Liwicki, Switzerland
Yong Ma, Omron, Japan
Udo Mahlmeister, Identix, USA
Yuasa Mayumi, Toshiba, Japan
Ernst Mucke, Identix, USA
Karthik Nandakumar, MSU, USA
Jiri Navratil, IBM Research, USA
Michel Neuhaus, Switzerland
Unsang Park, MSU, USA
Jeong-Seon Park, Korea University, Korea

Sharath Pankanti, IBM Research, USA
Marios Savvides, CMU, USA
Andreas Schlapbach, Switzerland
Conrad Sanderson, Australian National University, Australia
Shiguang Shan, CAS, China
Zhenan Sun, CAS, China
Umut Uludag, MSU, USA
Jason Thornton, CMU, USA
Krithika Venkataramani, CMU, USA
Chunyan Xie, CMU, USA
Fan Wei, NLPR, CAS, China
S.C. Yan, City University of Hong Kong, Hong Kong, China
Miki Yamada, Toshiba, Japan
O. Yamaguchi, Toshiba, Japan
Zhanfeng Yue, University of Maryland, USA
Tao Zhao, Sarnoff Corporation, USA
Junping Zhang, Fudan University, China
Yongfang Zhu, MSU, USA
Shaohua Zhou, Siemens Corporate Research, USA

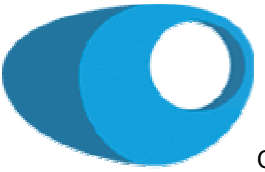
Sponsors (in Alphabetical Order)











Omniperception











Table of Contents

Iris

Iris Recognition at a Distance 1
Craig Fancourt, Luca Bogoni, Keith Hanna, Yanlin Guo, Richard Wildes, Naomi Takahashi, and Uday Jain

Iris Recognition Using Fourier-Wavelet Features 14
Ping S. Huang, Chung-Shi Chiang, and Ji-Ren Liang

Specific Texture Analysis for Iris Recognition 23
Emine Krichen, Lorène Allano, Sonia Garcia-Salicetti, and Bernadette Dorizzi

A Study on Iris Image Restoration 31
Byung Jun Kang and Kang Ryoung Park

Face I (Short)

Eye Perturbation Approach for Robust Recognition of Inaccurately Aligned Faces 41
Jaesik Min, Kevin W. Bowyer, and Patrick J. Flynn

On Combining Textural and Geometrical Scores for Discriminative Face Authentication 51
José Luis Alba-Castro and Daniel González-Jiménez

A One Bit Facial Asymmetry Code (FAC) in Fourier Domain for Human Recognition 61
Sinjini Mitra, Marios Savvides, and B.V.K. Vijaya Kumar

Face Recognition with the Multiple Constrained Mutual Subspace Method 71
Masashi Nishiyama, Osamu Yamaguchi, and Kazuhiro Fukui

A Flexible Object Model for Recognising and Synthesising Facial Expressions . . . 81
Andreas Tewes, Rolf P. Würtz, and Christoph von der Malsburg

Face Reconstruction Across Different Poses and Arbitrary Illumination Conditions 91
Sen Wang, Lei Zhang, and Dimitris Samaras

Stepwise Reconstruction of High-Resolution Facial Image Based on Interpolated Morphable Face Model 102
Jeong-Seon Park and Seong-Whan Lee

Illumination Invariant Face Recognition
Using Linear Combination of Face Exemplars 112
Song-Hyang Moon, Sang-Woong Lee, and Seong-Whan Lee

Video-Based Face Recognition Using Bayesian Inference Model 122
Wei Fan, Yunhong Wang, and Tieniu Tan

Finger-I (Short)

A Hybrid Swipe Fingerprint Mosaicing Scheme 131
Yong-liang Zhang, Jie Yang, and Hong-tao Wu

A Study on Multi-unit Fingerprint Verification 141
Kangrok Lee, Kang Ryoung Park, Jain Jang, Sanghoon Lee, and Jaihie Kim

A Fingerprint Authentication System Based on Mobile Phone 151
Qi Su, Jie Tian, Xinjian Chen, and Xin Yang

Fingerprint Quality Indices for Predicting Authentication Performance 160
Yi Chen, Sarat C. Dass, and Anil K. Jain

Registration of Fingerprints by Complex Filtering
and by 1D Projections of Orientation Images 171
Kenneth Nilsson and Josef Bigun

A Feature Map Consisting of Orientation and Inter-ridge Spacing
for Fingerprint Retrieval 184
Sung-Oh Lee, Yong-Guk Kim, and Gwi-Tae Park

A Graph Matching Based Approach to Fingerprint Classification
Using Directional Variance 191
Michel Neuhaus and Horst Bunke

Fingerprint Singular Points Detection and Direction Estimation
with a "T" Shape Model 201
Tong Liu, Pengwei Hao, and Chao Zhang

Face-I

Face Detection Based on the Manifold 208
Ruiping Wang, Jie Chen, Shengye Yan, and Wen Gao

Local and Global Feature Extraction for Face Recognition 219
Yongjin Lee, Kyunghee Lee, and Sungbum Pan

Video-Based Face Recognition Using Earth Mover's Distance 229
Jiangwei Li, Yunhong Wang, and Tieniu Tan

Face Recognition Based on Recursive Bayesian Fusion of Multiple Signals and Results from Expert Classifier Sets 239
Michael Hild and Ryo Kuzui

Fingerprint

Principal Deformations of Fingerprints 250
Sergey Novikov and Oleg Ushmaev

Fingerprint Mosaicking by Rolling and Sliding 260
Kyoungtaek Choi, Hee-seung Choi, and Jaihie Kim

Minutiae Matching Based Fingerprint Verification Using Delaunay Triangulation and Aligned-Edge-Guided Triangle Matching 270
Huimin Deng and Qiang Huo

Security and Smartcard

An Asymmetric Fingerprint Matching Algorithm for Java Card™ 279
Stefano Bistarelli, Francesco Santini, and Anna Vaccarelli

Scenario Based Performance Optimisation in Face Verification Using Smart Cards 289
Thirimachos Bourlai, Kieron Messer, and Josef Kittler

Characterization, Similarity Score and Uniqueness Associated with Perspiration Pattern 301
Aditya Abhyankar and Stephanie Schuckers

Fuzzy Vault for Fingerprints 310
Umut Uludag, Sharath Pankanti, and Anil K. Jain

Short Oral

Infrared Face Recognition by Using Blood Perfusion Data 320
Shi-Qian Wu, Wei Song, Li-Jun Jiang, Shou-Lie Xie, Feng Pan, Wei-Yun Yau, and Surendra Ranganath

On Finding Differences Between Faces 329
Manuele Bicego, Enrico Grosso, and Massimo Tistarelli

Face Detection Using Look-Up Table Based Gentle AdaBoost 339
Cem Demirkır and Bülent Sankur

Post-processing on LDA's Discriminant Vectors for Facial Feature Extraction 346
Kuanquan Wang, Wangmeng Zuo, and David Zhang

An Integrated Prediction Model for Biometrics 355
Rong Wang, Bir Bhanu, and Hui Chen

Active Shape Models with Invariant Optimal Features (IOF-ASMs) 365
Federico Sukno, Sebastián Ordás, Costantine Butakoff, Santiago Cruz, and Alejandro Frangi

Palmprint Authentication Using Time Series 376
Jian-Sheng Chen, Yiu-Sang Moon, and Hoi-Wo Yeung

Ear Biometrics by Force Field Convergence 386
David J. Hurley, Mark S. Nixon, and John N. Carter

Short Oral-4

Towards Scalable View-Invariant Gait Recognition: Multilinear Analysis for Gait 395
Chan-Su Lee and Ahmed Elgammal

Combining Verification Decisions in a Multi-vendor Environment 406
Michael Beattie, B.V.K. Vijaya Kumar, Simon Lucey, and Ozan K. Tonguz

Gait Recognition by Combining Classifiers Based on Environmental Contexts . . . 416
Ju Han and Bir Bhanu

Addressing the Vulnerabilities of Likelihood-Ratio-Based Face Verification 426
Krzysztof Kryszczuk and Andrzej Drygajlo

Practical Biometric Authentication with Template Protection 436
Pim Tuyls, Anton H.M. Akkermans, Tom A.M. Kevenaar, Geert-Jan Schrijen, Asker M. Bazen, and Raymond N.J. Veldhuis

A Study of Brute-Force Break-ins of a Palmprint Verification System 447
Adams Kong, David Zhang, and Mohamed Kamel

Modification of Intersession Variability in On-Line Signature Verifier 455
Yasunori Hongo, Daigo Muramatsu, and Takashi Matsumoto

MOC via TOC Using a Mobile Agent Framework 464
Stefano Bistarelli, Stefano Frassi, and Anna Vaccarelli

Fusion

Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks 474
Norman Poh and Samy Bengio

A Classification Approach to Multi-biometric Score Fusion 484
Yan Ma, Bojan Cukic, and Harshinder Singh

A Generic Protocol for Multibiometric Systems Evaluation
on Virtual and Real Subjects 494
Sonia Garcia-Salicetti, Mohamed Anouar Mellakh, Lorène Allano, and Bernadette Dorizzi

Multi-modal

Multi-biometrics 2D and 3D Ear Recognition 503
Ping Yan and Kevin W. Bowyer

Biometric Authentication System
Using Reduced Joint Feature Vector of Iris and Face 513
Byungjun Son and Yillbyung Lee

An On-Line Signature Verification System
Based on Fusion of Local and Global Information 523
Julian Fierrez-Aguilar, Loris Nanni, Jaime Lopez-Peñalba, Javier Ortega-Garcia, and Davide Maltoni

Human Recognition at a Distance in Video
by Integrating Face Profile and Gait 533
Xiaoli Zhou, Bir Bhanu, and Ju Han

4:50 – 5:50 Oral-7 (Palm and Finger Surface)

Identity Verification Utilizing Finger Surface Features 544
Damon L. Woodard and Patrick J. Flynn

Palmprint Authentication Based on Orientation Code Matching 555
Xiangqian Wu, Kuanquan Wang, and David Zhang

A Novel Palm-Line Detector 563
Li Liu and David Zhang

Speaker and Gait

A New On-Line Model Quality Evaluation Method for Speaker Verification 572
Javier R. Saeta and Javier Hernando

Improving Speaker Verification Using ALISP-Based Specific GMMs 580
Asmaa El Hannani and Dijana Petrovska-Delacrétaz

Multimodal Speaker Verification Using Ear Image Features
Extracted by PCA and ICA 588
Koji Iwano, Taro Miyazaki, and Sadaoki Furui

Modelling the Time-Variant Covariates for Gait Recognition 597
Galina V. Veres, Mark S. Nixon, and John N. Carter

Face II

Robust Face Recognition Using Advanced Correlation Filters
with Bijective-Mapping Preprocessing 607
Marios Savvides, Chunyan Xie, Nancy Chu, B.V.K. Vijaya Kumar, Christine Podilchuk, Ankur Patel, Ashwath Harthattu, and Richard Mammone

Photometric Normalisation for Face Verification 617
James Short, Josef Kittler, and Kieron Messer

Experimental Evaluation of Eye Location Accuracies and Time-Lapse Effects
on Face Recognition Systems 627
Haoshu Wang and Patrick J. Flynn

Experiments in Mental Face Retrieval 637
Yuchun Fang and Donald Geman

Poster I

Fingerprint Image Segmentation Based on Quadric Surface Model 647
Yilong Yin, Yanrong Wang, and Xiukun Yang

A Fingerprint Matching Algorithm Based on Radial Structure
and a Structure-Rewarding Scoring Strategy 656
Kyung Deok Yu, Sangsin Na, and Tae Young Choi

A Novel Algorithm for Distorted Fingerprint Matching
Based on Fuzzy Features Match 665
Xinjian Chen, Jie Tian, and Xin Yang

Minutiae Quality Scoring and Filtering
Using a Neighboring Ridge Structural Analysis on a Thinned Fingerprint Image . . 674
Dong-Hun Kim

Hardware-Software Codesign of a Fingerprint Identification Algorithm 683
Nicolau Canyellas, Enrique Cantó, G. Forte, and M. López

Fingerprint Matching Using the Distribution
of the Pairwise Distances Between Minutiae 693
Chul-Hyun Park, Mark J.T. Smith, Mireille Boutin, and Joon-Jae Lee

A Layered Fingerprint Recognition Method 702
Woong-Sik Kim and Weon-Hee Yoo

Super-template Generation Using Successive Bayesian Estimation for Fingerprint Enrollment 710
Choonwoo Ryu, Youngchan Han, and Hakil Kim

Secure Fingerprint Matching with External Registration 720
James Reisman, Umut Uludag, and Arun Ross

Palmprint Recognition Using Fourier-Mellin Transformation Based Registration Method 730
Liang Li, Xin Yang, Yuliang Hi, and Jie Tian

Parametric Versus Non-parametric Models of Driving Behavior Signals for Driver Identification 739
Toshihiro Wakita, Koji Ozawa, Chiyomi Miyajima, and Kazuya Takeda

Performance Evaluation and Prediction for 3D Ear Recognition 748
Hui Chen, Bir Bhanu, and Rong Wang

Optimal User Weighting Fusion in DWT Domain On-Line Signature Verification 758
Isao Nakanishi, Hiroyuki Sakamoto, Yoshio Itoh, and Yutaka Fukui

Gait Recognition Using Spectral Features of Foot Motion 767
Agus Santoso Lie, Ryo Shimomoto, Shohei Sakaguchi, Toshiyuki Ishimura, Shuichi Enokida, Tomohito Wada, and Toshiaki Ejima

VALID: A New Practical Audio-Visual Database, and Comparative Results 777
Niall A. Fox, Brian A. O'Mullane, and Richard B. Reilly

Audio-Visual Speaker Identification via Adaptive Fusion Using Reliability Estimates of Both Modalities 787
Niall A. Fox, Brian A. O'Mullane, and Richard B. Reilly

Speaker Identification Using the VQ-Based Discriminative Kernels 797
Zhenchun Lei, Yingchun Yang, and Zhaohui Wu

Exploiting Glottal Information in Speaker Recognition Using Parallel GMMs . . . 804
Pu Yang, Yingchun Yang, and Zhaohui Wu

Biometric Recognition Using Feature Selection and Combination 813
Ajay Kumar and David Zhang

Evaluation of Biometric Identification in Open Systems 823
Michael Gibbons, Sungsoo Yoon, Sung-Hyuk Cha, and Charles Tappert

Exploring Similarity Measures for Biometric Databases 832
Praveer Mansukhani and Venu Govindaraju

Indexing Biometric Databases Using Pyramid Technique 841
Amit Mhatre, Sharat Chikkerur, and Venu Govindaraju

Classification Enhancement via Biometric Pattern Perturbation 850
Terry Riopka and Terrance Boult

Calculation of a Composite DET Curve 860
Andy Adler and Michael E. Schuckers

Privacy Operating Characteristic
for Privacy Protection in Surveillance Applications 869
P. Jonathon Phillips

Poster II

Headprint – Person Reacquisition
Using Visual Features of Hair in Overhead Surveillance Video 879
Hrishikesh Aradhye, Martin Fischler, Robert Bolles, and Gregory Myers

A Survey of 3D Face Recognition Methods 891
Alize Scheenstra, Arnout Ruifrok, and Remco C. Veltkamp

Influences of Image Disturbances on 2D Face Recognition 900
Henning Daum

Local Feature Based 3D Face Recognition 909
Yonguk Lee, Hwanjong Song, Ukil Yang, Hyungchul Shin,
and Kwanghoon Sohn

Fusion of Appearance and Depth Information for Face Recognition 919
Jian-Gang Wang, Kar-Ann Toh, and Ronda Venkateswarlu

Gabor Feature Based Classification Using 2D Linear Discriminant Analysis
for Face Recognition 929
Ming Li, Baozong Yuan, and Xiaofang Tang

Multi-resolution Histograms of Local Variation Patterns (MHLVP)
for Robust Face Recognition 937
Wenchao Zhang, Shiguang Shan, Hongming Zhang, Wen Gao, and Xilin Chen

Analysis of Response Performance Characteristics for Identification
Using a Matching Score Generation Model 945
Takuji Maeda, Masahito Matsushita, Koichi Sasakawa, and Yasushi Yagi

Pose Invariant Face Recognition Under Arbitrary Illumination
Based on 3D Face Reconstruction 956
Xiujuan Chai, Laiyun Qing, Shiguang Shan, Xilin Chen, and Wen Gao

Discriminant Analysis Based on Kernelized Decision Boundary for Face Recognition 966
Baochang Zhang, Xilin Chen, and Wen Gao

A Probabilistic Approach to Semantic Face Retrieval System 977
Karthik Sridharan, Sankalp Nayak, Sharat Chikkerur, and Venu Govindaraju

Authenticating Corrupted Facial Images on Stand-Alone DSP System 987
Sang-Woong Lee, Ho-Choul Jung, and Seong-Whan Lee

Evaluation of 3D Face Recognition Using Registration and PCA 997
Theodoros Papatheodorou and Daniel Rueckert

Dynamic Approach for Face Recognition Using Digital Image Skin Correlation . 1010
Satprem Pamudurthy, E Guan, Klaus Mueller, and Miriam Rafailovich

Rank-Based Decision Fusion for 3D Shape-Based Face Recognition 1019
Berk Gökberk, Albert Ali Salah, and Lale Akarun

Robust Active Shape Model Construction and Fitting for Facial Feature Localization 1029
Zhenghui Gui and Chao Zhang

Comparative Assessment of Content-Based Face Image Retrieval in Different Color Spaces 1039
Peichung Shih and Chengjun Liu

A Principled Approach to Score Level Fusion in Multimodal Biometric Systems 1049
Sarat C. Dass, Karthik Nandakumar, and Anil K. Jain

A Score-Level Fusion Benchmark Database for Biometric Authentication 1059
Norman Poh and Samy Bengio

Fusion for Multimodal Biometric Identification 1071
Yongjin Lee, Kyunghee Lee, Hyungkeun Jee, Younhee Gil, Wooyong Choi, Dosung Ahn, and Sungbum Pan

Between-Source Modelling for Likelihood Ratio Computation in Forensic Biometric Recognition 1080
Daniel Ramos-Castro, Joaquin Gonzalez-Rodriguez, Christophe Champod, Julian Fierrez-Aguilar, and Javier Ortega-Garcia

The Effectiveness of Generative Attacks on an Online Handwriting Biometric . . 1090
Daniel P. Lopresti and Jarret D. Raim

Vulnerabilities in Biometric Encryption Systems 1100
Andy Adler

Securing Electronic Medical Records Using Biometric Authentication 1110
Stephen Krawczyk and Anil K. Jain

A Novel Approach to Combining Client-Dependent and Confidence Information in Multimodal Biometrics 1120
Norman Poh and Samy Bengio

Author Index 1131

Iris Recognition at a Distance

Craig Fancourt, Luca Bogoni*, Keith Hanna, Yanlin Guo,
Richard Wildes**, Naomi Takahashi, and Uday Jain

Sarnoff Corp., 201 Washington Road, Princeton, NJ 08540, USA
cfancourt@sarnoff.com

Abstract. We describe experiments demonstrating the feasibility of human iris recognition at up to 10 m distance between subject and camera. The iris images of 250 subjects were captured with a telescope and infrared camera, while varying distance, capture angle, environmental lighting, and eyewear. Automatic iris localization and registration algorithms, in conjunction with a local correlation based matcher, were used to obtain a similarity score between gallery and probe images. Both the area under the receiver operating characteristic (ROC) curve and the Fisher Linear Discriminant were used to measure the distance between authentic and imposter distributions. Among variables studied, database wide experiments reveal no performance degradation with distance, and minor performance degradation with, in order of increasing effect, time (one month), capture angle, and eyewear.

Introduction

Both theory and practice suggest that human iris patterns may serve as a sufficient basis for constructing very accurate biometric systems [1,2]. Experimental data to date suggest a remarkable level of performance; no false matches and 2% false non-matches in over 2 million cross comparisons of 200 irises in an independent evaluation conducted according to accepted best practices [3].

However, the performance of any real system must also take into account the distribution of the distance metric between the same iris (authentic) captured at different times, and in possibly different conditions. For a given comparison algorithm, while the distribution of imposter scores is somewhat invariant to independent noise caused by changes in the acquisition system or environment, the distribution of authentic scores is highly dependent on the details of the acquisition. Daugman [4] showed that variations in camera and subject-camera distance greatly reduced the separation between authentic and imposter distributions, to the point of increasing the probability of a false match to 1 in 4 million. However, those studies still involved subjects acquired at relatively small distances. As the distance from subject to camera increases, and in more realistic scenarios, the expected variation in the resulting iris images also increases, resulting in potentially large differences in resolution, illumination, and eye pose, to name a few.

Here, we present a first study of iris recognition at a distance, and the effect of controlled variations in capture conditions expected in a real world scenario where the subject is far from the acquisition system. To this end, we collected the iris images of

* Now at Siemens Corp

** Now at Dept. of Computer Science & Centre for Vision Research, York Univ

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1–13, 2005.
© Springer-Verlag Berlin Heidelberg 2005

250 subjects using a telescope and infrared camera, at distances of 5 m and 10 m, while varying environmental illumination, capture angle (eye pose), time between acquisitions, eye motion, and eyewear. However, in this paper we do not present results involving illumination or eye motion.

Table 1. Independent variables and their values for this study

Variable	Values
Distance	5 m, 10 m
Illumination	Low, diffuse, overhead
Angle	0°, 30°
Time	Month 1, month 2
Eye motion	Fixating, tracking
Eyewear	None, glasses, contacts

We now describe the various components of the system, including the data acquisistion, iris localization, registration, and matching algorithms. We then present a set of experiments designed to isolate the effect of the various experimental conditions on iris recognition accuracy.

Data Collection

A schematic diagram of the data collection staging area is shown in Figure 1, while photos of key equipment are shown in Figure 2. During data collection the subject was seated and the subject's head was supported by a chin rest with a forehead stabilizer. To collect fixating data, the subject was instructed to focus at one of two spots on a distant wall, located at 0° and 30°, respectively, where 0° was in-line with the camera. To collect tracking data, the subject followed a slowly moving LED light projected onto a black surface directly in front of the subject.

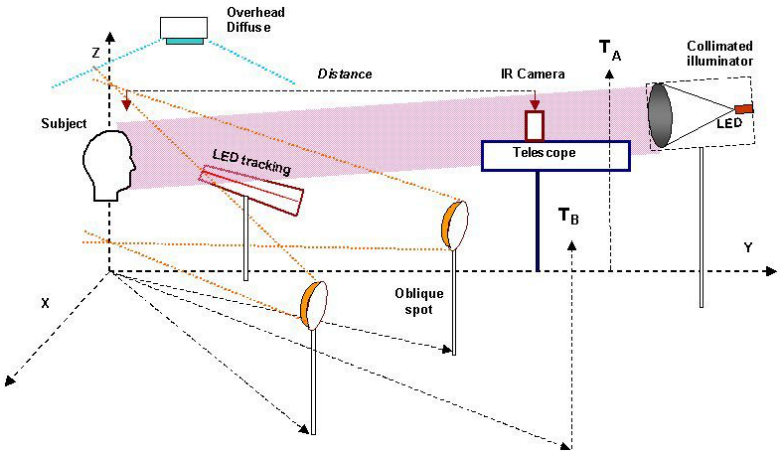


Fig. 1. Schematic diagram of the data collection staging area. Ta and Tb represent distant spots for subject focus during 0 and 30 degree acquisition, respectively

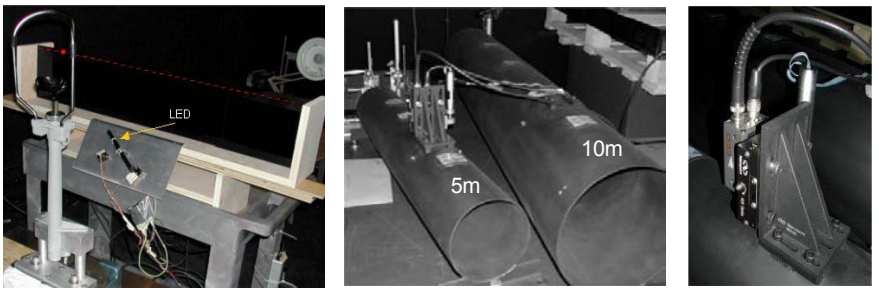


Fig. 2. Data acquisition equipment: (left) chin rest with forehead stabilizer and moving LED light projector for tracking experiments; (center) elliptical mirror telescopes for 5 m and 10 m acquisition; (right) closeup of infrared camera

The subject was illuminated from a distance by a 880 nm collimated infrared source, resulting in a measured irradiance at the eye of ~0.4 mW/cm^2. Overhead and oblique visible light sources were used to vary environmental illumination. An infrared image of a region centered around the eye was captured using one of two custom-built elliptical mirror telescopes connected to an infrared camera (Hitachi KP-M2RN). The telescope designed for 5 m acquisition had a diameter of 10.16 cm, while the one for 10 m acquisition had a diameter of 20.32 cm. Both were designed to produce an image with 128 pixel resolution across a typical iris diameter, out of a total resolution of 640x480 pixels. This is comfortably larger than the 70 pixels estimated as the minimum required for recognition [4]. The measured modulation transfer function is shown in Figure 3, which is very close to the theoretical limit.

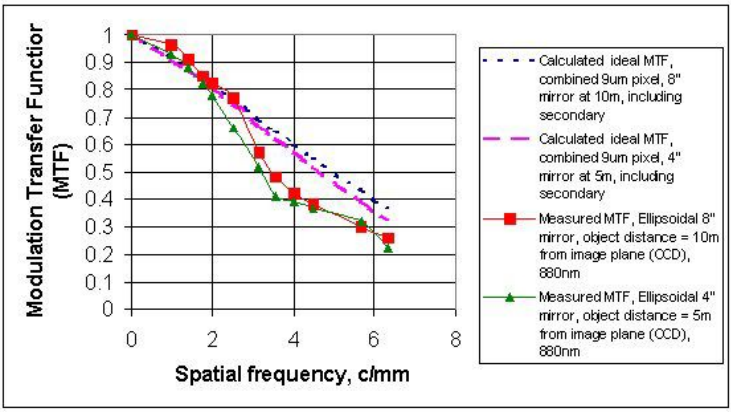


Fig. 3. Measured vs. theoretical modulation transfer function for the 5 m and 10 m acquisition system

Two separate databases were collected. Database I involved 50 subjects and 50 irises, resulting in 1996 video sequences. The evaluation of Database I provided a basis for the sub-sampling criteria used for the Database II acquisition. Database II involved 200 subjects and 247 irises, resulting in 4720 video sequences. There were no common subjects between the two databases. Within each database, each subject

participated in two sessions separated in time by at least one month. Each experiment was captured at 12 frames per second, for 10 seconds, resulting in 120 frames per video sequence.

For experiments involving fixating subjects, two images were chosen from each video sequence: one from frames 1-60, and a second from frames 61-120. They were chosen on the basis of the minimum energy change between neighboring frames, in order to eliminate rapid changes that could affect iris visibility or focus, such as eye blinks and movements, and pupil dilation or contraction. For experiments where the gallery and probe refer to the same experiment, the gallery refers to the chosen image from frames 1-60, while the probe refers to the chosen image from frames 61-120. For experiments where the gallery and probe refer to different experiments, only the chosen image from frames 1-60 was used.

Iris Localization

Two algorithms were developed for automatic iris localization, partly based on work by Camus and Wildes [5]. Both algorithms are based on a multiple hypothesis approach that selects candidate seeds for the iris center loci and then verifies the presence of the iris. The first algorithm uses a seed that is region-based while the second uses edges. Both demonstrated good results in the localization of the iris. The region-based approach was more sensitive to the presence of multiple highlights from eyewear and was limited in performance when the view was not normal. However, it was a very fast algorithm, running at a 50 Hz frame rate on a 750 MHz Pentium computer. The second approach robustly localized the pupillary and limbic boundaries in the presence of highlights (it was able to recover exact pupillary shape even when multiple highlights were present), and when the eye was looking temporally or nasally. However, this second method was rather slow. The first method demonstrated a success rate of 98% on 670 images, with all failures arising from cases with glasses. The second method showed a success rate of 97% on 4574 images of subjects both with and without glasses, as well as other conditions randomly selected from the database. The second method is described below.

Edge Based Seed Method

A computer vision algorithm extracts edge and curvature information from the iris image, which is then used to locate a pair of roughly concentric elliptical boundaries, representing the pupil and iris boundaries.

The algorithm first finds the edges present in the image. The edge detection method utilizes oriented filters in order to compute edge strengths along a pre-set number of orientations, from which the dominant orientation can be found at each image location. Contours of a limited number of possible orientations are formed by linking together neighboring edge pixels with orientations that are close to one another.

Taking each contour in turn, a search is done at each contour endpoint for neighboring contours whose orientations help to complete longer, circular contours. On finding a set of candidate continuous contours, an ellipse is fit to the points belonging to those contours. A confidence measure is computed for each hypothetical

ellipse found in this manner, by comparing hypothetical ellipse points to the oriented edge pixels found in the first step of the algorithm. The ellipse with the highest confidence measure is chosen as the boundary of the inner circle, and the process is repeated to look for a larger ellipse which is roughly concentric with the one already found. Features to restrict the larger ellipse include a minimal and maximal size in relation to that of the smaller ellipse, the degree to which the centers of the two ellipses do not coincide, and the eccentricity of the ellipse.

Similar to the edge focusing method [6], analysis is performed at multiple levels of a Gaussian pyramid, starting at the coarsest level and proceeding to the finest, where the best hypothesis at each level is used as a seed for the next finer level. The advantage of this order of computation lies in cases where edges are diffuse at fine resolution scales, but where the corresponding locations in coarser scales yield more definite and continuous edges. An edge that may not have been considered a contour at a finer resolution would be found at a coarser resolution, allowing contour-linking to occur, in turn leading to a hypothesis that is selected as the one to be tested at finer resolutions.

Manual Initialization

A set of experiments conducted at an early stage revealed that the pupil localization algorithms enabled localization of the pupillary boundary in most images under a broad set of conditions. However, the localization of the limbic boundary, the border between the iris and the sclera is more difficult due to low image contrast in the limbic region, occlusion due to eyelashes and eyelids, and severe foreshortening of the boundary when the subject is looking nasally or temporally. While several algorithms were developed to handle most of the limbic boundary localization, the reliability concerns suggested the use of a manual method to bootstrap the automatic method of performing limbic localization in a sequence. This was performed on all images by manual selection of eight points each along the limbic (sclera) and pupil boundaries, and three points each along the upper and lower eyelids.

Registration of Gallery and Probe

This process recovers and applies the image transformation that warps the probe image into the coordinate frame of the gallery image. It comprises a concatenation of transformations that account for scale, rotation, translation, skew, and foreshortening. These are captured through 2D homography. In addition, a radial warping compensates for pupil dilation or constriction. Finally, optical flow is used to provide small corrections. The transformation parameters (homography, pupil dilation/constriction, and optical flow) that registered gallery and probe images were recovered via a hierarchical coarse-to-fine gradient based estimator [7].

As part of the initial steps for alignment, the shape of the pupil (and in particular its eccentricity and locus of the center) is employed as a first order approximation for the alignment. The introduction of pupillary alignment as part of the first step proved to be important to guarantee correct alignment. In particular, it was experimentally observed to be the most effective manner to introduce the first order of alignment in the presence of geometric deformations. Large geometric deformations occur when the

subject is not looking directly toward the camera but is either fixating at 30° or tracking a target.

Once the global alignment was performed, small alignment corrections, sub-pixel to two-pixels, were achieved using a recovered optical flow. Flow is applied only at lower resolutions and is smoothed prior to its application using a Gaussian kernel. The smoothing eliminates any discontinuities in the flow field, thus representing a transformation that is more consistent with the physiological characteristics of tissue. If the flow field were allowed to be discontinuous, then flow vectors would attempt to align local iris tissue in an effort to match iris texture, even when the gallery and probe images are of different irises. This could significantly modify the match results for imposters, since the alignment algorithm itself would then be introducing a correlation between the gallery and probe images. The matching results, presented next, demonstrate that the smoothed alignment is capable of compensating for a broad range of deformations yielding good separation between authentics and impostors, even when two images have been acquired at different angles (fixating at 0° vs. 30°).

Similarity Score Between Gallery and Probe

Once gallery and probe iris images were aligned, the distance between them was measured. The simplest comparison metric is a single correlation value between corresponding pairs of pixels where iris tissue is present in both images. The problem with this approach, however, is that it assumes that the distribution of intensities is stationary across different regions of the images; an assumption that is easily violated due to non-uniform illumination.

A better approach is to compute the average correlation coefficient over sub-blocks. This ensures stationarity within the region of comparison. Daugman [8] implemented a similar approach, except he used Gabor filters rather than Laplacian filters, used polar coordinate systems in his matching, and binary rather than gray-scale matching. In the method used here, the similarity score is obtained as the average local correlation coefficient between corresponding sub-blocks of the gallery and probe images at the finest resolution image scale

$$score = E_{m,n}\left[C(m,n)\cdot B(m,n)\right] \tag{1}$$

where E[] denotes expectation over the image plane. The local normalized correlation coefficient [-1,1] is

$$C(m,n) = \frac{\sum_{i,j=1}^{12} G(m+i,n+j)\cdot P(m+i,n+j) - \sum_{i,j=1}^{12} G(m+i,n+j)\cdot \sum_{i,j=1}^{12} P(m+i,n+j)}{\sqrt{\left\{\sum_{i,j=1}^{12} G^2(m+i,n+j) - \left(\sum_{i,j=1}^{12} G(m+i,n+j)\right)^2\right\}\cdot\left\{\sum_{i,j=1}^{12} P^2(m+i,n+j) - \left(\sum_{i,j=1}^{12} P(m+i,n+j)\right)^2\right\}}} \tag{2}$$

where and G and P are the gallery and probe images, respectively. The binary indicator, B, is true only when all the mask pixels of the corresponding sub-block are true

$$B(m,n) = 1 \quad \textit{iff} \quad \sum_{i,j=1}^{12} M(m+i,n+j) = 144 \tag{3}$$

where M is the overall mask, obtained as the logical AND between the gallery and probe masks. Although this is a conservative approach, it eliminates edge effects

where the local correlation coefficient might be determined by only a few valid pixels. It is also robust to minor errors in the semi-automatic determination of the mask. Several different sub-block sizes of 4x4, 8x8, … up to 32x32 pixels were tried, with the best results obtained with a sub-block size of 12x12 pixels. Overlapping sub-blocks were allowed, and this resulted in each similarity score being based on an average of 6,300 local correlation scores, compared to an average overall mask size of 10,660 pixels at 5 m, and 10,900 pixels at 10 m. Note that the overall score is within [-1,1], and negative correlation scores can occur due to coincidental contrast reversal.

Performance Measures

For *N* subjects, there were *N* authentic scores, and *N* (*N*-1) imposter scores. The latter result is because the registration methods are non-symmetric. The set of authentic scores and the set of imposter scores were viewed as distributions, and the distance between them was used to measure performance.

Separation between the authentic and imposter distributions was measured in terms of the area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of true versus false positives, and the area under the curve is a number between 0.5 and 1 which represents the discrimination power over a range of thresholds. It is the probability that a randomly selected subject from the gallery set will have a larger similarity score when compared against him/herself versus an impostor in the probe set.

When there is little or no overlap between the measured distributions of the impostor and authentic similarity scores, the area under the ROC curve ceases to be a useful metric for comparison. One possible alternative is to fit parametric distributions to the measured scores, and then calculate the ROC from the theoretical distributions. Although the similarity score is obtained from the sum of many local correlation scores, the central limit theorem may not apply because the scores are not entirely independent.

We therefore opted for a simpler performance measure; the Fisher Linear Discriminant (FLD). The FLD provides a useful measure of the separation of two class conditional distributions, under the assumption that the distributions are symmetric and completely characterized by their first and second-order moments. It is a measure of the ratio of the between class scatter to the within class scatter. It thus provides a useful performance measure even when there is no overlap between the measured distributions.

$$FLD = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{4}$$

Analysis of Database I

These ten experiments explore the effect of three fundamental variables on the similarity score distributions when acquisitions are subject to changes in time, capture angle, and eyewear. They were acquired at a distance of 10 m under background illumination (IR only). The results are summarized in Table 2. Time refers to one of two

sessions per subject, separated by approximately one month. Experiments with the same Time were acquired during the same session. We now proceed to describe the experimental results in greater detail.

Table 2. Database I experiments. Time refers to one of two sessions per subject, separated by ~1 month. Eyewear designations are: NG (no glasses), G (glasses), or NGB (no glasses baseline). The NGB set consists of the 24 subjects who wore glasses in general, but in this set are not wearing glasses. An ROC area of 1 indicates perfect separation of imposter and authentic distributions

Experiment Info								Score Distributions				Distances	
ID	# Subjects	Gallery			Probe			Imposters		Authentics		ROC area	FLD
		Time	Angle (°)	Eyewear	Time	Angle (°)	Eyewear	μ	σ	μ	σ		
1	50	1	0	NG	1	0	NG	0.17	0.06	0.75	0.08	1	32.8
2	50	1	0	NG	2	0	NG	0.16	0.06	0.63	0.08	1	20.6
3	50	1	0	NG	1	30	NG	0.15	0.06	0.50	0.07	1	13.1
4	50	1	0	NG	2	30	NG	0.14	0.06	0.47	0.08	0.993	10.2
5	24	1	0	NGB	2	0	NGB	0.14	0.06	0.77	0.06	1	57.0
6	23	1	0	G	2	0	G	0.15	0.07	0.44	0.12	0.970	4.4
7	24	1	0	G	1	0	NGB	0.16	0.06	0.51	0.12	0.996	7.0
8	23	1	0	G	2	0	NGB	0.16	0.06	0.49	0.10	0.994	7.8
9	24	1	0	NGB	1	0	G	0.16	0.07	0.51	0.11	0.996	7.6
10	23	1	0	NGB	2	0	G	0.15	0.07	0.46	0.13	0.955	4.4

Experiment 1 establishes baseline performance of the system. Both the gallery and the probe consist of all subjects, acquired during the same session, and with no subject wearing glasses. It identifies the degradation due to noise in the system at the acquisition, alignment and matching components. The ROC area indicates perfect separation in this case.

Experiment 2 studies the impact of time. Both the gallery and probe consist of all subjects, with no subject wearing glasses, but where the gallery was captured in session 1, and the probe in session 2. The ROC area reflects perfect separation, but the FLD decreases relative to experiment 1, due to a reduction in the mean of the authentics. Degradation reflects changes due to geometrical mis-alignment, pupil dilation and other physiological changes that might impact the iris appearance.

Experiment 3 studies the impact of capture angle. Both the gallery and the probe consist of all subjects, acquired during the same session, and with no subject wearing glasses. In the gallery the subject is facing the camera (0°), while in the probe the subject is turned by 30°. Relative to the baseline experiment 1, the ROC remains at 1, but the FLD decreases, again due to a reduction in the mean of the of the authentics. Degradation reflects primarily geometrical mis-alignment.

Experiment 4 studies the impact of time and angle, effectively combining the variables in experiments 2 and 3. The gallery consists of all subjects from session 1, facing the camera (0°). The probe consists of the same subjects, but acquired during session 2, and turned by 30°. Unlike experiment 3, there is now some overlap between the imposter and authentic distributions. However, only a slight performance decrease relative to experiment 3 is observed, indicating the potential for iris recognition at a distance for vastly different capture angles.

Experiment 5 establishes a baseline for the subset of subjects who normally wear glasses, but in this experiment are not wearing glasses. The gallery was captured in session 1, while the probe was captured in session 2.

Experiment 6 studies the effect of time on subjects that wear glasses. Both the gallery and probe are wearing glasses. The gallery was captured in session 1, while the probe was captured in session 2. Compared with experiment 5, glasses have a dramatic effect on both the authentic mean and variance, and results in overlap between the authentic and imposter distributions. Note that the number of subjects differs between experiments 5 and 6 because one subject wore contacts in session 2.

Experiment 7 establishes a baseline for experiment 8. The gallery consists of subjects wearing glasses in session 1, while the probe consists of the same subjects not wearing glasses in the same session.

Experiment 8 studies the effect of both time and a change in eyewear. The gallery consists of subjects wearing glasses in session 1, while the probe consists of those same subjects not wearing glasses in session 2. Relative to experiment 7, there is almost no difference in performance. However, performance is better than experiment 6, indicating that having a view of the subject without glasses in at least one session, is preferable to both sessions with glasses.

Experiment 9 establishes a baseline for experiment 10. The gallery consists of subjects that normally wear glasses, but are not wearing them in session 1, while the probe consists of the same subjects wearing glasses in the same session.

Experiment 10 measures the effect of both time and a change in eyewear. The gallery consists of subjects that normally wear glasses, but are not wearing them in session 1, while the probe consists of the same subjects wearing glasses in session 2. It is the reverse of experiment 8, and thus measures symmetry in the matching process. Compared to both experiment 8 and 9, performance decreases, mostly due to a drop in the mean of the authentics, but also due to a small increase in the variance. Thus, for this system, it is better that the reference image is without glasses.

Analysis of Database II

The following six experiments, summarized in Table 3, are designed to study the effect of time and distance. All acquisitions were performed with the subject fixating at a target at 0° under background illumination (IR only), with no eyewear (subjects who normally wear glasses were asked to remove them).

Experiment 11 defines a baseline for the scalability of the dataset. The gallery consists of subjects acquired at either 5 m or 10 m during session 1, while the probe consists of the same subjects acquired during session 2 (some subjects were acquired at the same distance in both sessions, while others were acquired at 5 m in one session, and 10 m in the other session). It addresses the degradation of the verification process subject to changes in both time and distance. In this case the degradation reflects changes due to geometrical mis-alignment and pupil dilation and other physiological changes that might impact the iris appearance, as well as distance.

Table 3. Database II experiments. Time refers to one of two sessions per subject, separated by ~1 month. An ROC area of 1 indicates perfect separation of imposter and authentic distributions

Experiment Info						Score Distributions				Distances	
ID	# Subjects	Gallery		Probe		Imposters		Authentics		ROC area	FLD
		Time	Dist. (m)	Time	Dist. (m)	μ	σ	μ	σ		
11	50	1	5,10	2	5,10	0.15	0.08	0.65	0.10	0.9996	15.5
12	270	1	5,10	2	5,10	0.14	0.08	0.67	0.09	0.99996	18.8
13	67	1	5	1	5	0.15	0.08	0.81	0.07	1	40.0
14	67	1	5	2	5	0.17	0.09	0.70	0.08	1	16.2
15	68	1	10	1	10	0.13	0.06	0.78	0.06	1	52.9
16	68	1	10	2	10	0.13	0.06	0.68	0.08	1	28.0

Experiment 12 studies the scalability to a larger population of the results from experiment 11, to which the other conditions are identical. This experiment includes subjects from both Databases I and II, and results in the largest possible number of comparisons. The authentic and imposter distributions and ROC curve for this experiment are shown in Figure 4. The ROC area and FLD distances are comparable to experiment 11, indicating that performance is consistent with an increasing number of subjects. It is interesting to note that a single subject was responsible for 33 of the worst 100 imposter scores, including the highest impostor score.

Experiment 13 constitutes the comparison benchmark for experiment 14. Both the gallery and the probe consist of subjects acquired during the same session at a distance of 5 m. It also serves to measure the noise in the 5 m acquisition.

Experiment 14 measures the effect of time at a fixed distance of 5 m. The gallery consists of subjects from session 1, while the probe consists of the same subjects from session 2. The ROC area indicates perfect separation, while the FLD indicates decreased separation relative to experiment 13, mostly due to a lowering of the average authentic score. However, there is also a small but unusual increase in the average imposter score relative to the baseline.

Experiment 15 constitutes the comparison benchmark for Experiment 16. Both the gallery and probe consist of subjects acquired during the same session at a distance of 10 m. The ROC area indicates perfect separation, while the FLD indicates no performance degradation relative to the 5 m capture in experiment 13, indicating that the system noise does not increase as a function of distance.

Experiment 16 measures the effect of time at a fixed distance of 10 m. The gallery consists of subjects from session 1, while the probe consists of the same subjects acquired during session 2. The ROC area indicates perfect separation, while the FLD indicates decreased separation relative to experiment 15. However, once again, no performance degradation is observed relative to the same experiment performed at 5 m in experiment 14. In fact, the FLD shows that 10 m performance is better than 5 m performance for the same conditions.

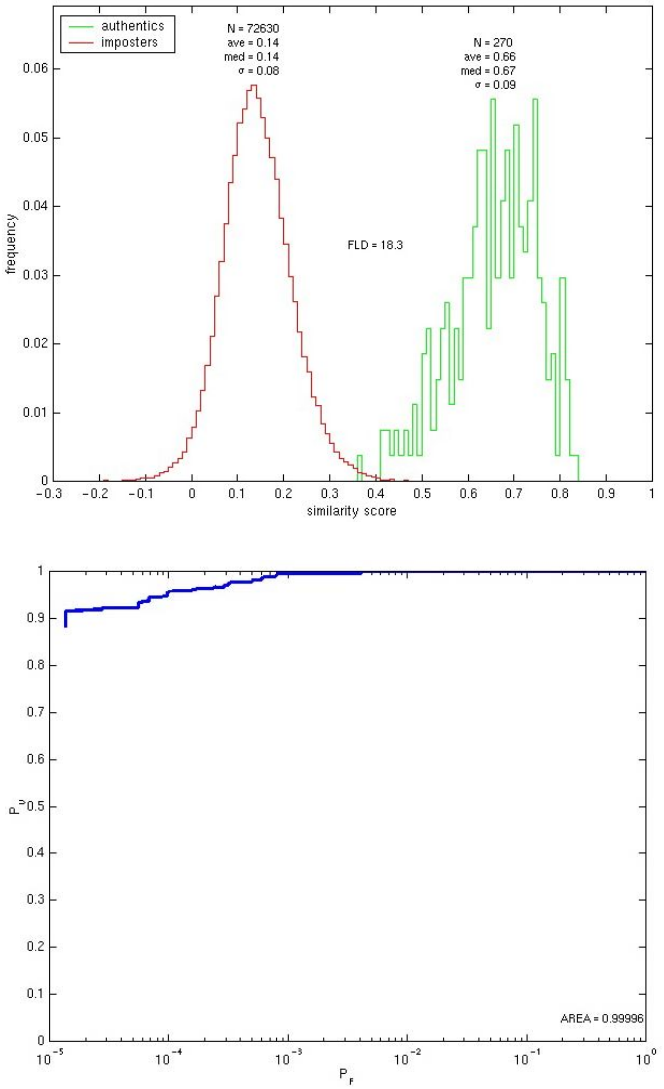


Fig. 4. Experiment 12 results: (top) similarity score distributions for authentics (green) and imposters (red); (bottom) ROC curve

Conclusions

In this paper we have reviewed the effects of different variables on iris recognition at a distance by exploring a set of representative experiments. Some general observations, as well as the results of the experiments, are summarized here.

Increasing the distance from 5 m to 10 m does not impact performance. This is somewhat expected, given that the hardware was designed to provide similar and sufficient spatial resolution at both distances. However, while these experiments have

demonstrated that it is possible to acquire iris images at a large distance, it is important that these images are in focus and present good contrast resolution.

Time does impact the results, however, in no case did the effect of time alone cause overlap between the imposter and authentic distributions.

Angle, relative to the acquisition device, introduced greater performance loss compared to time, but angle alone did not lead to an overlap between the imposter and authentic distributions. However, this analysis was limited to a change in pose of 30°. Larger pose changes, and/or improvement of these results, will most likely require an alignment procedure that takes into account the 3-D structure of the iris, in order to better correct for foreshortening and varying illumination across the surface. The tracking experiments, where angle is continuously varied, are likely to be very useful for this purpose.

Eyewear effected performance more than the other variables considered. This is most likely due to specularities caused by glasses. Simple enhancement, such as thresholding the pixel intensity, suggests a modest improvement. However, a proper modeling of the data distribution for the highlights and iris data in conjunction with an opportunistic approach may yield significant improvement.

For different capture conditions of the probe and gallery, such as varying angle or eyewear, false negatives or positives are typically caused by a lowering of authentic scores, rather than a raising of imposter scores, relative to baseline experiments. This is consistent with Daugman's observation that the distribution of imposters depends more on the uniqueness of the iris code, while the distribution of authentics depends more on the acquisition system. However, small increases in the imposter scores do occur, as was observed in experiment 14, and are most likely due to a correlated noise source or artifact present in both gallery and probe sessions.

Scalability of results from 50 to 270 subjects, compared over both time and distance, suggest that the underlying behavior seen is indicative of the trends expected in a larger population.

A single subject with an unusual iris pattern or capture condition may be responsible for many false matches, as was seen in experiment 12. Such subjects are sometimes called wolves [9]. The appearance of portions of two irises at a certain scale may in fact be correlated. However, it is likely that a multi-scale match approach would have identified a difference in iris texture at other scales, thereby differentiating impostors more effectively.

Algorithmic limitations, such as the alignment and mask delineation, have a large impact on the quality of the matching and scoring. This is of particular significance in the presence of specularities. The automatic localization algorithms that have been developed show promise but further tuning is required. Being able to automatically and robustly extract iris boundaries will provide a basis for extending the set of experiments that can be performed. In addition, further experiments could be performed on the database aimed at comparing and improving the current matching process (and employed algorithms in general).

The results from these experiments provide further evidence of the viability of iris verification as a biometric tool. In addition, the large databases acquired, covering a broad spectrum of environmental variables, provide a means for developing new algorithms. However, prior to actually fielding a system for iris recognition at a distance, several topics still need investigation. First, the data acquisition process should allow

more freedom of movement of the subjects, by removing head constraints, and allowing the subjects to stand and walk. Clearly this would require expanding the current hardware to a system capable of acquiring images from a moving subject. Second, further refinement of the current algorithms is needed for localizing, registering, and matching irises captured under very different environmental conditions.

Acknowledgements

This work was sponsored by the Defense Advanced Research Projects Agency under Department of Interior contract #DABT63-00-C-1042. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

References

1. Wildes R., Iris Recognition: An Emerging Biometric Technology, *Proc. of the IEEE,* vol.85, pp.1348-1363, 1997.
2. Ma L., Tan T.N., Wang Y, and Zhang D., Personal identification based on iris texture analysis, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1519-1533, 2003.
3. Mansfield, T., Kelly, G., Chandler, C. and Kane, J., Biometric Product Testing Final Report, Technical Report, Centre for Mathematics and Scientific Computing, National Physics Laboratory, Teddington, Middlesex, UK, 2001.
4. Daugman J., How iris recognition works, *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21-30, January 2004.
5. Camus T. and Wildes R., Reliable and fast eye finding in close-up images, in *Proc. Int. Conf. on Pattern Recognition*, pp. 389-394, 2002.
6. Bergholm F., Edge focusing*, IEEE Trans. Pattern Analysis and Machine Intelligence*, 19:726-741, 1987.
7. Bergen J.R., Anandan P., Hanna K., and Hingorani R., Hierarchical model-based motion estimation, in *Proc. 2nd European Conference on Computer Vision*, pp. 237-252, 1992.
8. Daugman J. and Downing C., Epigenetic randomness, complexity, and singularity of human iris patterns*. Proceeding of the Royal Society Biological Sciences*, **268**, pp. 1737-1740, 2001.
9. Doddington G., Liggett W., Martin A., Pryzybocki, and Reynolds D., Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation, in *Proc. of 5th Int. Conf. of Spoken Language Processing*, 1998.

Iris Recognition Using Fourier-Wavelet Features

Ping S. Huang, Chung-Shi Chiang, and Ji-Ren Liang

Department of Electrical Engineering, Chung Cheng Institute of Technology,
National Defense University, Taoyuan 335, Taiwan, R.O.C.
{pshuang,g950302,g941311}@ccit.edu.tw
http://www.ccit.edu.tw/~elec/

Abstract. This paper presents an effective iris recognition system for iris localization, feature extraction, and matching. By combining the shift-invariant and the multi-resolution properties from Fourier descriptor and wavelet transform, the Fourier-Wavelet features are proposed for iris recognition. A similarity measure is adopted as the matching criterion. Four wavelet filters containing Haar, Daubechies-8, Biorthogonal 3.5, and Biorthogonal 4.4 are evaluated and they all perform better than the feature of Gaussian-Hermite moments. Experimental results demonstrate that the proposed features can provide promising performance for iris recognition.

1 Introduction

Biometrics nowadays has become an important technique to verify human's identification by his/her own biological features. The advantages provided are that they cannot be stolen or forgotten like passwords. Current applications of biometrics include face, fingerprint, iris, palm-prints, and gait etc. The lowest error recognition rate has been achieved by iris recognition [1] that has received increasing attention in recent years.

The eye appearance consists of sclera, iris, and pupil. The boundaries of sclera, iris and pupil are like circles with varied radii. Sclera is the outside portion of the eye occupying about 30% area of the eye. The central portion of the eye is the pupil including 5% area of the eye. Iris is the colored portion of the exterior eye, which is embedded with tiny muscles that affect the pupil size, about 65% area of the eye [2]. The color, texture, and patterns of each person's iris are considered as unique as the fingerprint. The formation of iris patterns is random and unrelated to any genetic factors [3]. Therefore, iris is a stable and long-lasting feature of biometrics suitable for human identification and verification.

Most commercial iris recognition systems use patented algorithms developed by John Daugman [4] to generate the iris code using 2D Gabor filters. By contrast, Fourier descriptor and wavelet transform are widely applied in pattern recognition [5]. A merit of the Fourier transform is shift-invariant [6, 7], that is, the information in frequency domain is not affected by its translation in spatial domain. However, the Fourier transform lacks the property of multi-resolution. The wavelet transform [8] is used to analyze the local frequency for that low frequency has long-time resolution and high frequency provides short-time resolution. In this paper, we propose to combine the shift-invariant and the multi-resolution properties of these two transforms as Fourier-Wavelet features for iris recognition.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 14–22, 2005.
' Springer-Verlag Berlin Heidelberg 2005

In general, iris recognition procedures can be divided into three parts: iris localization, feature extraction, and feature matching (identification and verification). Before any matching algorithms are applied, it is essential to detect and localize iris patterns accurately. Most of the approaches are using edge detection to localize the iris boundaries. Here, we adopt Daugman's integro-differential algorithm [4] instead. In the stage of feature extraction, 1D Fourier transform and 1D wavelet transform are sequentially adopted to extract the iris features and then an iris feature vector is created. In feature matching stage, Daugman proposed a method to generate iris binary code and Hamming distance is used to distinguish the authentic one from the imposter [4]. One commonly used similarity measure is adopted for feature matching. Since it is easier and faster to match two features than coding, we use it for feature matching in this work. The test data used is the CASIA iris image database collected by Institute of Automation, Chinese Academy of Science [9].

2 Image Pre-processing

The human iris is an annular portion between the pupil (inner boundary) and the sclera (outer boundary). Both the inner and the outer boundaries of a typical iris can be approximated as circles. Despite those two circles are not true concentric, they can be treated as concentric for that the iris patterns are changeless near the outer boundary. Based on this, the first step of iris recognition is to localize the pupil center representing the concentric point of the inner and outer boundaries. The pupil localization is implemented by the following methods.

2.1 Iris Localization

Since the pupil is normally darker than any other portion in an eye image. By adjusting the threshold value to generate a binary image, the pupil region is filled only with pixels of value 1. According to the region gravity method [10], the pupil center is given by

$$X_0 = \frac{1}{A}\sum_{(x,y)\in R} x \quad \text{and} \quad Y_0 = \frac{1}{A}\sum_{(x,y)\in R} y, \tag{1}$$

where (X_0, Y_0) is the coordinates of the pupil center, A is the area of region R (including the pupil region). That is, the pupil center is calculated by the averaged summation of pixel coordinates with value 1 along X and Y axis respectively.

Daugman [4] proposed an effective algorithm to find the inner and outer boundaries using an integro-differential operator given by

$$\max_{(r,x_0,y_0)} \left| G_\sigma(r) * \frac{\partial}{\partial r} \oint_{r,x_0,y_0} \frac{I(x,y)}{2\pi r} ds \right| \tag{2}$$

where * denotes the convolution and $G_\sigma(r)$ is a smoothing function such as a Gaussian of scale σ. $I(x,y)$ is a raw input image processed by integro-differential operators by searching over the image domain (x,y) for the maximum in the blurred partial derivative. The parameter space of center coordinates and radius (x_0, y_0, r)

defines the path of contour integration, with respect to increasing radius r, of the normalized contour integral of $I(x, y)$ along a circular arc ds of radius r and center coordinates (x_0, y_0).

To reduce the influence from eyelashes and eyelids, we use Daugman's algorithm and the method is described as follows. At first, the original eye image is segmented into two parts, left and right, based on the pupil center along Y axis. The ranges of the right part is set from $-\pi/6$ to $\pi/6$ and the left part is set from $5\pi/6$ to $7\pi/6$. By applying Daugman's integro-differential operator with increasing radius r, starting from the pupil center to the margin and summing the gray values of pixels in every sector of each contour, then the average gray values of each contour are calculated. The inner and outer boundaries are where the sharp peak values located.

2.2 Iris Normalization

The pupil will dilate or constrict when eye images are captured with flash light or in the darker circumstance. On the other way, it may be affected by illumination, light source, and distance. Even captured from the same people, the iris size may change by the mentioned factors. Such deformation will influence the matching result. This problem is solved by converting the annular iris template into a fixed size. The iris region is transformed from the Cartesian coordinates (x, y) to the polar coordinate (r, θ). The rectangular texture size is 64×512 in our experiments.

2.3 Feature Enhancement

After normalization, iris templates still have low contrast and non-uniform illumination problems. To eliminate the background brightness, the iris template is divided into non-overlapped 16×16 blocks and their means constitute coarse estimates of background illumination for individual blocks. By using bicubic interpolation, each estimated value is expanded to the size of a 16×16 block. Then each template block can be enhanced to a uniform light condition by subtracting from the background illumination. After that, the lighting corrected images are enhanced by histogram equalization. The images showing temporary results of the feature enhancement are demonstrated in Figure 1.

3 Feature Extraction

Despite all normalized iris templates have the same size and uniform illumination; there are still eyelashes and eyelids on them. Therefore, the region of interest (ROI) is selected to remove the influence of eyelashes and eyelids. The features are extracted only from the upper half section (32×512). This way can eliminate most of the influence and provide purer iris templates for feature extraction.

3.1 1-D Fourier Transform

A well-known property of the Fourier transform is shift-invariant of power spectrum, that is, the information in frequency domain is not affected by its translation in spatial

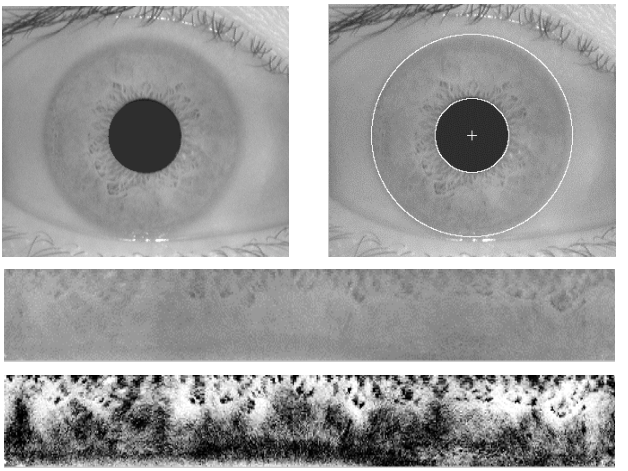


Fig. 1. The upper left image is the original image from CASIA. The upper right is the image where pupil center, inner and outer boundaries are located. The middle image is the normalized iris template with size of 64x512 in polar coordinate. The lower one is the image after feature enhancement

domain [6, 7, 11]. Since the spectra of Fourier transform for circularly shifted signals are the same, the difference between the same but shift-variant iris templates can be mitigated. The 1-D Fourier transform is performed at the iris template along the axis of the polar angle θ to obtain its spectrum. The iris template has been represented by $I(r,\theta)$ in polar coordinates. The symbol θ in the Cartesian coordinate represents the annular portion $[0,2\pi]$ of the iris region. By using the 1-D Fourier transform, the θ values along each r become shift-invariant and the result can be represented by

$$\mathbb{F}'(r,\phi) = \left| FT_{\theta}(I(r,\theta)) \right| \tag{3}$$

where $I(r,\theta)$ represents the normalized iris template in the polar coordinate, and FT_{θ} depicts that the 1-D Fourier transform is individually applied to each θ. $\mathbb{F}'(r,\phi)$ denotes the power spectrum after 1-D Fourier transform, a shift-invariant feature.

Since the ROI size of each iris template is 32×512, the number of spectrum coefficients remains the same size as the ROI. Due to that the coefficients of each row are symmetric, we delete the DC value and keep half of the coefficients as the template spectrum coefficients with the size of 32×255.

3.2 1-D Wavelet Transform

The results of wavelet transform can represent pattern features at different levels. Here, we apply the wavelet transform at the template spectrum coefficients along the axis of radius (each column), so that we can query the pattern feature database from coarse scales to fine scales. Wavelet transform is well suited for localized frequency analysis, because the wavelet basis functions have short time resolution for high fre-

quencies and long time resolution for low frequencies [5]. The wavelet transform can be represented as follows and we can obtain the wavelet coefficients, $W\mathbb{F}''(r,\phi)$, by

$$W\mathbb{F}''(r,\phi) = WT_r(\mathbb{F}'(r,\phi)) \tag{4}$$

3.3 Feature Vector

After the multi-resolution analysis by the wavelet transform, the accomplished feature dimensionality depends on the applied levels of the wavelet transform. The feature vector $\mathbf{V}$ is given by

$$\mathbf{V} = [\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N] \tag{5}$$

where $\mathbf{R}_1$ to $\mathbf{R}_N$ represent the row feature vectors and N denotes the maximum row numbers derived from the multi-resolution analysis of wavelet families.

4 Matching

The main goal of iris recognition is to match the feature and determine whether the feature comes from the authentic one or the imposter. The general approach is to transform the feature vector into the iris code and calculate the Hamming distance as matching criterion [4]. In this paper, the feature vector is a set of wavelet coefficients, so we adopt the similarity measure as the matching criterion.

4.1 Similarity Measure

The similarity measure used to compute the similarity of two feature vectors is represented by

$$\delta_{Cov}(\mathbf{V},\mathbf{F}) = 1 - \left(\frac{\mathbf{V}}{\|\mathbf{V}\|} \bullet \frac{\mathbf{F}}{\|\mathbf{F}\|}\right) \tag{6}$$

where $\mathbf{V}$ and $\mathbf{F}$ represent two feature vectors, $\|\bullet\|$ denotes the Euclidean norm. The term $(\mathbf{V}/\|\mathbf{V}\|) \bullet (\mathbf{F}/\|\mathbf{F}\|)$ computes the cosine similarity between two vectors $\mathbf{V}$ and $\mathbf{F}$. The range of $(\mathbf{V}/\|\mathbf{V}\|) \bullet (\mathbf{F}/\|\mathbf{F}\|)$ is $[0,1]$. The more similar the two vectors are, the larger the $\delta_{Cov}(\mathbf{V},\mathbf{F})$ value is.

5 Experimental Results

After feature extraction, the matching step is to distinguish the imposter from the authentic one or to verify if the test sample is genuine. By comparing two iris feature vectors using the similarity measure in (6), we can obtain a false acceptance rate (FAR) versus false reject rate (FRR) curve and the crossover point is the threshold value where the equal error rate (EER) is the smallest. The way to verify the template comes from the authentic one or the imposter depends on the threshold value matching. By using the adjustable threshold, the receiver operating curve (ROC) [12] can be created. The ROC curve plots the FAR against (1-FRR) to demonstrate the per-

formance of a matching algorithm. One commonly used performance measure derived from the ROC curve is the area under the ROC curve (denoted as Az). The Az reflects how well the intra- and inter-class distributions and the ranges are from 0.5 to 1. In the ideal ROC curve, the Az should be 1. And if the Az is equal to 0.5, it denotes that the intra- and inter-class are inseparable.

5.1 Iris Database

Our iris database is authorized from the CASIA iris image database collected by Institute of Automation, Chinese Academy of Science [9]. Figure 2 shows four iris samples from the CASIA iris database. The database comprises 108 classes of eye images, and each class has 7 images (total 756 images). Whole classes come from 108 different eyes, and each image with the resolution of 320×180 in gray level. There are two groups in each class, the first one has 3 and the second one has 4 images. Each group denotes different time capturing. In iris localization process, we can successfully localize the iris regions of the 756 images (successful rate 100%) and normalize them to polar coordinates.

Among those 756 images (108 classes), we choose 468 images coincided with our normalized algorithm that top 50% of templates is useful. Those 468 images are divided into 68 classes, and each of them has 5 to 7 images. The first 2 images of each class are selected to be the training set, so the total training set has 136 images. The remaining 3 to 7 images of each class are used to be the test set with 332 images.

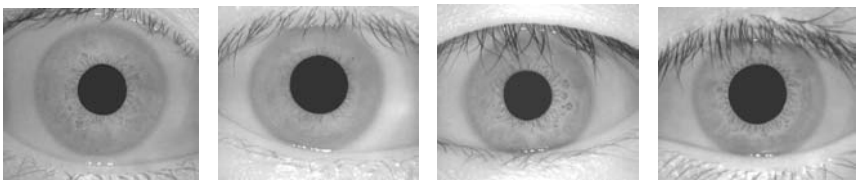


Fig. 2. Iris samples from CASIA Iris Database

5.2 Recognition Results

As mentioned above, wavelet transform is widely used for its multi-resolution property [13] that is suitable for localized frequency analysis. In the first experiment, different kinds of wavelet families are evaluated for the 1-D wavelet transform in Section 3.2. To compare the performance achieved by different wavelet filters, we use four wavelet filters containing Haar, Daubechies-8, Biorthogonal 3.5, and Biorthogonal 4.4. The feature vectors are obtained by the 1^{st}-scale decomposition including only low frequency coefficients. This is because the high frequency scale coefficients contain feature details and noise. The experimental results are shown in Table 1, and the ROC curve is shown in Fig. 3. Figure 3 shows that the areas under the ROC curve accomplished by different wavelet families are all near 1. It means that the performance of FAR and FRR achieved by each wavelet is very good. From Table 1, the correct recognition rates obtained from those wavelet families are similar and satisfied. Based on the results in Figure 3 and Table 1, we can conclude that each of those four wavelet transforms can provide promising performance.

Table 1. Recognition rate of different wavelet transform filters

Wavelet filter	Az (area under the ROC curve)	Correct Recognition rate (%)
Haar	0.9997	93.58
Daubechies-8	0.9998	95.47
Biorthogonal 3.5	0.9997	92.15
Biorthogonal 4.4	0.9997	93.21

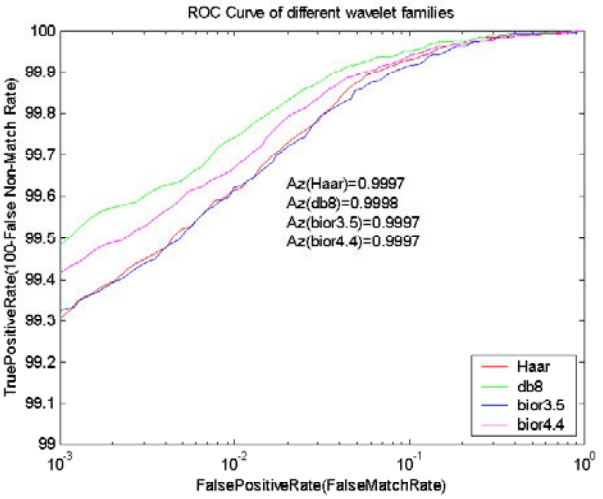


Fig. 3. The ROC curve of the proposed method using different wavelet families

Based on the first experiment, in the second experiment, we select the Haar wavelet as our testing wavelet filter in the next experiment. For each wavelet scale, we only choose the low frequency coefficients to form the feature vector. This is because the high frequency coefficients contain edge information and noise that is useless for recognition. The dimensionality of each feature vector will reduce to half of each upper scale. The experimental results in Table 2 show the best correct recognition rate of wavelet scale appears at the 1st-scale decomposition. The more dimensionality the feature vector contains, the more feature information it preserves.

Table 2. Recognition rate of wavelet transform in different scales (1 to 3 scales)

Wavelet scale	Dimensionality of features	Correct recognition rate (%)
1st-scale	4080	93.58
2nd-scale	2040	92.34
3rd-scale	1020	89.92

By using the Daubechies-8 wavelet, the ROC curve and the area under the ROC curve (Az) of our proposed method is shown in Fig. 4. In the third experiment, the total trials are 39440. Figure 4 illustrates the ROC curves of 1st-scale of Daubechies-8 wavelet and Li Ma et al.'s method [12] that adopted Gaussian-Hermite moments to characterize local variations of intensity signals. Here, we implement the Li's algorithm without using Fisher linear discriminant (FLD) and the downsampling factor is

set to 8. The correct recognition rate is shown in Table 3. Table 4 lists three typical operating states of the proposed method and shows that the false reject rate of our proposed algorithm is only 0.48% when if one and only one false match occurs in 100,000 trials. The experimental results demonstrate that our propose algorithm is suitable to distinguish one iris feature to another.

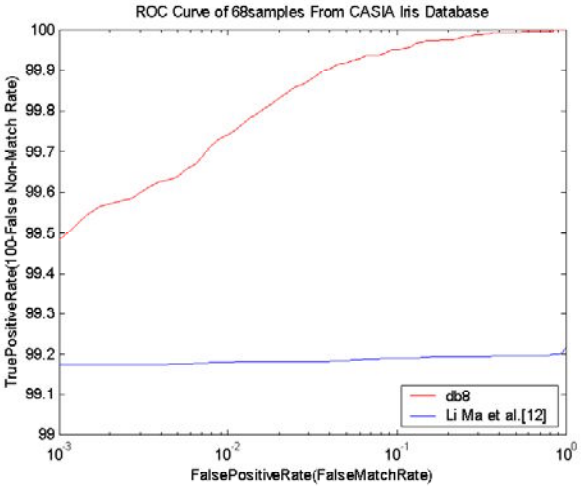


Fig. 4. The ROC curves of the proposed method and Li Ma et al. [12]

Table 3. Performance comparison of the Li's method

Method	Proposed	Li Ma et al.[12]
Dimensionality of features	4080	2560
Az (area under the ROC curve)	0.9998	0.9920
Correct recognition rate (%)	95.47	87.2

Table 4. The typical operating states of the proposed method with 1st-scale of Daubechies-8 wavelet

FAR (%)	FRR (%)
0.001	0.48
0.01	0.3
0.1	0.08

6 Discussion

This paper presents an effective iris recognition system for iris localization, feature extraction, and matching. In iris localization process, we have modified Daugman's algorithm and can successfully localize the iris regions of the 756 images (successful rate 100%) in the experiments. By combining the shift-invariant and the multi-resolution properties from Fourier descriptor and wavelet transform, the Fourier-Wavelet features are proposed for iris recognition. Four wavelet filters containing Haar, Daubechies-8, Biorthogonal 3.5, and Biorthogonal 4.4 are evaluated and the

best performance is achieved by using the Daubechies-8 filter. According to Table 4, the false reject rate of our proposed algorithm is only 0.48% when if one and only one false match occurs in 100,000 trials. Experimental results demonstrate that all four Fourier-Wavelet features perform better than the feature of Gaussian-Hermite moments. Although the proposed features can provide promising performance for iris recognition, a larger database is needed to further verify its robustness in the future.

Acknowledgement

The authors would like to thank the Institute of Automation, Chinese Academy of Science for providing the CASIA iris image database and the National Science Council of the Republic of China for financially supporting this work under Contract No. NSC 93-2218-E-014-003.

References

1. A. Jain, R. Bolle and S. Pankanti, Eds.: Biometrics - Personal Identification in Networked Society. Kluwer Academic Publishers (1999)
2. B. Miller.: Vital Signs of Identity. Vol. 31.IEEE Spectrum (1994) 22-30
3. R. P. Wildes.: Iris Recognition: An Emerging Biometric Technology. Vol. 85, no. 9. Proceedings of the IEEE (1997) 1348-1363
4. J. G. Daugman.: How Iris Recognition Works, Vol. 14, no. 1. IEEE Transactions on Circuits and Systems for Video Technology (2004) 21-30
5. T. D. Bui and G. Chen.: Invariant Fourier-Wavelet Descriptor for Pattern Recognition. Vol. 32, no. 7. Pattern Recognition (1999) 1083-1088
6. R. N. Bracewell.: The Fourier Transform and Its Application. The McGraw-Hill Companies (2000)
7. D. Casasent and D. Psaltis.: Position, Rotation, and Scale Invariant Optical Correlation. Vol. 15, no. 7. App. Opt. (1976) 1795-1799
8. S. Mallat.: A Wavelet Tour of Signal Processing. Academic Press Publishing Company (1998)
9. Institute of Automation, Chinese academy of Science, CASIA Iris Image Database. http//www.sinobiometrics.com/chinese/chinese.htm.
10. T. Bernier and L. Jacques-Andr.: A New Method for Representing and Matching Shapes of Natural Objects. Vol. 36, no. 8. Pattern Recognition (2003) 1711-1723
11. A. Grossmann and J. Morlet.: Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape. Vol. 15, no. 4. SIAM Journal of Math. Anal (1984) 723-736
12. Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang.: Local Intensity Variation Analysis for Iris Recognition. Vol. 37. Pattern Recognition (2004) 1287-1298
13. Y. Y. Tang, B. F. Li, H. Ma, and J. Liu.: Ring-Projection-Wavelet-Fractal Signatures, A Novel Approach to Feature Extraction. Vol. 45, no. 8. IEEE Trans. on circuits and systems-II: Analog and digital signal processing (1998)

Specific Texture Analysis for Iris Recognition

Emine Krichen, Lorène Allano, Sonia Garcia-Salicetti, and Bernadette Dorizzi

Institut National des Télécommunications, 91011 Evry, France
{emine.krichen,lorene.allano,sonia.salicetti,bernadette.dorizzi}
@int-evry.fr

Abstract. In this paper, we present a new method for iris recognition based on specific texture analysis. It relies on the use of Haar wavelet analysis to extract the texture of the iris tissue. The matching is done using a specific correlation based on local peaks detection. Experiments have been conducted on the CASIA database in verification mode and show an EER of 0.07%. Degraded version of the CASIA database results in an EER of 2.3%, which is lower than result obtained by the classical wavelet demodulation (WD) method in that database.

1 Introduction

Iris scan is a sufficiently mature biometrics to be used for identification purposes. John Daugman was the pioneer in iris recognition area. His method relies on the use of Gabor wavelets in order to process the image at several resolution levels. An iriscode composed of binary vectors is this way computed and a statistical matcher (logical exclusive OR operator) analyses basically the average Hamming distance between two codes (bit to bit test agreement) [1]. Besides, several methods based on Gabor filters have been proposed using a circular filter [2] or the zero cross representation [3]. Global aspect analysis was also used for iris recognition relying on Independent component analysis [4], while some recent methods are based on the extraction of key points [6] or local extreme points [5]. The system proposed by J.Daugman has been tested using a database of thousand people; some of the tests have been made by independent institutes [8]. Moreover, the system is tested every day in Emirates borders and very low error rates were observed in all these experimentations. To reach low error rates, strong constraints must be imposed to the users during the acquisition process to reduce false rejection rates (false acceptance rates are not affected by variation in image quality). People can accept these conditions in airport or in secured building accesses but not for daily use. It is interesting to try to release these constraints and to build a system operating on degraded quality images. Of course, we do not expect such system to obtain the accuracy level of the Daugman one, but to provide however, acceptable results, in verification mode. In this paper we propose a method based on an original texture analysis or texture correlation (TC), using Haar first level wavelet decomposition, and a normalized correlation process based on peak detection. In the framework of iris correlation, Wildes [9] proposes a method based on the correlation of small windows of the iris image at several levels of resolution. Kumar et al [10] have also used specific designed filters to perform correlation in the Fourier domain. Our method has the originality to perform correlation after previous extraction of iris texture. Working on texture images, using a correlation process will

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 23–30, 2005.
© Springer-Verlag Berlin Heidelberg 2005

allow us to deal in a better way with rotation, dilation or contraction of the pupil and also with blurred images. Experiments have been conducted using the CASIA database collected by the Institute of Automation of the Chinese Academy of Science. In this database, the images were acquired with a monochrome camera using NIR illumination. We have tested our method in verification mode. We have also performed experimentations using a degraded version of the CASIA database in which the images were blurred by Gaussian filter with fixed size. The content of this paper is as follows. In the next section, we will describe the texture analysis process, including iris segmentation, normalisation, enhancement and texture extraction. Section 3 presents our matching process and the results of the different experimentations are given in section 4. Finally, the conclusions and perspectives are provided in section 4.

2 Texture Analyses

2.1 Preprocessing

We use a hybrid method for iris segmentation using the Hough transform and integrodifferential operators as described in [7]. In the present work the iris normalisation is performed using a pseudo polar transformation [1]. We also detect eyelids and eyelashes in the original iris image, in order to isolate only the iris tissue pixels for the next processings (texture extraction and matching). Our eyelids and eyelashes detection is based on an histogram analysis. The histogram of the iris rim is represented in figure 1. We suppose that the iris distribution is Gaussian and we find the parameters of this Gaussian (mean, standard deviation) by eliminating iteratively the pixels of eyelids and eyelashes based on the fact that the iris tissue pixels are darker than the eyelids and spot reflections pixels, and whiter than the eyelashes pixels. Figure1 also shows the final position of eyelashes and eyelids thresholds. Figure 2 shows all the preprocessing process including iris segmentation, normalization and enhancement.

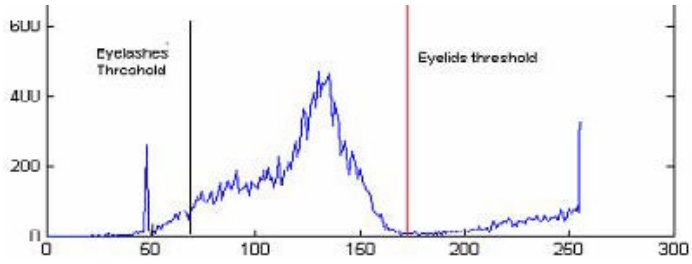


Fig. 1. The iris rim histogram and the eyelashes and eyelids thresholds

2.2 Texture Extraction

In order to extract the texture from the background, we have used the Haar wavelet decomposition [12]. At the first wavelet decomposition order, we obtain four images, each one having a size equals to one quarter of the original image, usually named gg (low frequencies in row and column), hh (high frequencies in row and column), hg (high frequencies in row and low frequencies in column) and gh (low frequencies in row and high frequencies in column), as shown in figure in figure 3.

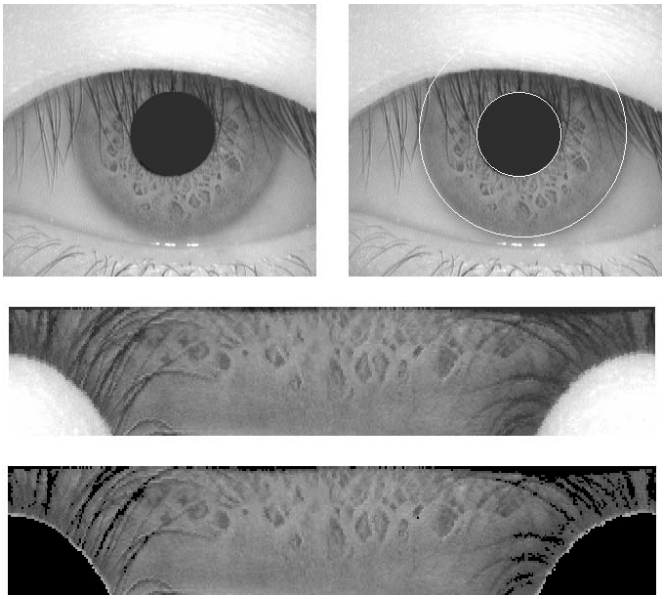


Fig. 2. The preprocessing steps including: The original iris image (top Left), the segmented image (top right), the normalized image at 64*360 pixels size (middle) and the enhanced image after detection of eyelids and eyelashes (Bottom)

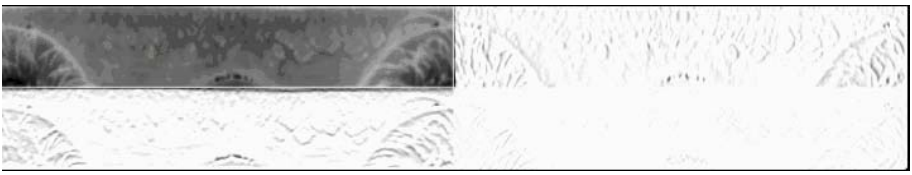


Fig. 3. The wavelet decomposition process: gg image (Top left), hg image (top right), gh image (bottom left), hh image (bottom right)

The texture information is contained in the high frequencies while the low frequencies pick up only the general aspect of the image with a quarter of the original resolution. To extract the iris texture we will thus not use gg images. In our initial approach we wanted to use hh images because they contain the high frequencies in both rows and columns. But we have noticed that the hh image contains a lot of noise. So, we focused only on the hg and the gh images. Because these two kinds of images are complementary (hg images contain the vertical texture information and gh images contain the horizontal texture information) we generate a new image from hg and gh images computing at each pixel the maximum grey level value between hg and gh (see equation 1). We finally resize the image at this new initial iris image using a linear interpolation.

$$I(x, y) = \text{maximum}(hg(x, y), gh(x, y)) \quad (1)$$

The iris texture image is illustrated in figure 4.



Fig. 4. Iris texture image by Haar wavelet decomposition

3 Matching Process

We will focus on the texture image obtained after the wavelet filtering as discussed in section 2. Our aim is to compute a similarity parameter between two iris texture images: an iris texture reference image Iref and an iris texture test image Itest. A rough correlation algorithm on the global image cannot be applied because of important intra class variability. This corresponds to rotation, dilation and contraction of the pupil which will results in fluctuation on the global texture image. Also the richness of the iris texture imposes us to work on little size images. The Itest image is divided into a set of 15*15 pixels size templates. Our approach is based on the matching of each template of Itest within Iref image. In the following we will explain our matching process on one template of Itest; generalization to other templates is straightforward. Given a template T1 from the set of templates extracted from Itest, we will try to find the position into Iref which is the more correlated to T1 and will associate to this position a correlation parameter. Instead of researching T1 in the whole Iref image, we will only take into account a template T2 with bigger size than T1 in order to deal with rotation, pupil contraction and dilation problems. The centre of T2 has the same coordinates in Iref the ones of as T1 in Iref. Namely T2 has a 21*33 pixels size, so we can deal with a pupil contraction and dilation of +- 3 pixels and a rotation of +/- 9° (see figure 5). Then we perform the correlation between T1 and T2 by measuring the matrix of correlation Cd = T1⊗T2.

If Itest and Iref come from the same person we will observe a peak of correlation into the Cd matrix; No such peak won't be normally observed when computing the correlation between images provided from different clients. We have used for peak detection the Peak to Side Lob Ratio [10] PSR as calculated in equation 2

$$PSR\ (T1,T2) = \frac{(\ Maximum(Cd)\ -\ mean(Cd)\ )}{Std(Cd)} \tag{2}$$

The richness of the iris textures varies from one person to another, and it's reflected in PSR. In order to compare irises from different person, we need to perform a normalisation of the PSR before to take the decision (Client/Impostors). To normalize our similarity parameters between images with high iris textures and images with few iris textures, we compute PSR (T1,Tint) for each T1 with Tint having the same size than T2 but extracted from the same image than T1. Finally we define the difference between the two PSR's, PSR (T1,T2) and PSR (T1,Tint) as the similarity parameter between T1 and T2. We also take into account Pt, the relative position of the peak of correlation into Cd. Pt is thus calculated by taking the origin at the centre of T1 (which is also the centre of T2). For each Itest we obtain a set of PSR's and Pt's, corresponding to the different sub-images extracted from Itest. If the comparison is done between two images provided by the same client, the Pt's (position of the

maximum) should normally be the same, or at least comparable. The mean of the Pt's can give the rotation parameter between the two images. We recall that in the chosen reference axes, rotations are transformed into translations. So, we will only take into account the PSR's in which the corresponding Pt is comparable to the mean of the Pt's, and also with T1 doesn't contain more than 10% of its pixels provided by eyelids and eyelashes detected in enhancement process to reduce intra class variability (see figure 6).

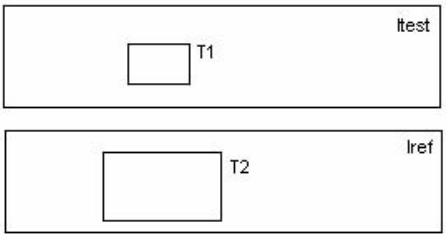


Fig. 5. An example of the extracted 15*15 template T1 and its corresponding 21*32 template T2 extracted from Iref

Then we compare each component of the extracted sub-set to a predefined threshold. We finally obtain the similarity measure (SM) parameter between Itest and Iref as follow:

$$SM\ (Iref, Itest) = \frac{Numbe\ of\ PSR\ superior\ to\ threshol}{Total\ PSR\ numbe} \tag{3}$$

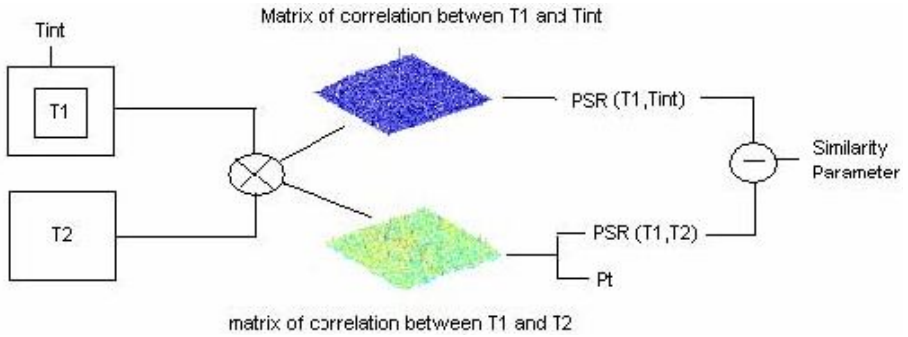


Fig. 6. The correlation process between one extracted sub-images T1 from Itest and its corresponding template T2 extracted from Iref

4 Experimentations

4.1 Experimentation on CASIA Database

We have tested texture correlation (TC) method, which the complete system is described in figure 7. The database comes from the National Laboratory of Pattern Recognition (NLRP) in China that is the CASIA Iris Image Database [12] collected by the Institute of Automation of the Chinese Academy of Science, in which the images

were acquired with a monochrome camera using NIR illumination. It is composed of 749 images from 107 different irises. For each iris, 7 images were captured in 2 sessions. The first session provides reference images (3 images) and the second session provides the test images, an each client.

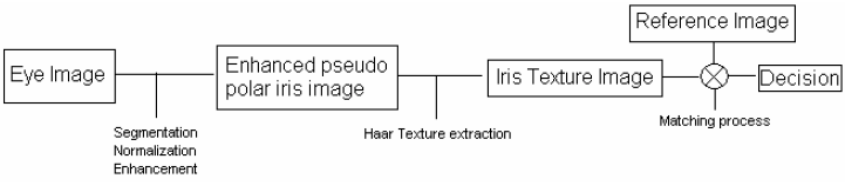


Fig. 7. The steps of our iris recognition system based on specific texture analysis, including iris segmentation, normalisation and enhancement (Step1), the texture extraction using Haar decomposition (Step 2) and the matching process to take the final decision

For each client and for each test images, we keep the minimum value of its similarity measure to the three references images of the client. This measure, when associated to a threshold gives two values FAR and FRR. Varying the threshold leads to the corresponding DET curves of the systems Figure 8. The EER of the system is 0.07%. This result is comparable (but quietly higher) to the one obtained in [13] with the WD.

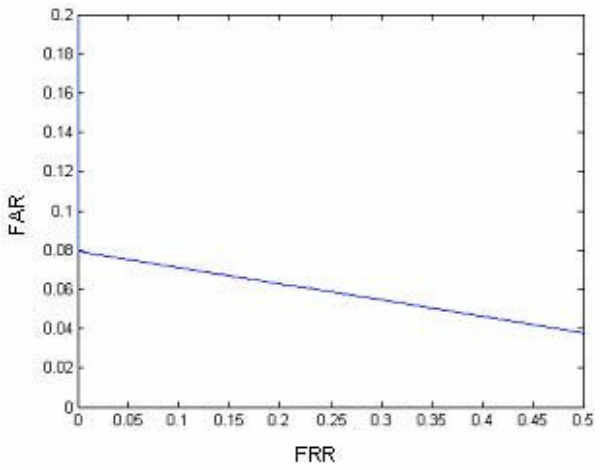


Fig. 8. DET curve using texture correlation algorithm on Casia database

4.2 Experimentation on Degraded CASIA Database

In our second experimentation, we have tested the ability of our method to deal with blurred images. In real applications, this kind of images can be obtained if the subject is not at Wright position from the camera lens or if there are some illumination effects. We have blurred the images through a convolution with a 21*21 Gaussian filter. Figure 9 illustrates the difference between an original Casia iris image and its corresponding blurred image. The DET curves of the two methods are shown in figure 10.

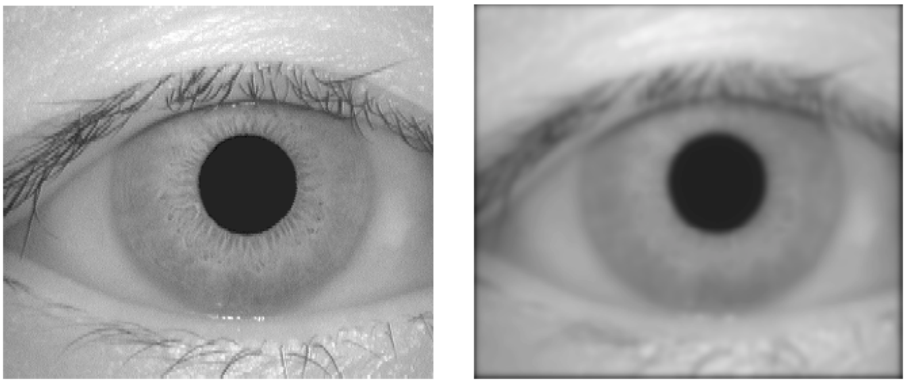


Fig. 9. Original iris image (left) and its blurred image (right) In this experimentation we have compared our method with the reference WD method inspired by Daugman works

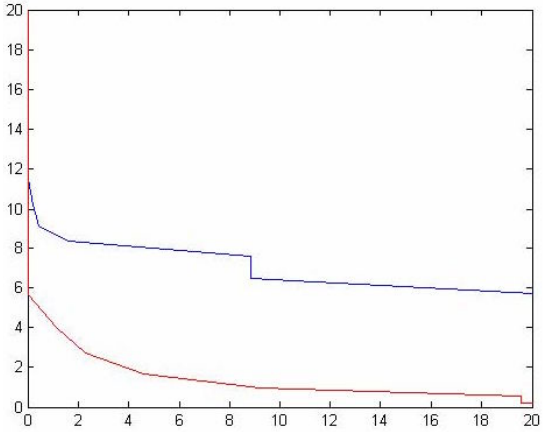


Fig. 10. DET curve for texture correlation (RED) and for wavelet demodulation (Bleue) on degraded CASIA database

We remark that the DET curve with TC method is roughly three times better than the WD method, which shows that TC method resists better to blurring. Please note that no particular optimization was made to code the WD method.

5 Conclusions

In this paper, we have presented a new method for iris recognition which relies on a specific texture analysis performed using a Haar wavelet transform while matching is done through local correlations within the different resulting images. Our aim is to provide an alternative to other methods relying on wavelet demodulation approach [1], which shows very good results, in identification on large scale database of the price of constrained acquisition which can be difficult to accept by the users. Of course releasing these constraints will lead to degradation in the quality of the images, and will introduce blurring, rotation, more pupil dilation and contraction Our

approach has been designed to cope with those kinds of distortions. Extraction of the iris texture information through Haar transformation has the effect of smoothing intra class differences. Matching is performed using local correlation on sub images extracted from the global one. A normalisation is also performed to again favour intra class stability. This way, we expect our system to provide better results in verification mode, when recording conditions are degraded and this exactly what we observe on our experiments realized on the CASIA database after performing some blurring on the original images. Our method allows to reduce by a factor of three the error rates of the conventional approach (WD).

This is very promising results which must be confirmed on a bigger size database and with other experiments conditions in particular combination of blurring and rotational effects.

References

1. John Daugman, "How iris recognition works", IEEE Transactions on Circuits and Systems fo Video Technology, VOl. 14, No. 1, January 2004.
2. Li Ma, Yunhong Wang, Tieniu Tan, " Iris recognition using circular symmetric filters", Proceedings of the 16th International Conference on Pattern Recognition, Volume 2, pp. 414 -417, 11-15 August 2002.
3. W.W. Boles and B. Boashash, "A Human Identification Technique Using Images of the Iris and Wavelet Transform", IEEE Trans. Signal Processing, Vol.46, No.4, 1998, pp.1185-1188.
4. Ya-Ping Huang, Si-Wei Luo, En- Yi Chen, "An efficient iris recognition system", Proceedings of the First Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002.
5. Jiali Cui, Yunhong Wang, Tieniu Tan, Li Ma, and Zhenan Sun, "An Iris Recognition Algorithm Using Local Extreme Points", Daviid Zhang and Anil K Jain (Eds.) LNCS 3072, pp.442-450, 2004
6. Wen Yang, Li Yu, Guangming Lu, and Kuanquan Wang, "Iris Recognition Based on Location of key points", Daviid Zhang and Anil K Jain (Eds.) LNCS 3072, pp.484-490, 2004
7. Emine Krichen, M.Anouar Mellakh, Sonia Garcia-Salicetti, Bernadette Dorizzi "Iris identification using wavelet packets", Pattern Recognition, 2004 Proceedings of the 17th International Conference on, Volume: 4, Aug. 23-26, 2004 Pages:335 " 338.
8. http://www.cl.cam.ac.uk/users/jgd1000/
9. R.P. Wildes, "Iris recognition: an emerging biometric technology", Proceedings of the IEEE, Volume 85, Issue 9, pp. 1348 -1363, September 1997.
10. B.V.K Vijaya Kumar, Chunyan Xie, Jason Thornton, "Iris Verification Using Correlation Filters", J. Kittler and M.S. Nixon (Eds.), AVBPA 2003, LNCS 2688, pp. 697-705, 2003.
11. http://www.sinobiometrics.com
12. A. Jensen, A. La Cour-Harbo "Ripplies in Mathematics The discrete Wavelet Transform", Springer 2001/
13. Zhenan Sun, Yunhong Wang, Tieniu Tan, Jiali Cui "Robust Direction Estimation of Gradient Vector Field for Iris Recognition", Pattern Recognition, 2004 Proceedings of the 17th International Conference on 2004.

A Study on Iris Image Restoration

Byung Jun Kang[1] and Kang Ryoung Park[2]

[1] Dept. of Computer Science, Sangmyung University,
7 Hongji-Dong, Jongro-ku, Seoul, Republic of Korea, Biometrics Engineering Research Center
9737001@smu.ac.kr
[2] Division of Media Technology, Sangmyung University,
7 Hongji-Dong, Jongro-ku, Seoul, Republic of Korea, Biometrics Engineering Research Center
parkgr@smu.ac.kr

Abstract. Because iris recognition uses the unique patterns of the human iris, it is essential to acquire the iris images at high quality for accurate recognition. Defocusing reduces the quality of the iris image and the performance of iris recognition, consequently. In order to acquire a focused iris image at high quality, an iris recognition camera must control the focal length of the moving lens. However, that causes the cost and size of iris camera to be increased and that needs complicated auto-focusing algorithm, also. To overcome such problems, we propose new method of iris image restoration. Experimental results show that the total recognition time is reduced as much as 390ms on average with the proposed restoration algorithm.

1 Introduction

The iris recognition uses the unique patterns of the human iris to recognize an individual with confidence[1][2]. For iris recognition, it is essential to acquire iris images at high quality to allow for accurate recognition. If a blurred iris image is acquired, the performance of the iris recognition is degraded. Optical defocusing is one of many factors which may make blurred iris images.

The acquisition of clear iris images is made difficult by the optical magnification requirements; restrictions on illumination, and the target motion, distance and size[3]. All of these factors reduce the possible depth of field of the optics, because they require a lower F number to accommodate both the shorter integration time and the light dilution associated with long focal length[3]. Iris recognition systems which use fixed-focus optical lens almost always produce defocused iris images. In order to acquire a focused iris image with high quality, the iris recognition camera controls the focal length of the moving lens with auto-focusing algorithm or selects the best focused iris image from frames in a video sequence[3]. However, the former causes the cost, size and complexity of iris camera to be increased. In addition, the latter gives in convenience to users, because the users have to move their heads back and forth with respect to the camera in order to acquire a good focused iris image. To overcome such problems, iris image restoration method was introduced by *J. van der Gracht*[4]. However, that needs additional hardware for cubic phase mask. In addition, that uses the Wiener filter for the iris image restoration. The Wiener filter does not only take much computation time, but also it is difficult that both $\Phi_L(u, v)$ (the normalized power spectrum of iris region) and $\Phi_N(u,v)$ (the normalized noise power spectrum) in Wiener filter are determined with accuracy in the case of iris, because iris patterns are

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 31–40, 2005.
© Springer-Verlag Berlin Heidelberg 2005

very random and various. In addition, their method does not consider the degree of blurring according to focus score. So, we propose a new iris image restoration method based on the assessed focus score. Consequently, the total recognition time is reduced as much as 390ms on average in the experiments.

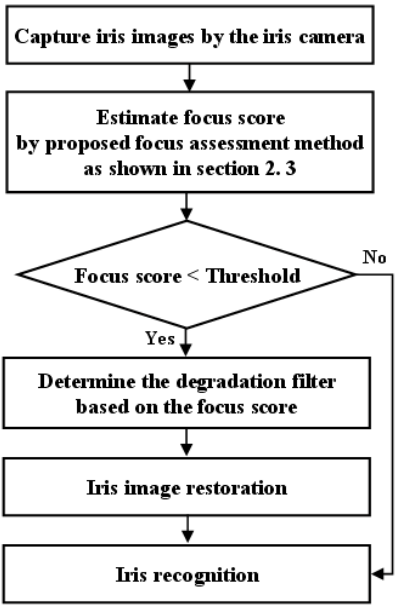


Fig. 1. The overview of the proposed algorithm

2 The Proposed Iris Image Restoration Method

2.1 The Overview of the Proposed Algorithm

The overview of the proposed algorithm is shown in Fig. 1. After capturing iris image, we estimate focus scores by the proposed focus assessment. If the measured focus score is lower than threshold, the acquired iris image is restored. In this case, we determine the degradation filter selecting the filter size based on the focus score. After that, we perform the iris image restoration with the determined filter, and the restored iris image is used for recognition. On the other side, if the focus score is higher than threshold (we used 80 as threshold.), it is used for recognition without iris image restoration.

2.2 Previous Researches on Focus Assessment

In previous research on focus assessment, it adopted the focusing method used for general scene (landscape or photographic scene)[5-12].

In general, defocused images contain more low frequency components than focused images in the frequency domain[8]. Therefore we can estimate degree of focus by measuring the high frequency components. Previous research uses the gradient

value of processed edge image as the focus score and this method is reported as Tenengrad method[9]. *Javis* uses the SMD (Sum Modulus Difference) for checking focus score[6]. *Nayer* adopts the SML (Sum Modified Laplacian) and the absolute values of 2nd order derivative (laplacian) are used as focus score in their method[7]. However, such methods are mainly target for the general scene, and they can generate the wrong focus score in case of iris image. Especially, in case of users with glasses, if the lens is positioned for focusing the scratched glasses surface or the glasses frame, such cases may make their focusing scores highest. To overcome such problems, *J.Daugman* proposes the (8 x 8) pixels sized convolution kernel for checking the focus score of iris image[3].

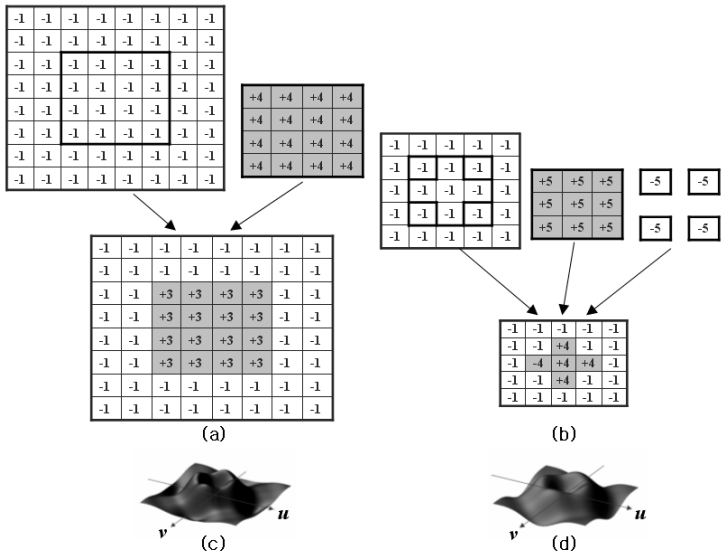


Fig. 2. (a) The (8 x 8) convolution kernel of *J.Daugman* for 2-D focus assessment. (b) The proposed (5 x 5) convolution kernel for 2-D focus assessment. (c) The Fourier spectrum of the (8 x 8) convolution kernel. (d) The Fourier spectrum of the proposed (5 x 5) convolution kernel

It consists of two square box functions, one of size (8 x 8) pixels and amplitude of -1, and the other one of size (4 x 4) pixels and amplitude of +4 as shown in Fig. 2(a). Because they are overlapped, the central region is summed to +3 [3]. So the 2-D Fourier transform of the overlapped convolution kernel composing of two square box functions is represented as Eq. (1), which consists of two 2-D "sinc" functions[3].

$$K(u,v) = \frac{\sin(u)\sin(v)}{\pi^2 uv} - \frac{\sin(2u)\sin(2v)}{4\pi^2 uv} \tag{1}$$

The *K(u,v)* in the 2-D Fourier domain is plotted in Fig. 2 (c). The total power in that band is the spectral measurement of focus. Finally, this summated 2-D spectral power is passed through a compressive nonlinearity of the form: $f(x) = 100x^2/(x^2+c^2)$, in order to generate a normalized focus score in the range of 0 to 100 for any image[3]. Here *x* is the total power spectrum measured by the (8 x 8) convolution kernel as shown in Fig. 2(a) and c is constant value (we use 600 as c).

2.3 Proposed Kernel for Fast Focus Assessment

The method by *J.Daugman* does not well grasp high frequency bands about the fine texture of iris images, and takes 15ms(in 300-MHz RISC processor) due to the large sized kernel. Because the focus assessment and iris image restoration should be finished within 33ms in our case for real-time processing, minimizing focus assessment time is important. In order to solve such problems, we propose (5 x 5) pixels sized convolution kernel as shown in Fig. 2(b). It consists of three square box functions, one of size (5 x 5) and amplitude of -1, one of size (3 x 3) and amplitude of +5, and four of size (1 x 1) and amplitude of -5. Because they are overlapped, the (5 x 5) convolution kernel looks like Fig. 2(b). So the 2-D Fourier transform of the convolution kernel composing of three square box function is represented as Eq. (2).

$$K'(u,v) = \frac{\sin(\frac{3}{2}u)\sin(\frac{3}{2}v)}{\frac{9}{4}\pi^2 uv} - \frac{\sin(\frac{5}{2}u)\sin(\frac{5}{2}v)}{\frac{25}{4}\pi^2 uv} - 4 \cdot \frac{\sin(\frac{1}{2}u)\sin(\frac{1}{2}v)}{\frac{1}{4}\pi^2 uv} \quad (2)$$

The $K'(u,v)$ in the 2-D Fourier domain is plotted in Fig. 2 (d). As in *J.Daugman'*s method, this summated 2-D spectral power is passed through a compressive nonlinearity of the form: $f(x) = 100x^2/(x^2+c^2)$, which can make a normalized focus score generated in the range of 0 to 100 for any image [3]. Here, x is the total power spectrum measured by the (5 x 5) pixels convolution kernel.

After the Fourier transform is performed, the central part in the frequency domain is the low frequency band. The high frequency band is located far from the central part in the frequency domain[13]. Both the (8 x 8) pixels convolution kernel and the (5 x 5) pixels convolution kernel use a high-pass filter with a similar shape. However, the (5 x 5) pixels convolution kernel contains more high frequency bands than the (8 x 8) pixels convolution kernel as shown in Fig. 3. With the (8 x 8) pixels convolution kernel, the calculated area of the high frequency band is almost 0.2144. And with the (5 x 5) pixels convolution kernel, the area of the high frequency band is almost 0.6076 as shown in Fig 3(b). From that, we can know our proposed kernel can detect fine and high frequency of iris texture much better than *Daugman*'s kernel.

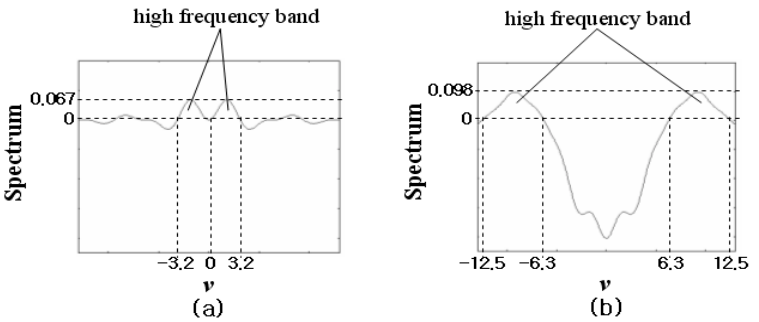


Fig. 3. When u=0.01 in Eq. (1) and Eq. (2), (a) the Fourier spectrum of the (8 x 8) convolution kernel. (b) the Fourier spectrum of the (5 x 5) convolution kernel

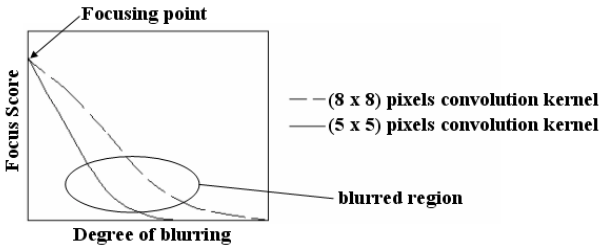


Fig. 4. When c=600 in compressive nonlinearity of the form: $f(x) = 100 \cdot x^2/(x^2+c^2)$, focus score vs. degree of blurring with the (8 x 8) convolution kernel and the (5 x 5) convolution kernel

With the (8 x 8) pixels convolution kernel, the convolution value is calculated per every fourth row and fourth column in the iris image[3]. Therefore the total multiplication count is 1,193,216(=8x 8x 158(the number of kernel movement steps in the X direction) x 118(the number of kernel movement steps in the Y direction)) in the image of size (640 x 480). With the proposed (5 x 5) pixels convolution kernel, the convolution value is calculated per every third row and third column. Therefore the total multiplication count is 842,700(=5x 5x 212(the number of kernel movement steps in the X direction) x 159(the number of kernel movement steps in the Y direction)) in the image of size (640 x 480). So we can know the proposed (5 x 5) pixels convolution kernel takes less execution time than the (8 x 8) pixels convolution kernel.

With iris images from the CASIA database [14], we have measured the focus score as the degree of blurring as shown Fig. 4. In general, if the curve has the shape that the slope near a focusing point and that in the blurred region are maintained to be steep, it is reported that the focusing algorithm shows good performance[8]. That is because in case that the slope is steep, the focus lens can reach the focused position fast and accurately. In addition, in case that the slope in the blurred region is also steep, the focus lens can determine its movement direction easily[8]. The proposed (5 x 5) pixels convolution kernel is more steep than the (8 x 8) pixels convolution kernel as shown in Fig. 4 and we can know the focusing performance of our method is better than *Daugman*'s one.

2.4 Restoration of the Defocused Iris Image

In the spatial domain, defocused image is represented as Eq. (3).

$$o(x,y) = h(x,y) * i(x,y) + n(x,y), \tag{3}$$

where $o(x,y)$ is blurred image by defocusing, $h(x,y)$ is 2-D point-spread function which causes blurring, $i(x,y)$ is clear (focused) image, and $n(x,y)$ is noise [13]. If it is transformed into the frequency domain, the defocused image is represented as Eq. (4).

$$O(u,v) = H(u,v) \bullet I(u,v) + N(u,v), \tag{4}$$

where $O(u,v)$ is the Fourier transform of the blurred iris image by defocusing, $H(u,v)$ is that of the degradation function (2-D point-spread function) which causes blurring,

I(u,v) is that of the clear (focused) image, and *N(u,v)* is that of noise. In our experiments, we do not consider *N(u,v)* by reducing it by (3 x 3) sized Gaussian filter, because *N(u,v)* is very smaller than *H(u,v)*. Convolution in the spatial domain is corresponding to multiplication in the frequency domain[13]. So, in the frequency domain, the clear image(*I(u,v)*) which is acquired from the defocused image is represented as Eq. (5).

$$I(u,v) = \begin{cases} \dfrac{O(u,v)}{H(u,v)} & H(u,v) \neq 0 \\ \dfrac{O(u,v)}{H(u,v)+c} & H(u,v) = 0,\ c \neq 0 \end{cases}, \tag{5}$$

where *c* is the constant for solving the zero-crossing problem.

In general, it is reported that the degradation function(*H(u,v)*) can be selected by empirical observation for the blurred image[15][16]. This study defines that the degradation function modeled such as Eq. (6) by observing the defocused images by an iris camera. In detail, we gathered several image of a bright point at the best focus and defocused position to provide an estimate of the degradation function.

$$H(u,v) = \begin{cases} Ae^{\left(\frac{u+v}{\sigma}\right)} & u < 0,\ v < 0 \\ Ae^{\left(\frac{u-v}{\sigma}\right)} & u < 0,\ v \geq 0 \\ Ae^{-\left(\frac{u-v}{\sigma}\right)} & u \geq 0,\ v < 0 \\ Ae^{-\left(\frac{u+v}{\sigma}\right)} & u \geq 0,\ v \geq 0 \end{cases} \tag{6}$$

As the degree of defocus is increased, pixels are more blurred by the degradation function. Therefore we determine the σ of the degradation filter in Eq. (6) according to the focus score which is measured by the (5 x 5) pixels convolution kernel as shown in Fig. 2 (b). For example, if the focus score is smaller, the degradation filter uses a bigger valued σ and vice versa. Here, A is constant value (we use 1 as A).

As shown in Eq. (5), we use the inverse filter, but previous research[4] does the Wiener filter. In generally, the Wiener filter has better performance than the inverse filter. However, because the Wiener filter takes much computation time, we use the inverse filter in order to process at real-time. Also it is difficult that both $\Phi_L(u, v)$(the normalized power spectrum of iris region) and $\Phi_N(u,v)$(the normalized noise power spectrum) in the Wiener filter are determined with accuracy due to various iris patterns signal.

3 Experimental Results

We have tested our iris image restoration algorithm with iris images from the CASIA Database [14]. The CASIA database has 756 iris images with size of (340 x 280) pixels from 108 eyes of 80 subjects.

In the first experiment, we have produced artificially blurred iris images from focused original iris images by Gaussian mask with radius of 2.5 pixels. We also produced restored iris images from blurred iris images by using the iris image restoration as shown in Fig. 5 (c). Dark dots occurred on the corneal specular reflection as over-

flow which happened while the iris image was being restored from the blurred image. The overflow occurred in the DC components (when *H(u,v)* is very low and *O(u,v)* is very high in Eq. (5)). In order to solve such problem, we reduced the DC components of the blurred image. After that, we increase the average brightness of the restored iris image as much as the average brightness of the blurred iris image. So, iris images are acquired as shown in Fig. 5(d).

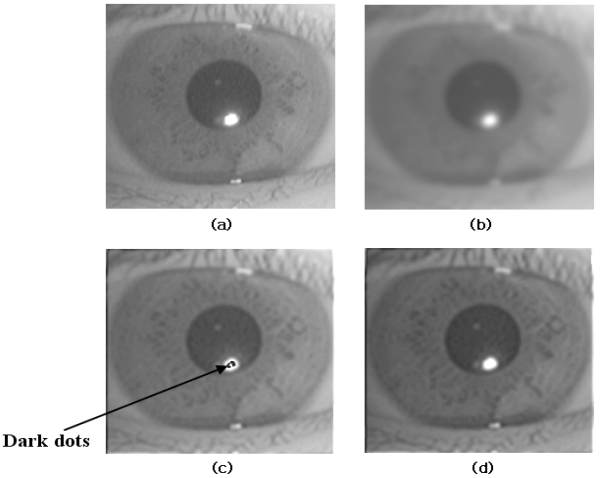


Fig. 5. (a) The focused original iris image. (b) The artificially blurred iris image by Gaussian filter with radius of 2.5 pixels. (c) The restored iris image from the defocused iris image. (d) The restored iris image after reducing the DC components of the blurred image(*O(u,v)*) and compensating the average brightness of the restored image(*I(u,v)*) as much as that of the blurred image(*O(u,v)*)

In order to evaluate the perform of our restoration method, we measured the pixel RMS error like Eq. (7).

$$RMS\ error = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\left|o(x,y)-\hat{o}(x,y)\right|, \tag{7}$$

where *o(x,y)* is the gray level of focused original images, and $\hat{o}(x,y)$ is that of blurred images or the restored images. The RMS error between blurred images and focused original ones was 3.56 on average, and that between restored images and focused original ones was 3.25 on average. The restored iris images showed lower RMS error than the blurred iris image.

In the second experiment, we have tested the recognition performance of iris images with our iris image restoration algorithm. We had 648 authentic tests and 69,338 imposter tests from CASIA database. First, we enrolled the focused original iris image in the iris recognition system with the Gabor filter having the frequency of $\pi/16$ and $\pi/8$ [3]. And as shown in Table 1, we computed the Hamming distance between the enrolled focused iris image and the blurred one or the restored one. The Hamming distance between the enrolled focused images and the blurred ones was 0.312 on average. And the Hamming distance between the enrolled focused images and the restored ones was 0.073 on average. The restored iris images show lower Hamming

distance than the blurred iris images. Also, some blurred iris images show the FRR(false reject error), whereas the restored iris images do not. From that, we can know that the restored iris images contain iris pattern information that are almost same to that of the original iris images. The FAR(false accept Error) cases do not happen in any case.

In the third experiment, we have tested the total recognition time with the BM-ET100 made by Panasonic[17] according to user's initial Z distance. We measured the total recognition time of the total 30 persons (each person tries to recognize 5 times) according to Z distance as shown in Table 2. The depth of field is increased with the proposed iris image restoration algorithm. Consequently, the original operating range of the BM-ET100 is 48-53cm, but in case of using our iris image restoration algorithm it becomes 46-56cm. The normal approaching speed of users is 5cm/sec ± 2 in order to locate the eye in the operating range, which is measured by position sensing device[18]. As shown in Table. 2, the recognition time with our iris image restoration algorithm is 1.324 sec on average and it is reduced as much as 390 ms compared to that without restoration algorithm.

Table 1. The examples of Hamming distance between the focused original iris image and the artificially blurred iris image or the restored iris image (threshold for recognition is 0.34)

Index	Between focused original image and the blurred image		Between focused original image and restored iris image	
	HD	Result	HD	Result
1	0.35232	False Reject	0.11032	True Accept
2	0.28324	True Accept	0.07732	True Accept
3	0.35932	False Reject	0.09821	True Accept
4	0.25352	True Accept	0.00771	True Accept

In the fourth experiment, we have measured the process time of our algorithm. The execution time of the focus assessment was almost 0ms on a 2.4-GHz Intel Pentium 4 processor, and the execution time of iris image restoration was 90 ms.

Table 2. Initial Z distance vs. the total recognition time

Initial Z distance (cm)	Recognition time of BM-ET100 without iris image restoration (sec)	Recognition time of BM-ET100 with iris image restoration (sec)
40	1.875	1.562
46	0.673	0.363
56	0.874	0.363
60	1.672	1.165
70	3.474	3.167
Average	1.714	1.324

In the last experiment, we have tested the recognition time and rate according to environmental lighting condition as shown in Table 3. The total recognition time and recognition rate are almost same irrespective of the change of environmental lighting condition with fluorescent lamp. That is because the BM-ET100 has the IR pass filter and the functionality of AE (Auto Exposure). To be notable, in case of the light condition below 500 Lux., the FRR and the recognition time is increased a little. That is

because the pupil is dilated too much due to dark environmental light and it can cause the increase of FRR, consequently. In this case, the recognition time is sum of focus assessment time, acquisition time of iris images, iris image restoration time, and recognition time.

Table 3. Lighting condition vs. the recognition time and recognition rate

Item \ Environmental Lighting condition(Lux.)		250	500	750	1000	1500
Recognition time (sec)	BM-ET100	1.76	1.75	1.72	1.72	1.71
	BM-ET100 + Proposed algorithm	1.35	1.35	1.32	1.32	1.32
Recognition rate (FAR) (%)	BM-ET100	0	0	0	0	0
	BM-ET100 + Proposed algorithm	0	0	0	0	0
Recognition rate Without glasses (FRR) (%)	BM-ET100	0.5	0.4	0.2	0.2	0.2
	BM-ET100 + Proposed algorithm	0.3	0.2	0.1	0.1	0.1
Recognition rate With glasses (FRR) (%)	BM-ET100	1.8	1.6	1.4	1.4	1.4
	BM-ET100 + Proposed algorithm	1.0	0.9	0.8	0.8	0.8

4 Conclusion

We have proposed iris image restoration which can overcome the limitation of the depth of field of the optics. In the experimental result, the total recognition time is reduced as much as 390ms on average with proposed image restoration algorithm and the time for restoration is about 90ms. However, it will be much reduced not generating the degradation filters dynamically, but defining them in advance. If an iris image includes hair or eyelash, proposed focus assessment method has the problem that wrong focus score is calculated. In addition, our image restoration method based on focus score may give poor results when comparing iris patterns that inherently have either high contrast or low contrast. To overcome such problem, we plan to research the method of using corneal specular reflection to measure the focus score. This paper supposed the 2-D spread function invariant in an iris image for image restoration. However, it is often the case that 2-D spread function is not invariant even in an image. Therefore we need study using variant 2-D spread function in future works.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

References

1. John G. Daugman, "High confidence visual recognition of personals by a test of statistical independence," IEEE Trans. Pattern Anal. Machine Intell., vol.15, no.11 (1993) 1148-1160
2. Anil K. Jain, "Biometrics: Personal Identification in Networked Society," Kluwer academic publishers (1998)
3. John G. Daugman, "How Iris Recognition Works," IEEE Trans. on Circuits and Systems for Video Technology, vol. 14, no. 1 (2004) 21-30

4. J. van der Gracht, V. P. Pauca, H. Setty, R. Narayanswamy, R. J. Plemmons, S. Prasad, and T. Torgersen, "Iris recognition with enhanced depth-of-field image acquisition," Proceedings of SPIE, vol. 5438 (2004) 120-129
5. Je-Ho Lee, Kun-Sop Kim, Byung-Deok Nam, Jae-Chon Lee, Yong-Moo Kwon and Hyoung-Gon Kim, "Implementation of a passive automatic focusing algorithm for digital still carmera," IEEE Transactions on Consumer Electronics, vol.41, no.3 (1995) 449-454
6. R. A. Javis, "Focus Optimization Criteria for Computer Image Processing," Microscope, vol. 24(2), 163-180
7. S. K. Nayar and Y. Nakagawa, "Shape from Focus," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.16, no.8 (1994) 824-831
8. Kang-Sun Choi, Jun-Suk Lee and Sung-Jae Ko, "New Auto-focusing Technique Using the Frequency Selective Weight Median Filter for Video Cameras," IEEE Trans. on Consumer Electronics, vol.45, no.3 (1999) 820-827
9. J. M. Tenenbaum, "Accommodation in computer vision," Ph. D. thesis, Stanford University (1970)
10. T. Haruki and K. Kikuchi, "Video Camera System Using Fuzzy Logic," IEEE Transactions on Consumer Electronics, vol.38, no.3 (1992) 624-634
11. K. Ooi, K. Izumi, M. Nozaki and I. Takeda, "An Advanced Auto-focusing System for Video Camera Using Quasi Condition Reasoning," IEEE Transactions on Consumer Electronics, vol.36, no.3 (1990) 526-529
12. K. Hanma, M. Masuda, H. Nabeyama and Y. Saito, "Novel Technologies for Automatic Focusing and White Balancing of Solid State Color Video Camera," IEEE Transactions on Consumer Electronics, vol.CE-29, no.3 (1983) 376-381
13. R. C. Gonzalez, R. E. Woods, "Digital Image Processing 2nd Edition," Prentice Hall (2002)
14. *http://www.sinobiometrics.com*
15. D. Kundur and D. Hatzinakos, "Blind image deconvolution," IEEE Signal Processing Magazine, vol.13, (1996) 43-64
16. Andreas E. Savakis, "Blur Identification by Residual Spectral Matching," IEEE Transactions on Image Processing, vol.2, no.2 (1993) 141-151
17. *http://www.panasonic.com/business/security/biometrics_data.asp*
18. *http://www.polhemus.com*

Eye Perturbation Approach for Robust Recognition of Inaccurately Aligned Faces

Jaesik Min, Kevin W. Bowyer, and Patrick J. Flynn

Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

Abstract. Extraction of normalized face from input images is an essential preprocessing step of many face recognition algorithms. Typical face extraction algorithms make use of the locations of facial features, such as the center of eyes, that are marked either manually or automatically. It is not guaranteed, however, that we always obtain the exact or optimal locations of the eye centers, and using inaccurate landmark locations, and consequently poorly registered faces, is one of the main causes of performance degradation in appearance-based face recognition. Moreover, in some applications, it is hard to verify the correctness of the face extraction for every query image. For improved performance and robustness to the eye location variation, we propose an eye perturbation approach that generates multiple face extractions from a query image by using the perturbed eye locations centered at the initial eye locations. The extracted faces are then matched against the enrolled gallery set to produce individual similarity scores. Final decisions can be made by using various committee methods – nearest neighbor, maximum vote, *etc.*– of combining the results of individual classifiers. We conclude that the proposed eye perturbation approach with nearest neighbor classification improves recognition performance and makes existing face recognition algorithms robust to eye localization errors.

1 Introduction

Many face recognition methodologies require, as an essential preprocessing step, the extraction of a normalized face region from the input image. In many appearance-based face recognition approaches, the face extraction is performed based on the locations of facial landmarks, such as eyes, nose, or mouth [1]. Once the coordinates of these landmarks are given, extraction of the face can be done through the processes of image scaling, rotation, intensity normalization, and aligning to a predetermined template, *etc.* that minimizes the variations unrelated to the identity.

The most prominent facial landmarks in 2D face images are the eyes [2], whereas it is the nose in 3D (depth) face images [3]. The locations of eye centers can be obtained either manually or automatically by using eye detection algorithms [4]. Often, however, the detected eye locations are unreliable; they are inaccurate and inconsistent across eye detectors. This causes sub-optimal face extraction, and consequently degrades recognition performance even with a good algorithm and images of well-posed faces [5]. In this paper we first investigates the effect of the accuracy of eye locations.

Minimizing the errors at the stage of localization is desired for this problem, but has a limit. An alternative solution is to take the existence of localization errors for granted,

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 41–50, 2005.
© Springer-Verlag Berlin Heidelberg 2005

and to design a recognition algorithm that is robust to the localization variation. In this paper we propose to produce multiple eye locations perturbed from the initial locations of both eyes and then use the extracted faces from these eye locations. We tested two representative face recognition algorithms, PCA and FaceIt, on a large number of face extractions that are generated from various sampling of eye locations. Then we compared the results of eye perturbation to the baseline.

The remaining sections of this paper are organized as follows. In Section 2, a number of related works are investigated. Sections 3 to 5 describe how we designed the experiments on eye perturbation and committee and discusses the effect of these factors on the performance. Section 6 shows compared results of the experiments. Section 7 summarizes our work and introduces topics to be addressed in future work.

2 Previous Works

The importance of eye localization as a preprocessing module in a face recognition system has been addressed by many researchers. Marques *et al.* [6] investigated the effect of eye position on a PCA-based face recognition algorithm. They used a total of 8 images and showed the sensitivity of the algorithm to the eye location deviations along various directions. As mentioned in their work, even the eye positions that are manually selected – or at least inspected – by human operators are unreliable and tend to deviate from a definition of the geometric eye center.

The role of eye locations in achieving high performance in face recognition systems received special focus in the paper by Riopka *et al.* [2]. They evaluated the effect of eye location accuracy through experiments of 3 different face recognition algorithms, that is, Principal Component Analysis (PCA), Elastic Bunch Graph Matching (EBGM), and FaceIt, on 1024 images from FERET database [7] by generating $17 \times 17 = 289$ perturbations of eye locations from the original locations and compared the recognition results. They first used ideal image data – that is, used the same image set for both gallery and probe sets – to measure the pure effect of eye perturbation. Then they applied the same perturbation to more realistic images. They report that using real image data did not degrade the performance drastically when the same eye perturbation is applied.

Some researchers have proposed solutions to the inaccurate localization problem. In the paper by Martinez [8], the gallery is augmented by perturbation and modeled by Gaussian Mixture Models (GMM). Shan *et al.* [9] defined robustness to misalignment in their paper and observed the effects of misalignment. They also proposed an enhanced Linear Discriminant Analysis (LDA) algorithm for face recognition that generated multiple ($9 \times 9 = 81$) virtual samples from each original training image by perturbation.

3 Experimental Design

A total of 600 subjects were selected partly from the FERET database [7] and partly from the University of Notre Dame (ND) database [10] so that their neutral expression face images are used in creating a training image set. Another 393 subjects from the ND database who participated between years of 2002 and 2003 were selected to create a test image set where each subject's earliest image is used as the gallery image and the



gallery probe

Fig. 1. Gallery and probe image samples of the University of Notre Dame database. Each image is either 1200×1600 or 1704×2272-pixel color image

latest image is put into the probe set (Figure 1). The elapsed time between the gallery and probe images ranges from 1 to 94 weeks, with 35 weeks on average. Both gallery and probe images are acquired under the same controlled environment, that is, lighting condition, background, and facial expression. There may exist slight and unintended pose variation and other variations over time.

As the recognition algorithms, we used two representative face recognition algorithms: PCA and FaceIt. For the PCA algorithm, we used the Version 5.0 code implemented at Colorado State University (CSU) [11]. The Mahalanobis Angle was selected as the distance metric, and no dimension reduction of the eigenspace was performed. For the FaceIt algorithm, we used the version G5, which was developed and distributed by Identix Incorporated.

The recognition performance is represented by a cumulative match characteristic (CMC) score, where CMC score at rank r is defined as the ratio of people whose correct match exists within r best matches. For example, a score of 85% at rank 1 means that 85% of people were correctly matched at the first choice. Similarly, a score of 90% at rank 3 means that 90% of people have their correct matches in the first three best matches. Therefore, a single recognition result gives different scores at different ranks, and the score at rank s is higher than or equal to the score at rank r, where $r < s$.

4 Effect of Inaccurate Localization

Previous studies [2, 9] show that PCA and LDA algorithms are sensitive to eye localization errors. Figure 2 shows the real examples of face extraction when eyes are localized by the eye locater module of the FaceIt software. Inaccurate localization yields undesirable, *e.g.*, scaled, rotated, or translated face templates. In this section, we investigate how the inaccurate localization affects the performance of algorithms. For this we set the manually marked eye locations as the ground truth and the automatically selected eye locations as the set of real-life samples, because it sounds more practical to get samples from a real eye locater rather than to add artificial random noises to the ground truth locations.

Originally, all of the 393 gallery images and 393 probe images are provided with ground truth eye locations. The probe images were also fed into the eye locater module of the FaceIt software to get the eyes localized automatically. These locations are compared to the ground truth eye locations of the same images. The difference between the

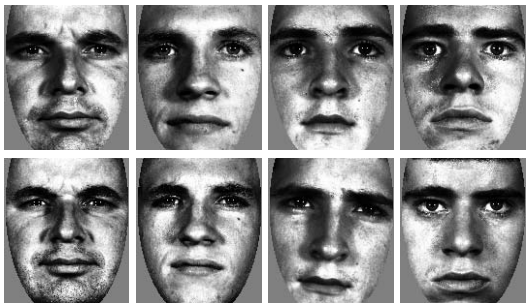


Fig. 2. Examples of poorly extracted faces due to the eye position deviation. Faces at the top row are from gallery images with ground truth eye positions and faces at the bottom are from corresponding probe images with automatically marked eye positions

manual and the automated eye locations is 10.7 pixels on average, with standard deviation of 5.7 pixels, while the average distance between two eye centers in the ground truth is 268.8 pixels.

We ran three face recognition algorithms at hand on this gallery set (with ground truth) and probe set (automatic markings). Figure 3 shows how the performance of each algorithm degrades with the inaccurate eye locations. As expected, the PCA algorithm degrades abruptly, confirming that it is highly dependent on the localization accuracy. The FaceIt and EBGM algorithms turned out to be relatively tolerant to the inaccurate localizations; the performance also degrades, but the amount of degradation is negligible. We do not know what FaceIt does to handle this problem, and EBGM adjusts itself to some degree. Similar experiments with FaceIt performed in [2] showed large degradation with the "weathered" image set. In the next session we propose a method of augmenting the probe data to solve the problem caused by misalignment.

5 Eye Perturbations

Using large and representative samples per class is the best way to assure better classification, but it is not always feasible [8]. Generating multiple versions of face templates from a limited number of originals, thus augmenting the dataset, is one promising solution, as in [8, 9]. To solve the problem of inaccurate eye localization as discussed in Section 4, we propose to augment the probe images by perturbing the initial eye locations.

In real-life applications, the gallery set resides in the database, thus its quality and metadata are under strict control. In contrast, the probe images usually are transient and its quality (along with that of metadata) is less controlled. Thus, it is more likely the probe images have bad eye localization than the gallery images do. Therefore, instead of augmenting the gallery set as in [8, 9], we propose to augment the probe images. By keeping the gallery set and augmenting the probe images, the face recognition system becomes more flexible in that the degree of dataset augmentation is easily adjustable accordingly; there is no need of rebuilding or remodeling of the system.

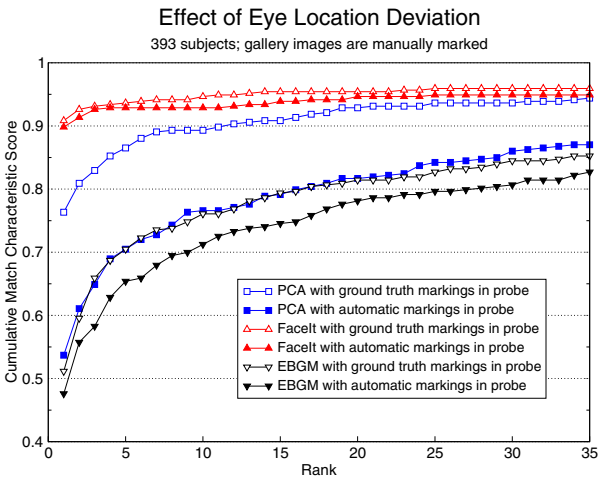


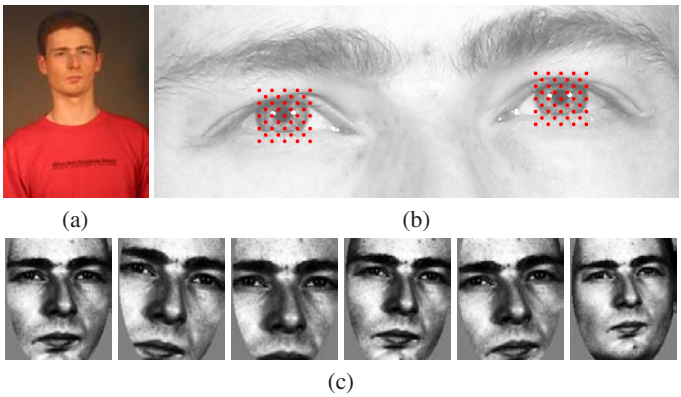
Fig. 3. The performance degradation due to inaccurate eye localization. The CMC at rank 1 of PCA algorithm dropped 76.3 % to 53.5 %. The EBGM (51.1 % to 47.6 %) and FaceIt (90.8 % to 89.8 %) were less affected by the deviation

All of 786 images used in our experiments are accompanied by the ground truth facial landmarks, which are marked by a number of different human operators and are highly reliable in spite of the existence of slight variations across operators and over time. The images also are provided with the machine-selected eye locations. For each original query image, we generate multiple normalized faces by perturbing the initial – either ground truth or machine-selected – eye locations (Figure 4). The sampling window size is set to 49×49 pixels so that it covers an area slightly wider than the iris. We sample 41 uniformly distributed locations for each eye, a total of $41\times41=1681$ pairs of eye locations, and thus generate the same number of normalized faces for each query image.

Each normalized face probe matches against the gallery set and produces distance measures to each of the 393 gallery images. So for each query image we will have 1681 individual classification results. A number of committee schemes to combine these results are available, such as nearest neighbor, k-nearest neighbors, weighted sum, maximum vote, *etc.* So far we tested the nearest neighbor and maximum votes. In the nearest neighbor (NN) scheme, we simply select the pair of probe and gallery with the minimum distance – or the highest similarity score in FaceIt terminology. In the maximum vote scheme, the gallery image that gets the maximum number of NN selections from 1681 individual normalized face probes is finally selected.

6 Results and Discussion

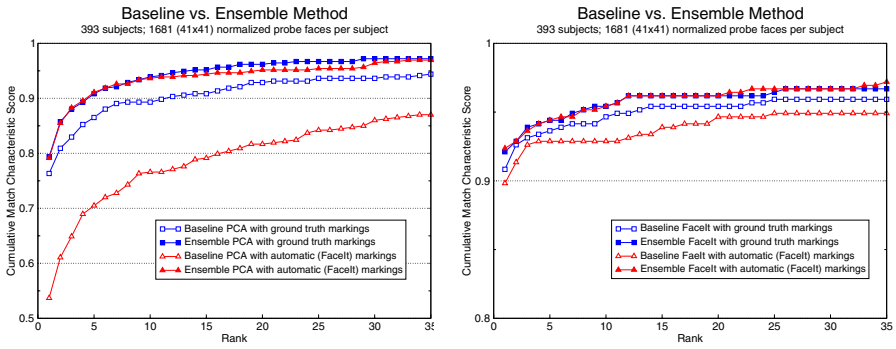
We compared the NN ensemble method to the baseline on the PCA and the FaceIt algorithms, where both ground truth and machine-selected eye markings are provided (Fig. 5 (a)). The NN ensemble PCA algorithm scored 79.4 % rank-1 CMC, marginally improved from that of the baseline PCA, 76.3 %. The improvement achieved by the



(a) (b)

(c)

Fig. 4. An example of multiple generation of normalized faces from a probe image. Given an original image (a), possibly with inaccurate eye locations, 41 sampling locations centered at the initial eye locations are selected for each eye as illustrated in (b). Six out of 1681 normalized faces are shown in (c)

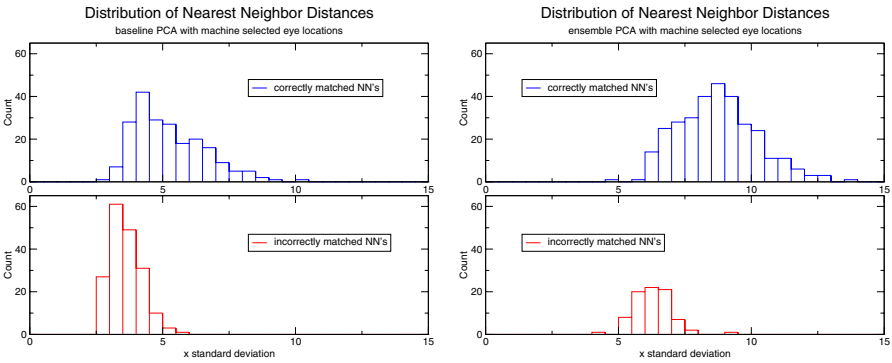


(a) PCA (b) FaceIt

Fig. 5. Comparison between the baseline and ensemble methods. The ensemble methods achieved both improvement (significant or marginal) and stability in performance

NN ensemble is just over 3 %, but considering that the baseline performance was with the ground truth, it promises greater improvement with machine-selected eye locations. The experiment of the ensemble PCA algorithm with machine-selected eye locations reached 79.1 % rank-1 CMC, which is a huge improvement from the baseline performance of 53.7 %. The comparison of baseline and ensemble FaceIt is shown in Fig. 5 (b). The baseline performance of FaceIt is already high enough, but we still observe marginal improvements, and the amount of improvement is a little higher in case of machine-selected eye locations, which was expected. The overall CMC curves shown in Figure 5 indicate that the ensemble method also achieved stability in performance as well as improvement, that is, we observe only negligible difference in performance between the ground truth and automatic markings.

At this point we need to analyze the mechanism of the maximum vote scheme, which yields low performance. The maximum-vote ensemble method was also applied



(a) baseline PCA (b) ensemble PCA

Fig. 6. The distribution of nearest neighbor distances in the (a) baseline and (b) ensemble PCA with the automatically marked eye locations. Each of the 393 probes have 393 distance measures to the gallery images, and this plot shows how the 393 NN distance outlies in the overall distribution of probe-to-gallery distances

with the PCA algorithm, but it achieved lower performance (63.6 % rank-1 CMC) than that of the baseline. In general, the NN pair of a probe and a gallery image is the extreme outlier in the distribution of distances between probe and gallery images. We investigated how far the NN distance lies in the distance distribution. In the baseline experiment the correctly matched NN distances lie at, on average, 5.2 standard deviation of the distance distribution, and the incorrectly matched NN distances lie at 3.6 standard deviation (Figure 6 (a)). In both cases, the NN distances are the extreme outliers in the distribution whose p-values are less than 0.001. This extremity of the NN distance gets further (Figure 6 (b)) in the ensemble scheme because it produces a better (or equal at least) NN distance and adds a huge amount of mediocre distances. This explains the poor performance achieved by the maximum vote committee method, where the newly produced NN distance just casts one vote equally as the other 1680 distances do. Therefore, hereinafter we discard the maximum vote committee scheme and focus on the NN scheme only.

Figure 7 shows two examples of successful NN match after the eye perturbation. At the top row, the original probe in the baseline was matched to a wrong gallery image, and the correct gallery image scored rank of 150; after the eye perturbation, one of the perturbed probes was matched to the correct gallery image. At the bottom row of the figure, which is the case where the machine-selected eye locations were provided, the rank score has jumped from 131 to 1.

However, augmenting the dataset not always improve the performance. It is possible that some of the enlarged data may match to wrong gallery images with smaller distances than that of correct match. In our experiments it actually happened (Figure 8), but the rank change is relatively small. The count and amount of performance improvement and degradation are summarized in Table 1. In the PCA case with ground truth, the number of instances of improvement is not much more than that of degradation, but the average amount of rank change is larger, which gives overall improvement. In the PCA case with machine-selected eye locations, both the number and the amount of rank



(a) (b) (c) (d)

Fig. 7. Successful cases of eye perturbation. The probe with initial eye localization (a) is matched to a wrong gallery image (b). After perturbation, a new probe image (c) is successfully matched to the correct gallery image (d). It is shown that the probe image (a) and the gallery image (d) are not well aligned



(a) (b) (c) (d)

Fig. 8. Examples of degradation after eye perturbation. The probe with initial eye localization (a) is matched to a correct gallery image (b). After perturbation, a new probe (c) picked up a wrong match (d) that has smaller distance

change is large, which explains the big jump in the CMC curve in Figure 5. The FaceIt rank results have similar patterns, although less obvious.

There also exist cases where the proposed method cannot be the solution. The subject in Figure 9, for example, has significant pose change between the gallery and probe images. Neither PCA nor FaceIt succeeded in matching this subject correctly both in the baseline and in the ensemble method because the problem here comes from the pose angle rather than from the localization accuracy.

7 Conclusions and Future Works

In this paper we showed the effect of inaccurate eye localization on the performance of face recognition and proposed a method that is robust to the effect. We first investigated the impact of eye localization accuracy through experiments with two sets of realistic localization data; a set of manual markings, which is used as the ground truth, and another set automatically marked by a commercial software, which served as the devi-

Table 1. The rank change between the baseline and the NN ensemble methods

	PCA				FaceIt			
	GT		Auto		GT		Auto	
	Count	Amount	Count	Amount	Count	Amount	Count	Amount
Improved	**69**	**36.4**	**160**	**44.2**	**22**	**16.7**	**31**	**53.6**
Degraded	**51**	**7.9**	**28**	**16.3**	**13**	**19.2**	**10**	**13.3**
Unchanged	273		205		358		352	



(a) (b) (c) (d) (e)

Fig. 9. A failed match after eye sampling. The various extractions of the probe images, (a) and (c), could not be matched to the correct gallery image, (e), because the pose difference between the original gallery and probe images is significant. Image (c) and (d) are the nearest neighbor pair after the eye perturbation

ation from the ground truth. By using large sets of images with substantial time lapse between the gallery and probe images, and by using real-life outputs of eye localization, we showed that, for some face recognition algorithms, the accuracy of eye localization is critical to the recognition performance.

Based on the baseline experimental results, we proposed an eye perturbation approach to make existing face recognition algorithms robust to the eye localization variation. A number of experiments with ground truth and machine-selected eye locations showed that achieving both improvement and robustness was successful.

It will be worth investigation to extend this experiments with image sets of larger variety. As mentioned in [2], the inaccurate eye localization may have the greatest impact on controlled pairs of gallery and probe images; using pairs of different conditions in the probe images – *e.g.*, uncontrolled probe images against controlled gallery – might attenuate the effect of inaccurate localization.

Currently the increased computational cost is the main problem of the proposed approach. We used a full-scale eye perturbation for a thorough investigation, but a smaller and sparser sampling may be enough for the intended purpose. Alternatively, the degree of perturbation may be parameterized so that the degree can be adjustable. We also plan to design an intelligent decision algorithm by modeling the distribution of NN distances as shown in Figure 6, so that it can decide the necessity and degree of the eye perturbation, methods of combining individual classifications, *etc.*

Acknowledgments

This work is supported by National Science Foundation grant EIA 01-20839 and Department of Justice grant 2004-DD-BX-1224.

References

1. Chellappa, R., Wilson, C., Sirohey, S.: Human and machine recognition of faces: A survey. Proceedings of IEEE **83** (1995) 705–740
2. Riopka, T., Boult, T.: The eyes have it. In: Proc. of the ACM SIGMM workshop on biometric methods and applications. (2003) 9–16
3. Chang, K., Bowyer, K., Flynn, J.: Face recognition using 2D and 3D facial data. In: ACM Workshop on Multimodal User Authentication. (2003) 25–32
4. Hsu, R., Abdel-Mottaleb, M., Jain, A.: Face detection in color images. IEEE Trans. Pattern Anal. and Mach. Intel. **24** (2002) 696–706
5. Zhao, W.: Improving the robustness of face recognition. In: Audio- and Video-Based Biometric Person Authentication. (1999) 78–83
6. Marques, J., Orlans, N., Piszcz, A.: Effects of eye position on eigenface-based face recognition scoring. Technical Report, MITRE Corporation (2000)
7. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. and Mach. Intel. **22** (2000) 1090–1104
8. Martinez, A.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. Pattern Anal. and Mach. Intel. **24** (2002) 748–763
9. Shan, S., Chang, Y., Gao, W., Cao, B., Yang, P.: Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution. In: International Conference on Automatic Face and Gesture Recognition. (2004) 314–320
10. Min, J., Flynn, P., Bowyer, K.: Assessment of time dependency in face recognition. TR-04-12, University of Notre Dame (2004)
11. Beveridge, R., Draper, B.: Evaluation of face recognition algorithms (release version 5.0) (URL: http://www.cs.colostate.edu/evalfacerec/index.html)

On Combining Textural and Geometrical Scores for Discriminative Face Authentication

José Luis Alba-Castro and Daniel González-Jiménez

Departamento de Teoría de la Señal y Comunicaciones
Universidad de Vigo, Spain
{jalba,danisub}@gts.tsc.uvigo.es

Abstract. In this paper, a combined shape-texture approach to discriminative face authentication is studied. Facial texture information is captured through Gabor responses (jets), similarly to the Elastic Bunch Graph Matching approach, but the points where filters are applied are located over lines that sketch the face. In this way, textural information is "shape-driven" and unlike other Gabor-based approaches, it shows a person-dependent behaviour. For every pair of face images, the score obtained through jets is combined with 3 measurements of pair-wise shape distortion. Discriminative Fisher methods are applied at the shape algorithm level and at the score combination level, in order to get a unified score ready for classification. Face verification results are reported on configuration I of the XM2VTS database.

1 Introduction

Most of face recognition systems rely on a compact representation that encodes global and/or local information. Global approaches are mainly represented by linear projection methodologies (among others: Principal Component Analysis (Eigenfaces [1]), Linear Discriminant Analysis (Fisherfaces [2]), Independent Component Analysis [3] , etc.). These methods are devoted to encode faces in an efficient manner, and characterize the spanned face space or manifold. Local approaches have been based on finding and characterizing informative features such as eyes, nose, mouth, chin, eyebrows, etc. If their mutual relationships are considered, then we have a local-global approach (Local Feature Analysis [4]). Inside this last group of methods we can find the Elastic Bunch Graph Matching (EBGM) algorithm [5]. It combines local and global representation of the face by computing multi-scale and multi-orientation Gabor responses (jets) from a set of the so-called fiducial points, located at specific face regions (eyes, tip of the nose, mouth,..., i.e. "universal features"). Finding every fiducial point is one of the most critical parts of EBGM, either in terms of accuracy or in terms of computational burden. This search relies on a matching process between the candidate jet and a bunch of jets extracted from the corresponding fiducial points in different faces. In this way, there are several variables that can affect the accuracy of the final fiducial point locations, as differences in pose, illumination conditions and insufficient representativeness of the stored bunch

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 51–60, 2005.
© Springer-Verlag Berlin Heidelberg 2005

of jets. Once fiducial points are adjusted, only textural information is used in the classifier. In [13], an easier way to locate grid-nodes was presented by taking advantage of illumination-invariant features from which geometry of face can be characterized.

As it will be seen later, our method chooses a set of points, $\mathcal{P}$, in an image $\mathcal{F}$ from lines that characterize the face. These sets are the base for shape distortion measurements and texture similarities. The set of selected points turned out to be quite robust against illumination conditions and slight variations in pose. Many of the points located belong to "universal" features, but many others are person-dependent. In this way we say that this method is inherently discriminative, in contrast to trainable parametric models. So, EBGM locates a pre-defined set of "universal" features and our approach locates a person-dependent set of features. As a byproduct of the correspondence algorithm [6], we extract two measurements of local geometrical distortion. Gabor jets are then calculated from the correspondent points and the final dissimilarity function compiles geometrical and local texture information. In order to include a global measurement of shape distortion, a Hausdorff-based distance has also been added to the final classifier.

Our method can be splitted into several steps. Given two faces, say $\mathcal{F}_1$ and $\mathcal{F}_2$, it normalizes both of them to a standard size, obtaining face sketches and choosing a set of points from those sketches, which must be matched. Later, Gabor responses are calculated at selected locations and this textural information is combined with geometrical distortions to compute the final dissimilarity score between $\mathcal{F}_1$ and $\mathcal{F}_2$. The different steps of our method are detailed in the next sections. Section 2 explains the operator used to extract face lines and introduces the global shape score between face images. The grid adjustment, the way we select points and the algorithm used to match these sets of points are also described in this section. The Gabor filters used to extract texture are explained in section 3. Section 4 introduces different geometrical terms and the *Sketch Distortion* concept used to measure dissimilarity between faces. Experimental results are given in section 5. Finally, conclusions are drawn in section 6.

2 Computation of Fiducial Points and Global Shape Distortion

In this work, shape information has been obtained using the ridges and valleys operator because of its robustness against illumination changes [7]. Moreover, the relevance of valleys in face shape description has been pointed out by some cognitive science works [8]. In this paper, we have used the ridges and valleys obtained by thresholding the so-called multi local level set extrinsic curvature (MLSEC) [9]. The MLSEC operator works here as follows: **i)** computing the normalized gradient vector field of the smoothed image, **ii)** calculating the divergence of this vector field, which is bounded and gives an intuitive measure of valleyness (positive values running from 0 to 2) and ridgeness (negative values from -2 to 0), and **iii)** thresholding the response so that image pixels where the

MLSEC response is smaller than -1 are considered ridges, and those pixels larger than 1 are considered valleys.

Now that the feature descriptor has been properly defined, we have a way of describing fiducial points in terms of positions where the geometrical image features have been detected.



Fig. 1. Left: Original Image. Center-left: Valleys and ridges image. Center-right: Thresholded ridges image. Right: Thresholded valleys image

Once we have extracted the ridges and valleys from two face images, a global shape score can be obtained. One of the most successful dissimilarity measurements for sets of points (or binary images) is the Hausdorff distance, that has been widely used for object matching in scene analysis [10]. It is well known that the standard Hausdorff distance is quite sensible to outliers, so some modifications [11] have been used to avoid such a problem. In this work, we have used a particular modification that can be referred as *Average Hausdorff Distance (AHD)*. Given two sets $\mathcal{A}$ and $\mathcal{B}$, the directed Average Hausdorff Distance $ahd(\mathcal{A},\mathcal{B})$ from the set $\mathcal{A}$ to the set $\mathcal{B}$, (assuming Euclidean distance between set elements) is:

$$ahd(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \min_{b\in\mathcal{B}}(\|a-b\|) \tag{1}$$

where $|\mathcal{A}|$ denotes the cardinal of the set $\mathcal{A}$. So, the (symmetric) *Average Hausdorff Distance* (AHD) can be formally written as:

$$AHD(\mathcal{A},\mathcal{B}) = \frac{1}{2}(ahd(\mathcal{A},\mathcal{B}) + ahd(\mathcal{B},\mathcal{A})) \tag{2}$$

The computation of $AHD(\mathcal{A},\mathcal{B})$ is easily performed as a double dot product: given our binary image $\mathcal{F}_1(x,y)$ that can be thought of as the output of any contour operator, with $\mathcal{A} = \{(x,y)|\mathcal{F}_1(x,y) = 1\}$; we can define $\overrightarrow{\mathcal{F}_1}$ as the binary vector associated to the binary image $\mathcal{F}_1$, and $\widehat{\mathcal{F}_1} = \frac{1}{|\mathcal{A}|}\overrightarrow{\mathcal{F}_1}$ the associated normalized vector. For a digital binary image, we can define the Distance Transform, $D(\mathcal{F}_1)$ [12], as a point-wise transform that contains, for each pixel, the distance between that pixel and the pixel of value 1 closest to it. The vector format for the distance transform $D(\overrightarrow{\mathcal{F}_1})$ can also be extended to the associated normalized image $D(\widehat{\mathcal{F}_1})$ with the same meaning. With these definitions, the AHD between binary images can then be calculated averaging inner products:

$$AHD(\mathcal{F}_1,\mathcal{F}_2) = \frac{1}{2}(<\widehat{\mathcal{F}_1}, D(\widehat{\mathcal{F}_2})> + <\widehat{\mathcal{F}_2}, D(\widehat{\mathcal{F}_1})>) \tag{3}$$

2.1 Point Matching

Once the ridges and valleys in a new image have been extracted, we must sample these lines in order to keep a set of points for further processing. There are some possible combinations, in terms of using just ridges, just valleys or both of them, so we will refer to the binary image, obtained as a result of the previous step, as the sketch from now on.

In order to select a set of points from the original sketch, a dense rectangular grid ($\mathcal{N}_x \times \mathcal{N}_y$ nodes) is applied onto the face image and each grid node changes its position until it finds the nearest line of the sketch. So, finally, we get a vector of points $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$,[1] where $\boldsymbol{p_i} \in \mathbb{R}^2$. These points sample the original sketch, as it can be seen in figure 2.

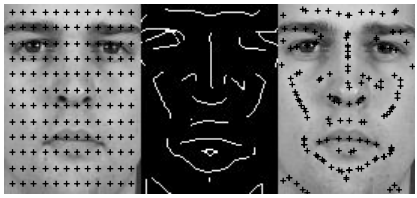


Fig. 2. Left: Original rectangular dense grid. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch

In order to compare feature vectors extracted at these positions, we must first compute the matching between points from both images. We have adopted the idea described in [6]. For each point i in the constellation, we compute a 2-D histogram h_i of the relative position of the remaining points, so that a vector of distances $\mathcal{D} = \{d_{i1}, d_{i2}, \ldots, d_{in}\}$ and a vector of angles $\boldsymbol{\theta} = \{\theta_{i1}, \theta_{i2}, \ldots, \theta_{in}\}$ are calculated for each point. As in [6], we employ bins that are uniform in log-polar space, i.e. the logarithm of distances is computed. Each pair $(\log d_{ij}, \theta_{ij})$ will increase the number of counts in the adequate bin of the histogram.

Once the sets of histograms are computed for both faces, we must match each point in the first set $\mathcal{P}$ with a point from the second set $\mathcal{Q}$. A point $\boldsymbol{p}$ from $\mathcal{P}$ is matched to a point $\boldsymbol{q}$ from $\mathcal{Q}$ if the term C_{pq}, defined as:

$$C_{pq} = \sum_k \frac{[h_p(k) - h_q(k)]^2}{h_p(k) + h_q(k)} \tag{4}$$

is minimized[2]. Finally, we have a correspondence between points defined by ξ:

$$\xi(i) : \boldsymbol{p_i} \Longrightarrow \boldsymbol{q_{\xi(i)}} \tag{5}$$

where $\boldsymbol{p_i} \in \mathcal{P}$ and $\boldsymbol{q_{\xi(i)}} \in \mathcal{Q}$.

[1] $n = \mathcal{N}_x \times \mathcal{N}_y$. Typical sizes for n are 100 or more nodes

[2] k in (4) runs over the number of bins in the 2D histogram

The vectors of angles $\boldsymbol{\theta} = \{\theta_{i1}, \theta_{i2}, \ldots, \theta_{in}\}$ are calculated taking the x-axis (the vector $(1,0)^T$) as reference. This is enough if we are sure that the faces are in an upright position. But, to deal with rotations in plane, i.e. if we do not know the rotation angle of the heads, we must take a relative reference for the shape matching algorithm to perform correctly. Consider, for the set of points $\mathcal{P} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \ldots, \boldsymbol{p_n}\}$, the centroid of the constellation $\boldsymbol{c_\mathcal{P}}$:

$$\boldsymbol{c_\mathcal{P}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{p_i} \tag{6}$$

For each point $\boldsymbol{p_i}$, we will use the vector $\overrightarrow{p_i c_\mathcal{P}} = \boldsymbol{c_\mathcal{P}} - \boldsymbol{p_i}$ as the x-axis, so that rotation invariance is achieved. Also, the angle between the two images, φ, can be computed as follows:

$$\varphi = \frac{1}{n} \sum_{i=1}^{n} \angle \left(\overrightarrow{p_i c_\mathcal{P}}, \overrightarrow{q_{\xi(i)} c_\mathcal{Q}} \right) \tag{7}$$

so that the system is able to put both images in a common position for further comparison. If we do not take this angle into account, textural extraction will not be useful for our purposes.

3 Local Texture Similarity

For this shape descriptor to be useful in face authentication, local texture information must be also taken into account. Gabor wavelets are biologically motivated convolution kernels that capture this kind of information and are also quite invariant to the local mean brightness, so an efficient face encoding approach will be to extract texture from these geometrically salience regions. The system uses a set of 40 Gabor filters, with the same configuration employed in [5]. These filters are convolution kernels in the shape of plane waves restricted by a Gaussian envelope, as it is shown next:

$$\psi_m \left(\overrightarrow{x} \right) = \frac{\left\| \overrightarrow{k}_m \right\|^2}{\sigma^2} \exp \left(- \frac{\left\| \overrightarrow{k}_m \right\|^2 \left\| \overrightarrow{x} \right\|^2}{2\sigma^2} \right) \left[\exp \left(i \overrightarrow{k}_m \cdot \overrightarrow{x} \right) - \exp \left(-\frac{\sigma^2}{2} \right) \right] \tag{8}$$

where $\overrightarrow{k}_m$ contains information about frequency and orientation of the filters, $\overrightarrow{x} = (x, y)^T$ and $\sigma = 2\pi$.

The region surrounding a pixel in the image is encoded by the convolution of the image patch with these filters, and the set of responses is called a jet, $\mathcal{J}$. So, a jet is a vector with 40 coefficients, and it provides information about a specific region of the image. Each coefficient, $\mathcal{J}_k$, can be expressed as follows:

$$\mathcal{J}_k \left(\mathcal{I} \left(x_0, y_0 \right) \right) = \sum_x \sum_y \mathcal{I}(x, y) \psi_k \left(x_0 - x, y_0 - y \right) \tag{9}$$

In the previous step, we have selected n points from the face image, but in order to avoid overlapping between responses of filters and to reduce computational time, we must leave just a few of them, from which we will extract textural information. So, we decided to establish a minimum distance D between each pair of nodes, so that all final positions are separated at least by D. As a consequence, the number of final points, n_D, will be less or equal than n. Let $\mathcal{P}' = \{\boldsymbol{p'_1}, \boldsymbol{p'_2}, \ldots, \boldsymbol{p'_{n_D}}\}$ denote the set of final points for textural extraction, and let $\mathcal{R} = \left\{\mathcal{J}_{\boldsymbol{p'_1}}, \mathcal{J}_{\boldsymbol{p'_2}}, \ldots, \mathcal{J}_{\boldsymbol{p'_{n_D}}}\right\}$ be the set of jets calculated for one face. The similarity function between two faces, $\mathcal{S}_{\mathcal{J}}(\mathcal{F}_1, \mathcal{F}_2)$ results in:

$$\mathcal{S}_{\mathcal{J}}(\mathcal{F}_1, \mathcal{F}_2) \equiv \mathcal{S}_{\mathcal{J}}\left(\mathcal{R}^1, \mathcal{R}^2\right) = \frac{1}{n_D} \sum_{i=1}^{n_D} < \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 > \tag{10}$$

where $< \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 >$ represents the normalized dot product between the i-th jet from $\mathcal{R}^1$ and the correspondent jet from $\mathcal{R}^2$, but taking into account that only the moduli of jet coefficients are used.

4 Discriminative Shape and Textural Distortion

Global shape distortion has been taken into account throughout the computation of $AHD(\mathcal{F}_1, \mathcal{F}_2)$. In this section, local shape distortions will be handled. So, we introduce two different terms here:

$$\mathcal{GD}_1(\mathcal{F}_1, \mathcal{F}_2) \equiv \mathcal{GD}_1(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{n} v_i C_{i\xi(i)} \tag{11}$$

$$\mathcal{GD}_2(\mathcal{F}_1, \mathcal{F}_2) \equiv \mathcal{GD}_2(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{n} w_i \left\| \overrightarrow{p_i c_{\mathcal{P}}} - \overrightarrow{q_{\xi(i)} c_{\mathcal{Q}}} \right\| \tag{12}$$

Equation (11) computes geometrical distortion by linearly combining the individual costs represented in (4). On the other hand, (12) calculates metric deformation by combining the norm of the difference vector between matched points[3].

Weighting vectors v and w can be simply set to the vector $\overrightarrow{1}$ or can be discriminatively calculated. When dealing with face shape distortion, it is obvious that regions related to face muscles are more likely to suffer slight displacements than others. Hence, the local contributions in $\mathcal{GD}_1$ and $\mathcal{GD}_2$ must be weighted accordingly. We have found the n components of v and w as the Fisher best discriminative direction between the local shape distortion vectors for evaluation clients and impostors. $\mathcal{GD}_1$ and $\mathcal{GD}_2$ can be seen as global shape distortion measurements, that should be large for faces of different subjects and small for faces representing the same person. If faces are in an upright position and are

[3] Note that the centroid of the constellation has been substracted from the point coordinates in order to deal with translation

scaled at the same size, adding the global distorsion $AHD(\mathcal{F}_1, \mathcal{F}_2)$ increases the discriminative power of the shape part of the classifier, as it will be seen at the results section.

Now we can think of linearly combining jet dissimilarity, $[1 - \mathcal{S}_{\mathcal{J}}(\mathcal{F}_1, \mathcal{F}_2)]$, with shape deformations, resulting in the final dissimilarity function $\mathcal{DS}(\mathcal{F}_1, \mathcal{F}_2)$:

$$\mathcal{DS}(\mathcal{F}_1, \mathcal{F}_2) = \lambda_1 [1 - \mathcal{S}_{\mathcal{J}}(\mathcal{F}_1, \mathcal{F}_2)] + \lambda_2 \mathcal{GD}_1(\mathcal{F}_1, \mathcal{F}_2) + \lambda_3 \mathcal{GD}_2(\mathcal{F}_1, \mathcal{F}_2) + \lambda_4 AHD(\mathcal{F}_1, \mathcal{F}_2) \tag{13}$$

with $\lambda_i > 0$. The combination of $\mathcal{GD}_1$ and $\mathcal{GD}_2$ is what we call *Sketch Distortion* (SKD). From (13) and using (10), (11) and (12), it follows that $\mathcal{DS}(\mathcal{F}_1, \mathcal{F}_2)$ is equal to:

$$\sum_{i=1}^{n} \left[\lambda_1 \frac{1 - < \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 >}{n_D} + \lambda_2 C_{i\xi(i)} + \lambda_3 \left\| \overrightarrow{p_i c_{\mathcal{P}}} - \overrightarrow{q_{\xi(i)} c_{\mathcal{Q}}} \right\| \right] + \lambda_4 \cdot AHD(\mathcal{F}_1, \mathcal{F}_2) \tag{14}$$

Equation (14) needs an explanation. The index i in (14) runs over the entire set of points, although only a subset of them was used to compute jet similarity, as it was explained in the previous section. When i refers to a point that was not used to calculate a jet, only geometrical dissimilarity is taken into account, as $1 - < \mathcal{R}_i^1, \mathcal{R}_{\xi(i)}^2 >$ is set to 0. Except for this, in (14) we can see that each contribution of jet dissimilarity is modified with a weighted geometrical distortion (the so-called *Local Sketch Distortion* or $LSKD$). A high value in $LSKD$ from the pair $(\boldsymbol{p_i}, \boldsymbol{q_{\xi(i)}})$ means that they are not positioned over the same face region, so that jet dissimilarity will also be high. This fact is more likely to occur when incoming faces do not represent the same person. Even if $LSKD$ is low, but faces do not belong to the same person, textural information will increase the dissimilarity between them. On the other hand, when faces belong to the same subject, low $LSKD$ values should be generally achieved, so that matched points are located over the same face region, resulting in a low jet dissimilarity. Thus, the measurement in (14) reinforces discrimination between subjects. Figures 3 and 4 give a visual understanding of this concept. Figure 3 shows two instances of face images from subject A, while faces in figure 4 belong to subject B. The visual geometric difference between the two persons is reflected in the Sketch Distortion term, whose values are shown in table 1.

The scores weighting vector $\overrightarrow{\lambda} = [\lambda_1, \lambda_2, \lambda_3, \lambda_4]^T$ is absolutely necessary to avoid that scores with weak performance provoke an useless score combination.

Table 1. Sketch Distortion (SKD) between the face images from figures 3 to 4

		Subject A		Subject B	
		Image 1	Image 2	Image 1	Image 2
Subject A	Image 1	0	1851	3335	3226
	Image 2	1851	0	3053	2821
Subject B	Image 1	3335	3053	0	1889
	Image 2	3326	2821	1889	0

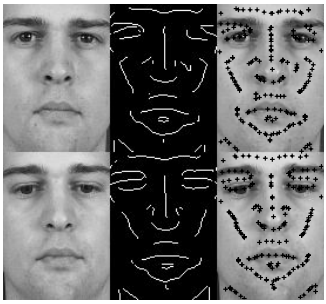


Fig. 3. Top: Left: First image from subject A. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch. **Bottom**: Left: Second image from subject A. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch

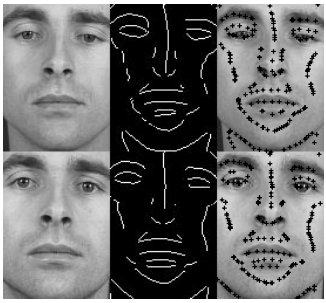


Fig. 4. Top: Left: First image from subject B. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch. **Bottom**: Left: Second image from subject B. Center: Valleys and ridges sketch. Right: Grid adjusted to the sketch

5 Results

In this section, we present the results using this new approach. From our previous work ([13]), the matching procedure between points has been improved, and the *Sketch Distortion* and *AHD* terms have been introduced. As before, we use the XM2VTS database [14] and the Lausanne protocol (configuration I) [15] for testing. The modifications mentioned above reduced the error rates, as shown in table 2. In the second row, although only textural information (T) is used, i.e. $\lambda_1 = 1, \lambda_{2,3,4} = 0$, some shape information still remains, because jets are extracted and compared at geometrically matched fiducial points. The next row shows the performace using only the AHD score. Rows 4th and 5th show the performance using the $\mathcal{GD}_1$ and the $\mathcal{GD}_2$ scores with Fisher weighting vectors (v and w) for balancing local shape distortion. The results in the sixth row ($T + SKD$) were achieved by using $\lambda_{1,2,3} = 1, \lambda_4 = 0$. Seventh row ($T + AHD$) shows performance with $\lambda_{1,4} = 1, \lambda_{2,3} = 0$. Next row presents the error rates with $\overrightarrow{\lambda} = [1, 1, 1, 1]^T$. Finally, the last row shows the results using the two vectors v and w mentioned above, and a second level of Fisher discriminative weighting for balancing individual scores λ_i.

Table 2. $FRR_{ev}(\%)$,$FAR_{ev}(\%)$,$FAR_{test}(\%)$ and $FRR_{test}(\%)$ (at EER threshold) for different configurations

Method	$FRR_{ev}(\%)$	$FAR_{ev}(\%)$	$FAR_{test}(\%)$	$FRR_{test}(\%)$
Previous work [13]	2.4	2.2	2.5	8.4
Textural (T)	3.17	2.36	2.5	5.11
AHD	8.67	6.08	11.75	7.22
$\mathcal{GD}_1$	13.5	6.41	29.75	11.12
$\mathcal{GD}_2$	13.17	7.21	38	12.09
$T + SKD$	3.33	1.73	5.75	4.23
$T + AHD$	4.17	2.76	4.75	4.93
$T + SKD$+AHD	2.67	2.02	4.25	4.26
Fisher combination	1.83	1.86	2.25	4.33

From this table we can highlight: **i)** Textural information extracted from person-dependent points performs better than any of the shape measurements tested, **ii)** $\mathcal{GD}_1$ and $\mathcal{GD}_2$, obtained as a byproduct of the point matching process do not perform well alone. Moreover, the direct combination of SKD with jet dissimilarity yields a worse performance than using Gabor responses alone, and the same for $(T + AHD)$ and $(T + SKD + AHD)$, but **iii)** both types of shape distortion help to reduce error rates when they are discriminatively combined with jet dissimilarity (a relative improvement of 13.53%).

We have also tested our verification system using the BANCA Database [16] on protocol MC and obtained an average WER of 4,89% with $\lambda_1 = 1, \lambda_{2,3,4} = 0$. An implementation of the EBGM algorithm from the CSU Face Identification Evaluation System (http://www.cs.colostate.edu/evalfacerec/index.html) on the same database and protocol gave an average WER of 8,79% [17]. With the above vector $\overrightarrow{\lambda}$, the main difference of both algorithms is the location of fiducial points, so it seems clear that our verification system selects more discriminative points.

6 Conclusions

In this paper, we have presented an inherently discriminative approach to automatic face recognition by combining shape and textural information. Fiducial points are located over lines that depict each individual face geometry, and shape differences between constellations of points from two faces are measured using the *Sketch Distortion* and the AHD terms. Gabor jets provide the textural information as defined in [5]. Two-level Fisher discriminative weightings were used to achieve results over the standard XM2VTS database. These results show that the method is comparable to the best ones reported in the literature and a clear improvement from those reported in [13].

References

1. Moghaddam, B. and Pentland, A., "Probabilistic Visual Learning for Object Representation," IEEE Trans. on PAMI, 19(7), 696-710, 1997

2. Belhumer, P.N., Espanha J.P. and Kriegman D.K., "Eigenfaces vs. Fisherfaces: Recognition using class-specific linear projection," IEEE Trans. on PAMI, 19(7), 711-720, 1997
3. Bartlett, M.S. and Sejnowski, M.S., "Viewpoint invariant face recognition using independent component analysis and attractor networks," NIPS, M.Mozer, et al.Editors, 1997, MIT Press
4. Penev, P. and Atick J., "Local feature analysis: a general statistical theory for object representation," Network: Computation in Neural Systems 7 (August 1996) 477-500
5. Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C. "Face recognition by Elastic Bunch Graph Matching." IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 775-779, 1997
6. Belongie, S., Malik, J., Puzicha J. "Shape Matching and Object Recognition Using Shape Contexts." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 24, April 2002
7. Pujol, A., López, A., Alba, José L. and Villanueva, J.J. "Ridges, Valleys and Hausdorff Based Similarity Measures for Face Description and Matching." Proc. International Workshop on Pattern Recognition and Information Systems, pp. 80-90. Setubal (Portugal), July 2001
8. Pearson, D.E., Hanna, E. and Martinez, K., "Computer-generated cartoons, " Images and Understanding, 46-60. Cambridge University Press, 1990
9. López, A. M., Lumbreras, F., Serrat, J., and Villanueva, J. J., "Evaluation of Methods for Ridge and Valley Detection," IEEE Trans. on PAMI, 21(4), 327-335, 1999
10. Huttenlocher, D.P. et al.,"Comparing images using the hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 3, pp. 850-863, 1993
11. Dubuisson, M.P. and Jain, A.K.,"A modified hausdorff distance for object matching," in Proceedings IEEE International Conference on CVPR, 1995
12. Paglieroni, D., "Distance transforms: Properties and machine vision applications," CVGIP:Graphical models and image processing, vol. 54, no. 1, pp. 56-74, 1992
13. González-Jiménez, D., Alba-Castro J.L., "Frontal Face Authentication through Creaseness-driven Gabor Jets," in Proceedings ICIAR 2004 (part II), pp. 660-667, Porto (Portugal), September/October 2004.
14. The extended xm2vts database. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/
15. Luttin, J. and Maître, G., "Evaluation protocol for the extended M2VTS database (XM2VTSDB)." Technical report RR-21, IDIAP, 1998.
16. The BANCA Database. http://www.ee.surrey.ac.uk/banca.
17. Face Authentication Test on the BANCA Database. http://www.ee.surrey.ac.uk/banca/icba2004/csuresults.html

A One Bit Facial Asymmetry Code (FAC) in Fourier Domain for Human Recognition

Sinjini Mitra[1], Marios Savvides[2], and B.V.K. Vijaya Kumar[2]

[1] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
smitra@stat.cmu.edu
[2] Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA 15213
msavvid@cs.cmu.edu, kumar@ece.cmu.edu

Abstract. The present paper introduces a novel set of biometrics based on facial asymmetry measures in the frequency domain using a compact one-bit representation. A simplistic Hamming distance-type classifier is proposed as a means for matching bit patterns for identification purposes which is more efficient than PCA-based classifiers from storage and computation point of view, and produces equivalent results. A comparison with spatial intensity-based asymmetry measures suggests that our proposed measures are more robust to intra-personal distortions with a misclassification rate of only 4.24% on the standard facial expression database (Cohn-Kanade) consisting of 55 individuals. In addition, a rigorous statistical analysis of the matching algorithm is presented. The role of asymmetry of different face parts (e.g., eyes, mouth, nose) is investigated to determine which regions provide the maximum discrimination among individuals under different expressions.

1 Introduction

Human faces have two kinds of asymmetry - intrinsic and extrinsic. The former is caused by growth, injury and age-related changes, while the latter is affected by viewing orientation and lighting direction. We are however interested in intrinsic asymmetry which is directly related to the individual facial structure while extrinsic asymmetry can be controlled to a large extent or can be pre-processed or normalized. Psychologists have long been interested in the relationship between facial asymmetry, attractiveness and identification. The more asymmetric a face, the less attractive it is and more recognizable ([1], [2]). This indicates the potential significance of asymmetry in automatic face recognition problems.

A commonly accepted notion in computer vision is that human faces are bilaterally symmetric ([3]) and [4] reported no differences whatsoever in recognition rates while using only the right and left halves of the face. However, a well-known fact is that manifesting expressions cause a considerable amount of facial asymmetry, they being more intense on the left side of the face ([5]). Indeed [6] found differences in recognition rates for the two halves of the face under a given facial expression.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 61–70, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Despite extensive studies on facial asymmetry, its use in human identification started in the computer vision community only in 2001 with the seminal work by Liu ([7]), who for the first time showed that certain facial asymmetry measures are efficient human identification tools under expression variations. This was followed by more in-depth studies ([8], [9]) which further investigated the role and locations of different types of asymmetry measures both for human as well as expression classifications. But people have not so far utilized the frequency domain for developing facial asymmetry measures for recognition. This is a natural extension given that there exists much correspondence between the two domains. We explore this in depth in this paper, with a view to developing a computationally and memory efficient biometric for face identification.

The paper is organized as follows. Section 2 describes the dataset used and Section 3 introduces the new asymmetry measures in the frequency domain. Section 4 presents some exploratory feature analysis and Section 5 contains the classification results along with a statistical analysis of the matching results. A discussion appears in Section 6.

2 Data

The dataset used here is a part of the "Cohn-Kanade AU-coded Facial Expression Database" ([10]), consisting of images of 55 individuals expressing three different kinds of emotions - joy, anger and disgust. Each person was asked to express one emotion at a time by starting with a neutral expression and gradually evolving into its peak form. The data thus consists of video clips of people showing an emotion, each clip being broken down into several frames. The raw images are normalized and cenetered using an affine transformation (details included in [8]). Each normalized image is of size 128×128. Some normalized images from our database are shown in Fig. 1. We use a total of 495 frames, which include 3 frames from each emotion for each subject ($55 \times 3 \times 3$). These are chosen from the most neutral (the beginning frame), the most peak (the final frame) and a middle frame in the entire sequence. Such a selection of frames is performed in order to be able to study the effects of extreme expression variations on the face identification routines based on the new biometric in an effective manner.



Fig. 1. Sample images from our database

3 The Frequency Domain

Many signal processing applications in computer engineering involve the frequency-domain representation of signals. The frequency spectrum consists of two components, the *magnitude* and *phase*. In 2D images particularly, the phase component captures more of the image intelligibility than magnitude and hence is very significant for performing image reconstruction ([11]). [12] showed that correlation filters built in the frequency domain can be used for efficient face-based recognition. Recently, the significance of phase has also been used in biometric authentication. [13] proposed correlation filters based only on the phase component of an image, which performed as well as the original filters. Later [14] demonstrated that performing PCA in the frequency domain by eliminating the magnitude spectrum and retaining only the phase not only outperformed spatial domain PCA, but also have attractive features such as illumination tolerance, can handle partial occlusions. All these point out the benefits of considering classification features in the frequency domain for potentially improved results.

Symmetry properties of the Fourier transform are often very useful ([15]). Any sequence $x(n)$ can be expressed as a sum of a *symmetric* part $x_e(n)$ and an *asymmetric* part $x_o(n)$. Specifically, $x(n) = x_e(n) + x_o(n)$, where $x_e(n) = \frac{1}{2}(x(n) + x(-n))$ and $x_o(n) = \frac{1}{2}(x(n) - x(-n))$. When a Fourier transform is performed on a real sequence $x(n)$, $x_e(n)$ transforms to the real part of the Fourier transform and $x_o(n)$ transforms to its imaginary part (Fourier transform of any sequence is generally complex-valued). The Fourier transform of a real symmetric sequence is thus real; that of a real and odd sequence is purely imaginary. Now, since phase is defined as $\theta = \tan^{-1}\left(\frac{I}{R}\right)$, it will be zero in case the imaginary component is zero. In other words, a symmetric sequence gives rise to zero-phase frequency spectrum. These observations therefore imply that the imaginary component of the Fourier transform can be considered as a measure of facial asymmetry in the frequency domain, and also establish a nice relationship between facial asymmetry and the phase component of the frequency domain. Given the role played by both phase and asymmetry in face-based recognition, this presents an opportunity to exploit this correspondence for the development of more refined classification tools. Note that, this holds for 1D sequences and hence we will consider 1D Fourier transforms of row slices of images. The 2D case is more complex and we will not address this issue in this paper.

3.1 Facial Asymmetry Code (FAC)

Following the notion presented in the earlier section, we wish to develop a simple frequency code to represent the asymmetry/symmetry in a frequency and this is done using the real and imaginary part of the Fourier transform. For each frequency x of the Fourier transform of a row slice of an image, we define a set of features as follows:

$$F(x) = \begin{cases} +1, & \text{if } I(x) > R(x) \\ -1, & \text{if } I(x) \leq R(x) \end{cases},$$

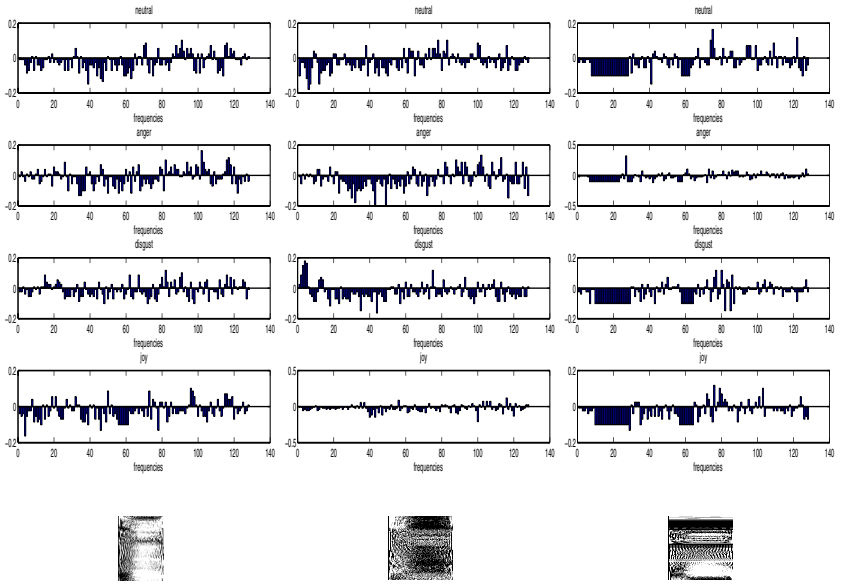
where $I(x)$ and $R(x)$ respectively denote the imaginary and the real part of the Fourier frequency x. Each of our asymmetry features is therefore of one bit per frequency and hence we call them one bit *Facial Asymmetry Code* (or, FAC for short). Note that, these features are very easy to compute and store requiring much less memory than usual quantified measures which are bigger in size. What the features describe are as follows: for a particular frequency, $F(x) = 1$ implies that that frequency has more asymmetry than symmetry, and vice versa if $F(x) = -1$. It is a very simplistic and compact representation of asymmetry of the different facial regions and we will show that these features are capable of devising efficient human recognition routines.

We consider two sets of features: (i) the frequency-wise FAC values - 128×128 matrix, and (ii) FAC computed on Fourier transforms of two-row averages of the original image - 64×128 matrix, denoted *ave FAC*. The averaging over rows for the latter case is done in an attempt to smooth out noise in the image which can possibly create artificial asymmetry artifacts and give misleading results. Averaging over more rows, on the other hand, can lead to over-smoothing and a loss of relevant information. The two-row blocks were selected as optimal after some experimentation.

4 Feature Analysis

For all the exploratory feature analysis, we consider a reduced dimension FAC set constructed as follows: the FAC bits are averaged over each row, so that if $b(x, y)$ denotes the bit at frequency (x, y), we compute $B(x) = \frac{1}{N} \sum_y b(x, y)$ where N denotes the number of columns in each image. This means that if $B(x) > 0$ for a particular row, the features in that row are more asymmetric and if $B(x) < 0$, the features in that row are more symmetric. This feature reduction technique is being used for two reasons. First, this helps us compare our results with those reported in [8] who also employed a similar row-averaging technique for feature analysis. Second, the frequency-wise values are noisy and do not depict a clear picture, whereas the row averages are much easier to study and at the same time provide an useful insight into the nature of the different features and their utility in classification. Figure 2 shows the pattern of variation of FAC for three people while expressing different emotions, framewise and generally over all frames. They give a preliminary but convincing idea that these measures may be helpful in recognizing people in the presence of expression variations owing to the existence of somewhat distinct patterns for each person. This hence constitutes a work parallel to that of [8], in a frequency domain framework instead.

We next studied the discriminative power of these asymmetry measures to determine which parts of the face are actually useful for the identification process. We used a variance ratio-type quantity called the *Augmented Variance Ratio* or AVR, which was also used by ([8]). AVR compares within class and between class variances and at the same time penalizes features whose class means are too close to one another. For a feature F with values S_F in a data set with C total classes, AVR is calculated as



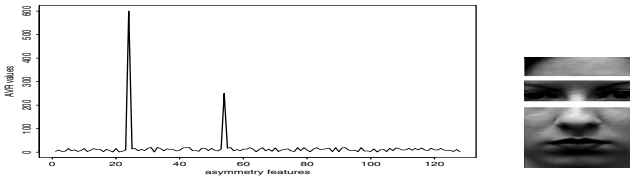
(a) Person 1 (b) Person 2 (c) Person 3

Fig. 2. FAC for the 4 expressions of 3 individuals. For the top 4 rows, +ve values denote more asymmetry and -ve values more symmetry. The features $0 - 128$ range from the forehead to the chin of a face. The last row shows FAC distribution over all images of each person - darker areas show more symmetry and lighter areas show more asymmetry across images of the same person

$$AVR(S_F) = \frac{Var(S_F)}{\frac{1}{C}\sum_{k=1}^{C} \frac{Var_k(S_F)}{min_{j \neq k}(|mean_k(S_F) - mean_j(S_F)|)}} \quad ,$$

where $mean_i(S_F)$ is the mean of the subset of values from feature F belonging to class i. The higher the AVR value of a feature, the more discriminative it is for classification. For our problem, the 55 subjects form the classes ($C = 55$).

Figure 3 shows the AVR values for the row-averaged FAC-based features, which clearly shows that features around forehead region just above the eyes contain the most discriminative information followed by the region around the nose bridge, pertaining to expression-invariant recognition of individuals. The other features have very low AVR values which signify that they are not very useful for human face recognition based on the FAC features. [8], on the other hand, reported that for their spatial asymmetry measures, the nose bridge is the most discriminating facial region for similar recognition tasks, followed by regions around the forehead and chin. Hence there exists some consistency in the location of asymmetry defined in different ways that helps distinguishing people under expression changes.



(a) (b)

Fig. 3. (a) AVR values for the FAC-based features, (b) The white strips on the image denote the regions corresponding to the two peaks in (a)

5 Results

We trained on the neutral frames of the 3 emotions of joy, anger and disgust from all the 55 individuals in the dataset and tested on the peak frames of all the 3 emotions from all the people. Hence this represents an expression-invariant human identification problem, similar to the one reported in [8] which uses an analogous simple measure of facial asymmetry in the spatial domain called *D-face* defined as:

$$D(x, y) = I(x, y) - I'(x, y) \ .$$

I denotes a normalized face and I' its reflected version along the face midline.

Since our features are essentially encoded as bit patterns, it seems natural to use a distance-type metric that is more effective for comparing bit patterns. Once such metric is the popular *Hamming distance* (we will denote it by HD, for short), which gives the count of bits that are different in two patterns. More generally, if two ordered list of items are compared, HD is the number of items that do not match identically. In our case, when comparing two FAC patterns, HD outputs the number of bits in two codes that do not match.

The results appear in Table 1 which show that our proposed HD classifier outperforms spatial D-face, an absolute improvement of $13-14\%$ was observed, and at the same time FAC is much more compact than the D-face representation. Furthermore, HD produced as good classification results as with the individual PCA approach ([12]). One advantage of this method is that it requires training for new images in the database unlike the global PCA method ([16]) in which the re-training and projections have to be done on the entire database each time a new person's data become available. Apart from the impressive results, the HD classifier also has a definite advantage over PCA-based method in that it is computationally much less intensive (involves Boolean exclusive-OR, operation

Table 1. Error rates using the HD classifier on the FAC-based features

Asymmetry features	Misclassification rates
FAC	4.24%
Ave FAC	4.54%
Spatial D-face	17.58%

only) and is much simpler to store than the eigenvectors of PCA which require floating point 32-bit representation. Moreover, only half of these codes need to be stored and used due to the *conjugate hermitian symmetry* arising from purely real sequences in the spatial domain, according to which frequencies are symmetric around the origin (real part is symmetric and imaginary part is odd-symmetric). So, in essence, we are just using half-bit codes in the matching routine. A comparison of the storage requirements shown in Table 2 clearly shows that HD requires upto 64 times less storage space than PCA for operation and even 16 times less storage space than the original normalized images. This alone establishes a firm basis for the utility of the HD classification algorithm based on FAC for performing face recognition in practice. Moreover, it lays the ground for some rigorous statistical analysis which we discuss in the next section.

Table 2. Storage requirements of HD and PCA classifiers for images of different sizes

Actual size	Image storage (bits)	PCA eigenvectors (bits)	HD (bits)
64×64	32768	131072	2048
128×128	131072	524288	8192

5.1 Statistical Analysis

The Hamming distance is computed using a matching of two one bit patterns. Hence if X is the random variable denoting the number of matched bits for a pair of FACs, then assuming that the individual bits are uncorrelated, X follows a *Binomial* distribution with parameters p (probability of a match) and n (total number of bits per image) with the distribution mass function

$$f(x) \;=\; \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n \;.$$

Now, if Y_i denotes the total number of matched bits for person i when matching N_i images of this person, then $Y_i \;=\; \sum_{k=1}^{N_i} X_k^i$, X_k^i is the number of matched bits for the k^{th} image of person i. Then $Y_i \;\sim\; Bin(nN_i, p_i), \quad i = 1, \ldots, 55,$ where p_i is the probability of a matched bit for person i. $p_i = p$ implies that every person has the same probability of match per bit. Note that, HD gives the number of mismatched bits for a pair of FAC, say Z, then $X = n - Z$.

We estimate p_i by the sample proportions of match, which is given by $\hat{p_i} = y_i/nN_i, \quad i = 1, \ldots, 55$. The 95% confidence interval for each p_i is then given by $\hat{p_i} \pm 1.96 \times \hat{\sigma_i}$, where $\hat{\sigma_i} \;=\; \sqrt{\frac{\hat{p_i}(1-\hat{p_i})}{nN_i}}$ using the normal approximation to binomial which is valid since we have a large number of samples. We compute these point and interval estimates for two cases: (i)"genuine", when matching two FAC belonging to the same person, and (ii) "impostor", where two FAC belonging to two different people are matched. This is done since it is reasonable to assume that the probability of bit-matching depends largely on the fact whether the bit patterns belong to the same individual or not. Figure 4 shows

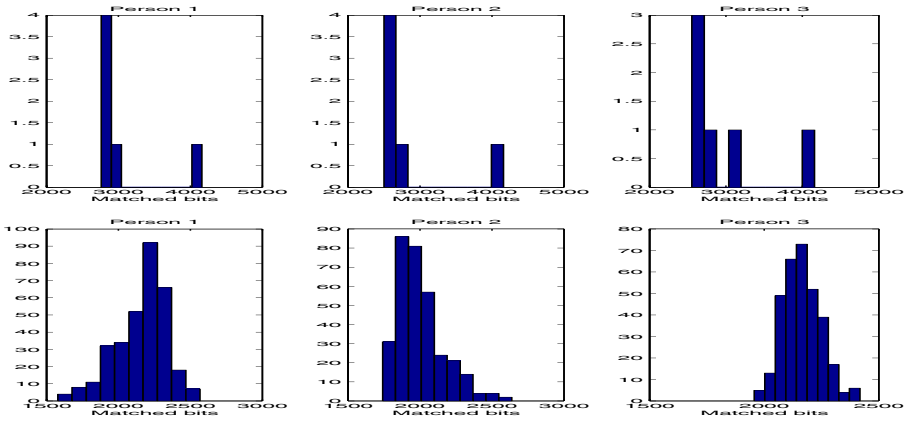
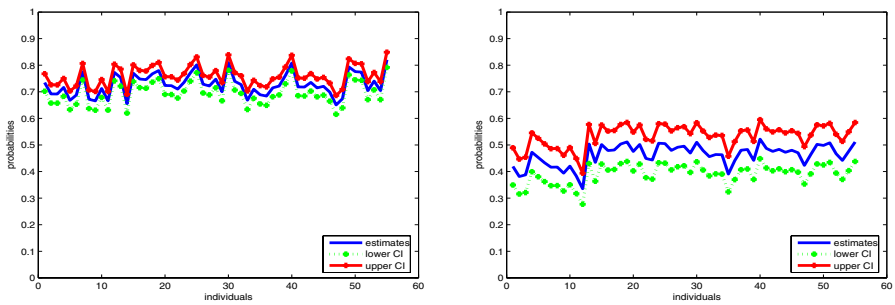


Fig. 4. Histograms of the number of matched bits for 3 people. The top panel is for the (i) genuine cases, and the bottom panel for the (ii) impostor cases

the distributions of Y_i for 3 people in the database. If we assume $p_i = p$, we get $\hat{p} = 0.7276$ for case (i) and $\hat{p} = 0.5625$ for case (ii) with confidence intervals of $(0, 1)$ for both, which is not very useful. This happens due to the fact that the variation among the number of matched bits for all the people is very large which off-sets the confidence intervals. It is thus desirable to form these estimates separately for each person for a more meaningful picture as well as a comparative study across people. Figure 5 shows the sample estimates along with the respective confidence intervals for the probabilities of matches for all the 55 individuals in the database for the two cases. As expected, the estimates for the genuine cases are considerably higher than the impostors ones. However, the upper confidence limits for latter seem a little higher than desirable (greater than 0.5 in some cases). This is attributed to inflated standard errors caused by variation in the impostor probabilities among different people. This happens because some people are more identical looking and hence more likely to be



(a) Genuine (b) Impostor

Fig. 5. The sample estimates of the probabilities of bit-matching ($\hat{p}$), along with the 95% confidence interval for each of the 55 individuals in the database

mistaken for each other than others. One way to rectify this will be to consider impostor probabilities for pairs of people taken at a time. The plots also suggest that people differ in expressing emotions to a considerable extent.

Statistical tests were also conducted to determine if there existed any *significant* difference in the $\hat{p_i}$ values for the two cases. A one-sided two-sample t-test ([17]) gave p-values < 0.0001 which indicated that the genuine cases have a significantly higher probability of bit-matching than the impostor cases. A pairwise comparison further showed that all the 55 people have significantly higher bit-matching probabilities for the genuine cases than the impostor cases, which is what one should expect. A plot of the p-values for the tests appear in Figure 6.

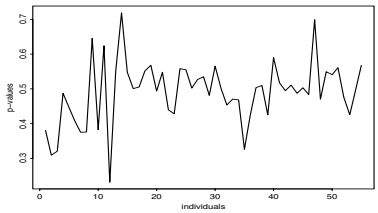


Fig. 6. The p-values for the tests comparing the genuine and impostor probabilities of bit-matching for the 55 individuals

6 Discussion

We have thus shown in this paper that facial asymmetry measures in the frequency domain offer a promising potential as an useful biometric in practice, especially in the presence of expression variations. Our proposed one bit FACs and the HD classifier are very efficient with regard to computation and storage requirements, in fact much more than other known features/classifiers (PCA, D-face) as we have demonstrated. In fact, FAC needs less space than the actual gray-scale intensity images also, thus instead of storing or transmitting those, one can compute their FAC and transmit them. This is very useful for mobile, low-bandwidth communication channels and low-memory devices such as smart-cards and system-on-chip implementations. As far as recognition is concerned, an error rate as low as 4.24% is very impressive and desirable indeed in any situation, especially given that the test images in our case are very different from the training ones. This in turn is very important for recognition routines in practice, for example, in biometric identification applications since surveillance photos captured at airports are expected to be quite diverse with respect to the expressions of an individual's face. Hence any algorithm that can deal with such variations is supposed to be attractive to users.

Moreover, the fact that we observed significant difference in the scores for genuine and impostor cases indicates that our method can be easily adopted to form efficient verification tools. This constitutes our next research direction, along with exploring whether FAC can handle illumination variations as well as it does expression variations and extension to a larger database.

References

1. Thornhill, R., Gangstad, S. W.: Facial attractiveness. Transactions in Cognitive Sciences **3** (1999) 452–460
2. Troje, N. F., Buelthoff, H. H.: How is bilateral symmetry of human faces used for recognition of novel views? Vision Research **38** (1998) 79–89
3. Seitz, S.M., Dyer, C.R.: View morphing. SIGGRAPH (1996) 21–30
4. Gutta, S., Philomin, V., Trajkovic, M.: An investigation into the use of partial-faces for face recognition. In: International Conference on Automatic Face and Gesture Recognition, Washington D.C. (2002) 33–38
5. Borod, J.D., Koff, E., Yecker, S., Santschi, C., Schmidt, J.M.: Facial asymmetry during emotional expression: gender, valence and measurement technique. Psychophysiology **36** (1998) 1209–1215
6. Martinez, A.M.: Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. PAMI **24** (2002) 748–763
7. Liu, Y., Schmidt, K., Cohn, J., Weaver, R.L.: Human facial asymmetry for expression-invariant facial identification. In: Automatic Face and Gesture Recognition. (2002)
8. Liu, Y., Schmidt, K., Cohn, J., Mitra, S.: Facial asymmetry quantification for expression-invariant human identification. CVIU **91** (2003) 138–159
9. Mitra, S., Liu, Y.: Local facial asymmetry for expression classification. In Proceedings of CVPR (2004)
10. Kanade, T., Cohn, J.F., Tian, Y.L.: Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition. (2000) 46–53
11. Hayes, M.H.: The reconstruction of a multidimensional sequence from the phase or magnitude of its fourier transform. ASSP **30** (1982) 140–154
12. Savvides, M., Vijaya Kumar, B.V.K., Khosla, P.: Face verification using correlation filters. In: 3rd IEEE Automatic Identification Advanced Technologies, Tarrytown, NY (2002) 56–61
13. Savvides, M., Kumar, B.V.K.: Eigenphases vs.eigenfaces. ICPR (2004)
14. Savvides, M., Kumar, B.V.K., Khosla, P.K.: Corefaces - robust shift invariant PCA based correlation filter for illumination tolerant face recognition. CVPR (2004)
15. Oppenheim, A.V., Schafer, R.W.: Discrete-time Signal Processing. Prentice Hall, Englewood Cliffs, NJ (1989)
16. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In Proceedings of CVPR (1991)
17. Casella, G., Berger, R.L.: Statistical Inference. 2nd edn. Duxbury (2002)

Face Recognition with the Multiple Constrained Mutual Subspace Method

Masashi Nishiyama[1], Osamu Yamaguchi[1], and Kazuhiro Fukui[2]

[1] Corporate Research & Development, Toshiba Corporation, Japan
{masashi.nishiyama,osamu1.yamaguchi}@toshiba.co.jp
[2] Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, Japan
kfukui@cs.tsukuba.ac.jp

Abstract. In this paper, we propose a novel method named the *Multiple Constrained Mutual Subspace Method* which increases the accuracy of face recognition by introducing a framework provided by ensemble learning. In our method we represent the set of patterns as a low-dimensional subspace, and calculate the similarity between an input subspace and a reference subspace, representing learnt identity. To extract effective features for identification both subspaces are projected onto multiple constraint subspaces. For generating constraint subspaces we apply ensemble learning algorithms, i.e. Bagging and Boosting. Through experimental results we show the effectiveness of our method.

1 Introduction

Recently, many face identification methods that perform recognition from a set of patterns instead of a single pattern have been proposed[1–5]. Since these methods are able to cope with variation in appearance under varying pose, a robust face identification application can be built.

To identify faces using a set of patterns, we have previously proposed the *Mutual Subspace Method* (MSM)[1]. In MSM, a set of patterns is represented as a low-dimensional subspace. To compare the input subspace with the reference subspace representing learnt identity, we calculate their similarity which is defined by the minimum angle between the input subspace and the reference subspace. These subspaces are generated using principal component analysis (PCA).

To improve the performance of MSM we have extended this method to the *Constrained Mutual Subspace Method* (CMSM)[5]. In CMSM, to extract effective features for identification, we project the input subspace and the reference subspace onto the constraint subspace, as shown in Fig. 1. Through this projection we can extract features that are insensitive to varying facial pose and illumination, while remaining sensitive to change in individual appearance. Using CMSM Sato et al.[6] illustrated the effectiveness in a practical security system, while Kozakaya et al.[7] demonstrated an implementation of a real-time system on an image processing LSI chip.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 71–80, 2005.
© Springer-Verlag Berlin Heidelberg 2005

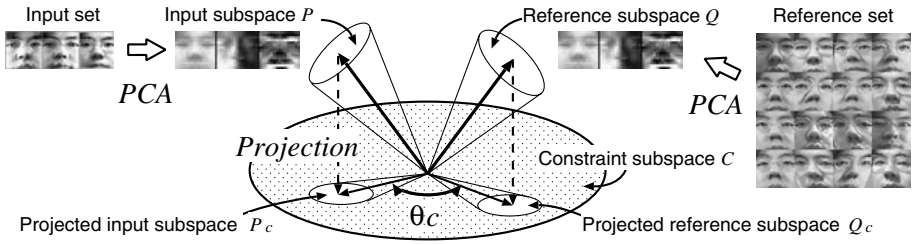


Fig. 1. Concept of CMSM. The input subspace and the reference subspace are generated from the set of patterns. Then, both subspaces are projected onto the constraint subspace. Finally, the similarity is determined with the angle θ_C

Although CMSM is effective, a large number of training patterns are required for generating the constraint subspace. Since variation in appearance is large under varying pose and illumination, it is difficult to acquire training patterns which sufficiently represent these variations. Therefore, we need a method of generating the constraint subspace which yields high performance from a limited number of acquired training patterns. In the field of machine learning, ensemble learning has been proposed[8, 9]. Ensemble learning derives recognition performance by combining hypotheses obtained from given training samples. Wang et al.[10] applied ensemble learning to face identification based on Linear Discriminant Analysis and demonstrated that they obtain high performance using only a few training patterns.

In this paper we propose a new method which generates multiple constraint subspaces by introducing the framework provided by ensemble learning. Using these constraint subspaces, we extend CMSM to the *Multiple Constrained Mutual Subspace Method* (MCMSM). In MCMSM, the input subspace and the reference subspace are projected onto each constraint subspace, and the similarity is calculated on each constraint subspace. By combining these similarities we finally determine the combined similarity as shown in Fig. 2. To generate constraint subspaces, we propose two approaches in which we apply the framework provided by ensemble learning.

This paper is organized as follows. First, we describe the method for applying MCMSM to face identification in section 2. Next, we describe two approaches for generating constraint subspaces in section 3. Then, we demonstrate the effectiveness of our method using MCMSM by experiments in section 4.

2 Identification Using MCMSM

2.1 Algorithm for Face Identification

In this section, we describe the procedure of our face identification method. First, an input set of face patterns is obtained from a video sequence. We locate the face pattern from the positions of the pupils and the nostrils obtained automatically by the method described in [1, 7]. The pattern is transformed to a vector by

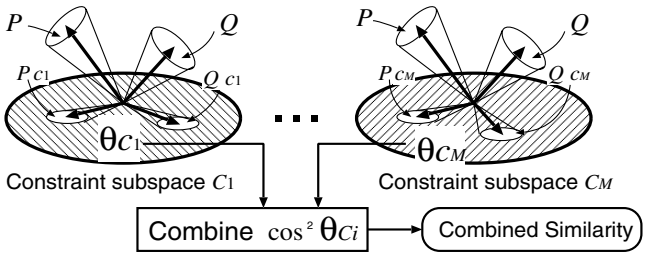


Fig. 2. Concept of MCMSM. The input subspace P and the reference subspace Q are projected onto each constraint subspace Ci. By combining M similarities($cos^2\theta_{Ci}$), which are calculated on Ci, we finally determine the combined similarity

raster-scanning of the pattern, and we apply PCA to the vectors to generate an input subspace. Let $\mathbf{x}$ be a vector and N_V be the number of the vectors, the basis vectors of the input subspace are the eigenvectors of the correlation matrix $\mathbf{A} = 1/N_V \sum_{i=1}^{N_V} \mathbf{x}\mathbf{x}^T$[12].

To compare the input subspace with the reference subspace, registered in a database for each individual, we calculate their combined similarity. This combined similarity is determined with similarities calculated on each constraint subspace. The identified person is determined as corresponding to the reference subspace of the highest combined similarity. The details of each process are described in the following section.

2.2 Projection onto Constraint Subspaces

To project the input subspace P onto M constraint subspaces, we carry out the following steps:

1. Project basis vectors of P onto the i-th constraint subspace C_i.
2. Normalize the length of each projected vector.
3. Apply Gram-Schmidt orthogonalization to the normalized vectors.

The orthogonal normalized vectors are basis vectors of the projected input subspace P_{C_i}. This procedure is repeated M times for each constraint subspace. The projected reference subspace Q_{C_i} can be obtained with the same procedure.

2.3 Calculation of the Similarity on Each Constraint Subspace

We define similarity S_{C_i} between the subspace P_{C_i} and the subspace Q_{C_i} as

$$S_{C_i} = \cos^2 \theta_{C_i} \ , \tag{1}$$

where θ_{C_i} represents the canonical angle between P_{C_i} and Q_{C_i}. The canonical angle is calculated using MSM[1]. The similarity S_{C_i} can be obtained from the largest eigenvalue λ_{max} of $\mathbf{X}$ using

$$\mathbf{X}\mathbf{a} = \lambda \mathbf{a} \ , \tag{2}$$

$$\mathbf{X} = (x_{mn}) \quad m, n = 1 \ldots N \ , \ and \tag{3}$$

$$x_{mn} = \sum_{l=1}^{N} (\psi_m, \phi_l)(\phi_l, \psi_n) \ , \tag{4}$$

where ψ_m is the m-th basis vector of subspace P_{C_i}; ϕ_l is the l-th basis vector of subspace Q_{C_i}; (ψ_m, ϕ_l) is the inner product of ψ_m and ϕ_l; N is the dimension of P_{C_i} and Q_{C_i}. The similarity S_{C_i}equals λ_{max}. If the input subspace and the reference subspace are identical, the canonical angle θ_{C_i} equals 0.

2.4 Combine Similarities

To combine similarities obtained on each constraint subspace, we define the combined similarity S_T as follows:

$$S_T = \sum_{i=1}^{M} \alpha_i S_{C_i} \ , \tag{5}$$

where M is the number of the constraint subspaces; α_i is the i-th coefficient of C_i; S_{C_i} is the similarity between P_{C_i} and Q_{C_i} projected onto C_i.

3 Generation of Multiple Constraint Subspaces with Ensemble Learning

In this section, we explain the method of generating a single constraint subspace for CMSM[5]. Next, we describe two approaches for generating multiple constraint subspaces with ensemble learning for MCMSM.

3.1 Generation of a Single Constraint Subspace

To allow for the variation in appearance for each individual, we acquire the sets of patterns while changing illumination and pose for L individuals. The variation of the patterns is represented as a subspace for each individual. We call this subspace the *training subspace*.

To generate a constraint subspace which separates the training subspaces by projection, we calculate eigenvectors using

$$(\mathbf{P}_1 + \mathbf{P}_2 + \ldots + \mathbf{P}_L)\mathbf{a} = \lambda \mathbf{a} \ , \tag{6}$$

$$\mathbf{P}_j = \sum_{k=1}^{N_B} \psi_{jk} \psi_{jk}^T \ , \tag{7}$$

where $\mathbf{P}_j$ is the projection matrix of the j-th training subspace; N_B is the dimension of training subspace; ψ_{jk} is the k-th basis vector of the j-th training subspace. The eigenvectors, selected in ascending order, are the basis vectors of the constraint subspace. For details of CMSM see [5, 7].

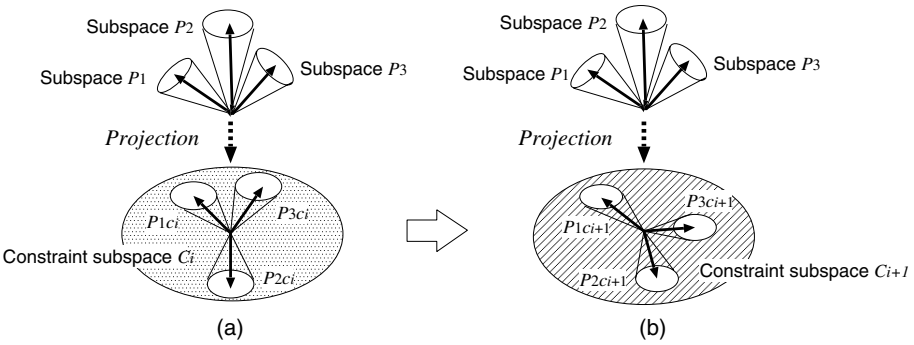


Fig. 3. Concept of the method for generating constraint subspaces using Boosting

3.2 Generation of Constraint Subspaces with Bagging

To generate constraint subspaces, we use Bagging[8], which is based on an ensemble learning algorithm. Multiple classifiers are constructed using random sampling in Bagging. To apply this framework to generating constraint subspaces, we randomly select $L'(< L)$ subspaces from L training subspaces. Each constraint subspace is generated independently using selected training subspaces.

Algorithm Using Bagging
To summarize: we generate M constraint subspaces by the following steps:

1. Select L' training subspaces randomly without replacement.
2. Generate a constraint subspace using selected L' training subspaces in eq.(6).
3. Until M constraint subspaces are generated, go to 1.

3.3 Generation of Constraint Subspaces with Boosting

In another method of generating constraint subspaces, we use Boosting[9]. Each classifier is constructed sequentially by reweighting the training patterns in Boosting. The current weight is given to training patterns which were classified incorrectly in the previous constructed classifier.

In applying this framework to generate constraint subspaces we must define how to calculate the weight for each training subspace. Consider similarities between training subspaces on the constraint subspace. As shown in Fig. 3(a), when the projected training subspace P_{1C_i} and the projected training subspace P_{3C_i} are similar on the constraint subspace C_i, the likelihood of the false identification is increased for these training subspaces. To cope with this problem, we aim to separate $P_{1C_{i+1}}$ and $P_{3C_{i+1}}$ on C_{i+1} as shown in Fig. 3(b). To achieve this, we generate C_{i+1} by assigning large weight to P_{1C_i} and P_{3C_i}, thereby increasing their importance and decreasing the remaining error.

Algorithm Using Boosting
To summarize: we generate M constraint subspaces by the following steps:

1. Define the initial weight $W_1(j)$.
2. Generate the i-th constraint subspace C_i using i-th weight $W_i(j)$ and the projection matrix $\mathbf{P}_j$ of the j-th training subspace as

$$(W_i(1)\mathbf{P}_1 + \ldots + W_i(L)\mathbf{P}_L)\mathbf{a} = \lambda\mathbf{a} \ . \tag{8}$$

3. Calculate the next weight $W_{i+1}(j)$ using C_i.
4. Until M constraint subspaces are generated, go to 2.

The weight $W_{i+1}(j)$ is calculated using

$$W_{i+1}(j) = \frac{S'_j}{\sum_{j=1}^{L} S'_j} \quad and \tag{9}$$

$$S'_j = \sum_{j' \neq j}^{L} \beta_{jj'} \ , \tag{10}$$

where $\beta_{jj'}$ equals $\theta_{C_{ijj'}}$; $\theta_{C_{ijj'}}$ is the angle between P_j and $P_{j'}$ projected onto the C_i. To generate a constraint subspace using only similar training subspaces, we can set threshold T to be

$$\beta_{jj'} = \begin{cases} \cos^2\theta_{C_{ijj'}} & T \leq \cos^2\theta_{C_{ijj'}} \\ 0 & T > \cos^2\theta_{C_{ijj'}} \end{cases} \ . \tag{11}$$

4 Empirical Evaluation

4.1 Performance for Varying Illumination

To illustrate the performance of our face identification method, the lighting condition was changed dynamically. We collected a video sequence at 5 frames per second for each person under each lighting condition. We set 10 difference lighting conditions using 7 light sources (A-G); see Fig. 4. A image of the set of each lighting condition is shown in Fig. 5(a). A video sequence consisted of 140 face images which were captured in arbitrary facial pose, e.g. translation, yaw and pitch. The size of each image was 240×320 pixels and 50 different individuals' data were collected. From each image a 30×30 pixel pattern, as shown in Fig. 5(b), was extracted. This pattern was histogram equalized, resized to 15×15 pixels by subsampling, a vertical gradient operator was applied, and finally the pattern was transformed to a $15 \times (15-1) = 210$-dimensional vector.

We divided the data into two groups that each consisted of 25 individuals' patterns. The first group was used for identification and the second for generating constraint subspaces. In the first group, we divided the patterns into input sets and reference sets for each person. An input set consisted of 10 patterns for each lighting condition and a reference set consisted of 70 patterns for each lighting condition. We used 7 input sets per person for each lighting condition. In the second group, to learn variation of patterns under varying illuminations, a

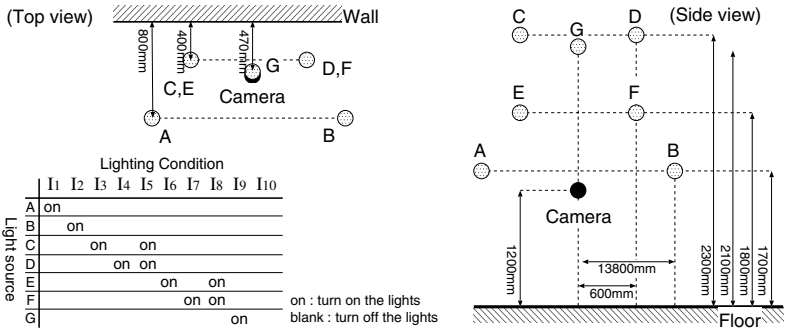
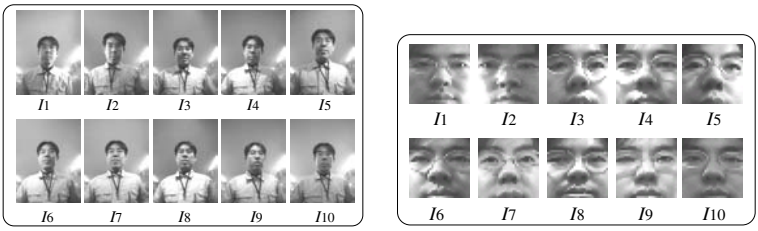


Fig. 4. Setting of light sources and camera



(a) captured images (b) face patterns

Fig. 5. Examples of the captured images and the face patterns

training subspace was generated by using all lighting condition patterns. A set of training patterns consisted of 140 patterns per lighting condition. We generated 25 training subspaces.

We compared the performance of MCMSM with those of conventional methods.

(a)Nearest Neighbor (NN)
The similarity was determined with the smallest Euclidean distance between the pattern in the input set and the pattern in the reference set.

(b)Subspace Method[11] (SM)
The similarity was determined using the average of the angle calculated between each pattern of the input set and the reference subspace. We generated the 40-dimensional reference subspace for each reference set.

(c)Mutual Subspace Method[1] (MSM)
The similarity was determined using the angle between the input subspace and the reference subspace. We generated the 7-dimensional input subspace for each input set and the 7-dimensional reference subspace for each reference set.

(d)Constrained MSM[5] (CMSM)
The similarity was determined with MSM after projection onto a single constraint subspace. The constraint subspace was generated with $L = 25$ training subspaces. We set the dimension of the training subspace to $N_B = 30$, and the dimension of the constraint subspace to $N_C = 170$.

(e)Multiple CMSM with Bagging (MCMSM-Bagging)
The similarity was determined with MSM after projecting onto multiple constraint subspaces. Each constraint subspace was generated from $L' = 8$ training subspaces selected randomly. We used $M = 10$ constraint subspaces. The coefficient of the combining was $\alpha_i = 1/10$.
(f)Multiple CMSM with Boosting (MCMSM-Boosting)
The similarity was determined with MSM after projecting onto multiple constraint subspaces. Each constraint subspace was generated from weighted training subspaces. We used $M = 10$ constraint subspaces. The initial weight $W_1(j)(j = 1 \ldots 25)$ was $1/25$, the threshold T was $3.5\sigma_i$, and σ_i was the standard deviation of similarities which were calculated between training subspaces. The coefficient α_i was $1/10$.

Table 1 shows the evaluation result of each method in terms of the correct match rate (CMR) and the equal error rate (EER). CMR is the probability that an input set of the right person is correctly accepted. EER is the probability that the false acceptance rate (FAR) equals the false rejection rate (FRR). We can see that the methods (e) and (f) using MCMSM are superior to (a)-(d) with regard to CMR and EER. Figure 6 shows the receiver operating characteristic (ROC) curves, which indicate FAR and FRR of each method. The superiority of MCMSM (e) and (f) is also apparent from this.

Table 1. Experimental results under varying illumination (25 registered persons)

	Method	CMR(%)	EER(%)
(a)	NN	95.4	23.9
(b)	SM	95.4	12.9
(c)	MSM	95.4	9.8
(d)	CMSM	95.4	5.0
(e)	MCMSM-Bagging	98.2	4.0
(f)	MCMSM-Boosting	98.6	3.9

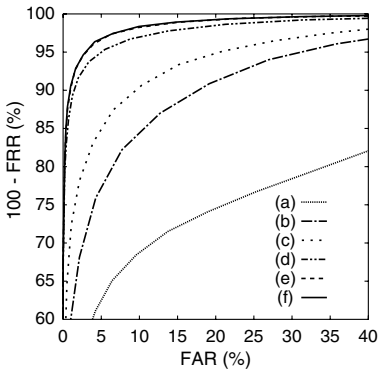


Fig. 6. ROC curves (25 registered persons, varying illumination)

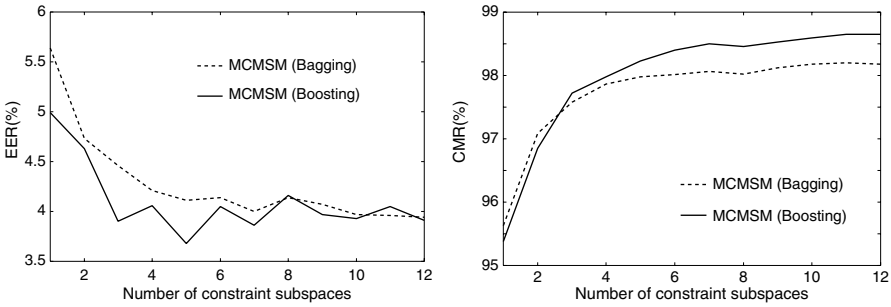


Fig. 7. Identification performance with increasing the number of constraint subspaces

Figure 7 shows the performance of MCMSM versus the number of constraint subspaces. We can see improved performance for both generating methods as the number of constraint subspaces increased.

4.2 Performance Assessment on a Large Database

To evaluate performance for a large number of individuals, we collected a total of 1000 input sets for 500 people. The facial pose changed irregularly at each input set although the lighting conditions were almost uniform. As before the dimension of the vector was 210. An input set consisted of 15 patterns and a reference set consisted of 125 patterns. We compared the performance of three methods: (i)CMSM, (ii)MCMSM-Bagging and (iii)MCMSM-Boosting. In these methods, we used the 7-dimensional input subspace and the 7-dimensional reference subspace. We used 500 training subspaces for generating the constraint subspace. The training subspace was generated with the reference set. The dimension of the training subspace was $N_B = 10$, and the dimension of the constraint subspace was $N_C = 170$. In (i), we used a single constraint subspace generated with 500 training subspaces. In (ii), we used $M = 10$ constraint subspaces. Each constraint subspace was generated from $L' = 30$ training subspaces. The coefficient α_i was 1/10. In (iii), we used $M = 10$ constraint subspaces. The initial weight $W_1(j)$ was 1/500, the threshold T was $5\sigma_i$, and the coefficient α_i was 1/10.

Table 2 shows the evaluation result of each method. We can see that the methods using MCMSM are superior to that using CMSM.

Table 2. Experimental results (500 registered persons)

	Method	CMR (%)	EER (%)
(i)	CMSM	94.7	2.3
(ii)	MCMSM-Bagging	96.2	1.6
(iii)	MCMSM-Boosting	96.8	1.6

5 Conclusion

This paper presented the *Multiple Constrained Mutual Subspace Method* in which we applied ensemble learning to the *Constrained Mutual Subspace Method*. To extract effective features for face identification, we project the input subspace and the reference subspace onto multiple constraint subspaces. In the experiment we obtained high performance compared with projecting onto a single constraint subspace. To generate constraint subspaces, we apply the framework provided by ensemble learning, i.e. Bagging, Boosting. We evaluated the algorithms on a database of varying illumination and a database with 500 individuals. The effectiveness of MCMSM is demonstrated on both databases.

References

1. Yamaguchi, O., Fukui, K., and Maeda, K.: Face recognition using temporal image sequence. Proceedings IEEE Third International Conference on Automatic Face and Gesture Recognition, (1998) 318-323
2. Shakhnarovich, G., Fisher, J.W., and Darrell, T.: Face Recognition from Long-Term Observations. Proceedings of European Conference on Computer Vision, (2002) 851-868
3. Wolf, L., and Shashua, A.: Learning over Sets using Kernel Principal Angles. Journal of Machine Learning Research, 4:913-931, (2003) 851-868
4. Arandjelovic, O., and Cipolla, R.: Face Recognition from Image Sets using Robust Kernel Resistor-Average Distance. The First IEEE Workshop on Face Processing in Video, (2004)
5. Fukui, K., and Yamaguchi, O.: Face Recognition Using Multi-viewpoint Patterns for Robot Vision. 11th International Symposium of Robotics Research, (2003) 192-201
6. Sato, T., Sukegawa, H., Yokoi, K., Dobashi, H., Ogata, J., and Okazaki, A.,: "FacePass" – Development of a Face-Recognition Security System Unaffected by Entrant's Stance. The Institute of Image Information and Television Engineers Transactions, Vol.56, No. 7, (2002), 1111-1117, (in Japanese).
7. Kozakaya, T., and Nakai, H.: Development of a Face Recognition System on an Image Processing LSI chip. The First IEEE Workshop on Face Processing in Video, (2004)
8. Breiman, L.: Bagging Predictors. Machine Learning, Vol. 24, No. 2, (1996) 123-140
9. Freund, Y., and Schapire, R.E.: A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences, Vol.55, No.1, (1997) 119-139
10. Wang, X., and Tang, X.: Random Sampling LDA for Face Recognition. Proceedings IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, (2004) 259-265
11. Watanabe, S., and Pakvasa, N.: Subspace method of pattern recognition. Proceedings of the 1st International Joint Conference on Pattern Recognition (1973) 25-32
12. Oja, E.: Subspace Methods of Pattern Recognition. Research Studies Press, England, (1983)

A Flexible Object Model for Recognising and Synthesising Facial Expressions

Andreas Tewes, Rolf P. Würtz, and Christoph von der Malsburg

Ruhr-Universität, Institut für Neuroinformatik, D–44780 Bochum, Germany

Abstract. We here introduce the Flexible Object Model to represent objects with structured deformation, such as the human face under variable expression. The model represents object shape and texture separately and extracts a data parameterisation autonomously from image sequences after initialisation by a single hand-labeled model graph. We apply the model to the representation, recognition and reconstruction of nine different facial expressions. After training, the model is capable of automatically finding facial landmarks, extracting deformation parameters and reconstructing faces in any of the learned expressions.

1 Introduction

Elastic matching of graphs labeled with Gabor wavelet features (EGM) [1] has proved a very successful basis for invariant object recognition, even when spatial deformation is involved as with face recognition under small changes in pose or expression. According to that concept, variation due to position, scale and in-plane orientation can be dealt with precisely, but intrinsic image deformations are not actively modeled and can only passively be followed. This leads to limited discriminatory power during recognition and precludes the possibility to reconstruct images from model data. Facial image deformations due to pose or expression are highly structured and should be represented by a parameterised model. To this end we have developed a Flexible Object Model (FOM). It continues to use elastic graphs to represent objects in individual images but parameterises these graphs, treating them as functions of pose and expression parameters. In this paper we present the FOM in general and apply it in chapter 3 to the description of the human face under nine different expressions. We demonstrate the power of the model by matching and reconstructing faces in a person-independent way. We conclude by discussing possible applications, among them improved facial recognition under variable expression.

2 The Flexible Object Model

The FOM, using Gabor wavelet-labeled graphs as fundamental data structure, distinguishes object shape (represented by the spatial arrangement of landmarks) from texture (represented by Gabor jets attached to the landmarks). While deformation of shape is described in a parameterised way relative to a reference model, the interrelationship between shape deformation and texture is characterised using a linear function mapping the former onto the latter. The FOM

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 81–90, 2005.
© Springer-Verlag Berlin Heidelberg 2005

therefore also includes mappings between shape deformation and texture, an idea developed earlier in our lab [2, 3]. Both the variations (of shape and texture) and the mappings between them are extracted by statistical procedures from video frame sequences for one or several persons performing different facial gestures. Compared to the concept of *Active Appearance Models*, which describes shape and texture variations using either one common set of parameters or one set for each [4], in the context of FOM only the shape variation is learned in a parameterised way while the texture is assumed to be fully determined by a given shape and map. Finally also the matching process, which uses the concept of EGM and is described in section 4 in detail, differs from the *Fitting* process in the context of AAMs.

2.1 Data Collection

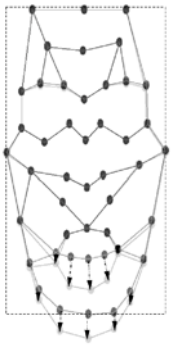
We used sample material collected by Hai Hong [3]. The sequences were taken under fairly controlled lighting conditions and in frontal pose. In each sequence the subject performs one of a number of facial gestures, starting and ending with neutral expression. The gestures were selected for ease of performance (shunning the difficulty of expressing emotional states) and attempting to cover the whole range of facial movements. In this study we have used only a subset of 9 of the 23 gestures originally collected [3] for each person.

We initialise the process of extracting model graphs from the frames of a sequence by manually locating the nodes of the standard graph over facial landmarks in the first frame. The system then automatically tracks these nodes from frame to frame with a method based on the phases of Gabor wavelets [5]. The link structure of the graphs is kept constant throughout. For the sake of scale invariance, the size of the reference graph is noted in terms of a bounding box and node displacement in x- and y-direction is measured relative to the width and height of that box, respectively. To encode local image texture, responses of several Gabor kernels differing in scale and orientation [1] were extracted at the landmark positions in each frame. We treat a set of 300 Gabor responses (real and imaginary part, 15 orientations $[\phi_{min} = 0, \phi_{max} = \frac{14}{15}\pi]$ and 10 scales $[k_{min} = 2^{-\frac{11}{2}}\pi, k_{max} = \frac{1}{2}\pi]$) as one real-valued vector, called Gabor jet.

For each frame, the normalised shift vectors of landmarks relative to the first frame as well as the Gabor jets at the node positions are noted. They together form the raw input data, see (1). The models of section 3 were created for individuals, the models in sections 4 and 5 were formed using video sequences for several persons. Figure 1, left side, shows two facial graphs superimposed on each other. The graph with black nodes represents the reference shape (first frame) while the one with grey nodes belongs to a deformation (which in this case obviously only affects mouth and chin).

2.2 Model Formation

To construct the FOM as a parameterised model of graph deformation, the raw data extracted from several video sequences are merged using *Principal Compo-*



$$\Longleftrightarrow \quad \begin{pmatrix} \frac{\Delta x_1}{w} \\ \frac{\Delta y_1}{h} \\ \vdots \\ \frac{\Delta x_N}{w} \\ \frac{\Delta y_N}{h} \end{pmatrix}$$

Fig. 1. Deformation from reference shape normalised by the width w and height h of the bounding box. N denotes the number of nodes while Δx_i and Δy_i indicates the displacement of node i along x- and y-direction

nent Analysis (PCA) [6] and a *Neural Gas* (NG) [7]. While the latter is suitable for forming sparse representations of the extracted deformations and for classification purposes, PCA is important for data compression and is particularly interesting for interpolating and extrapolating the deformations present in the samples. By this, different deformations which do not not occur simultaneously in the sample sequences can be superimposed, as illustrated in figure 4. In addition we are working with *Principal Curves* [8] to describe smooth transitions, although we don't elaborate on that here.

To represent our raw data we use the following notation. If the number of video frames and raw graphs is M we form the matrices

$$\underline{\underline{D}} := (\boldsymbol{d_1} \dots \boldsymbol{d_M}) ; \quad \underline{\underline{F^i}} := \left(\boldsymbol{j_1^i} \dots \boldsymbol{j_M^i}\right), \tag{1}$$

where the column vector $\boldsymbol{d}$ denotes the deformation as introduced in figure 1 and the column vector $\boldsymbol{j^i}$ indicates the feature vector belonging to the node with index i. Using PCA, we can now construct the following quantities

$$< \boldsymbol{D} > \equiv \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{d}_m \tag{2}$$

$$< \boldsymbol{F^i} > \equiv \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{j_m^i} \tag{3}$$

$$\underline{\underline{P}} := (\boldsymbol{P_1} \dots \boldsymbol{P_L}) \equiv \text{Principal Deformations} \tag{4}$$

$$\underline{\underline{Q^i}} := \left(\boldsymbol{Q_1^i} \dots \boldsymbol{Q_K^i}\right) \equiv \text{Principal Features at node } i \tag{5}$$

$$\underline{\underline{\tilde{D}}} := \underline{\underline{D}} - < \boldsymbol{D} > \underbrace{(1 \dots 1)}_{M \text{ times}} \equiv \text{Mean-Free Deformations} \tag{6}$$

$$\underline{\underline{\tilde{F}^i}} := \underline{\underline{F^i}} - < \boldsymbol{F^i} > \underbrace{(1 \dots 1)}_{M \text{ times}} \equiv \text{Mean-Free Features of node } i \tag{7}$$

where all vectors are taken as column vectors. To reduce the data dimensionality we use only the first **L** principal components to describe graph deformation and the first **K** principal components for the Gabor jets, respectively. Throughout this paper we have set $L = 7$ and $K = 20$, values that proved sufficient to reproduce the original data with little error.

Shape deformation is always accompanied by changing texture. We make the simple assumption of a linear mapping between the shape deformation and the feature vectors (or rather their mean-free versions), and see that assumption justified by our numerical results, see chapter 3. Using the matrices $\underline{\underline{A^i}}$ (one matrix per node) we can express and estimate this relationship as follows,

$$\underline{\underline{A}}^i \underline{\underline{P}}^T \underline{\tilde{D}} \stackrel{!}{=} (\underline{\underline{Q}}^i)^T \underline{\tilde{F}}^i \quad \Rightarrow \quad \underline{\underline{A}}^i \approx (\underline{\underline{Q}}^i)^T \underline{\tilde{F}}^i \left(\underline{\underline{P}}^T \underline{\tilde{D}} \right)^+, \tag{8}$$

where + indicates the Moore-Penrose inverse [9] of the term in brackets. By using homogeneous coordinates it is possible to squeeze all necessary operations into one matrix that maps a given deformation immediately onto the feature vector. This is important because it accelerates the computation and therefore makes it more suitable for the matching tasks introduced in chapter 4.

3 Flexible Model for Synthesising Facial Expressions

In this section we demonstrate the ability of the FOM to synthesise images of varying facial expression. To this end we have created a person-specific FOM, using as data nine video sequences with nine different facial expressions (each containing between 30 and 70 frames). Sample frames are shown in figure 2.

Figure 3 shows three sample frames, taken from the same sequence, with tracked landmarks.

After collecting the data from all nine sequences, we perform the PCA of steps (4) and (5), and estimate the shape-to-texture mappings according to (8). To demonstrate the resulting FOM we chose two of the principal components, added them with variable amplitude to the mean deformation (which is near to the neutral expression) and show in figure 4 reconstructions of the resulting data models. Reconstructions were obtained by the method of [10]. In the bottom row of the figure the PC amplitude runs from one negative standard deviation on the left through zero in the middle to one positive standard deviation on the right. The middle column shows the effect of another PC for positive amplitudes. Three of the gestures shown in figure 2 can be recognised among the reconstructions in the middle columns and bottom row. The diagonals of the figure were formed by superposition of the two PCs and show gestures not present in the input sequences, demonstrating the extrapolation alluded to above.

In the next section we will need a discrete set of "canonical" facial deformations. To this end we use a Neural Gas [7] for clustering and apply the following procedure. From each frame we obtain a shape deformation vector **d**. This we project into the subspace of the first $L = 7$ principal components. These shape vectors are clustered by a neural gas of 9 neurons, each neuron corresponding



Fig. 2. Facial Gestures shown at maximal extent



Fig. 3. Autonomously Tracked Landmarks within a gesture sequence

to a cluster. Figure 5 shows the deformed face graphs for the 9 neurons or clusters. From each neuron's deformation $\mathbf{d}$ we obtain Gabor jets by applying the matrices $\underline{\underline{A}}^i$ and reconstruct a facial image, shown for the 9 clusters or canonical gestures in figure 6.

4 Landmark Finding

Landmark finding, that is, the establishment of precise point correspondences between facial images and generic facial models, is a critical step for facial processing. It is difficult to achieve, especially in the presence of facial deformation. *Passive* mechanisms, such as classical elastic graph matching [1] have to be permissive in terms of deviations of image texture and shape relative to the model and thus lose reliability quickly beyond small deformations. The problem can only be solved by *actively* modeling the changes in texture and shape observed in images. For this purpose we here employ a FOM. For greater robustness we

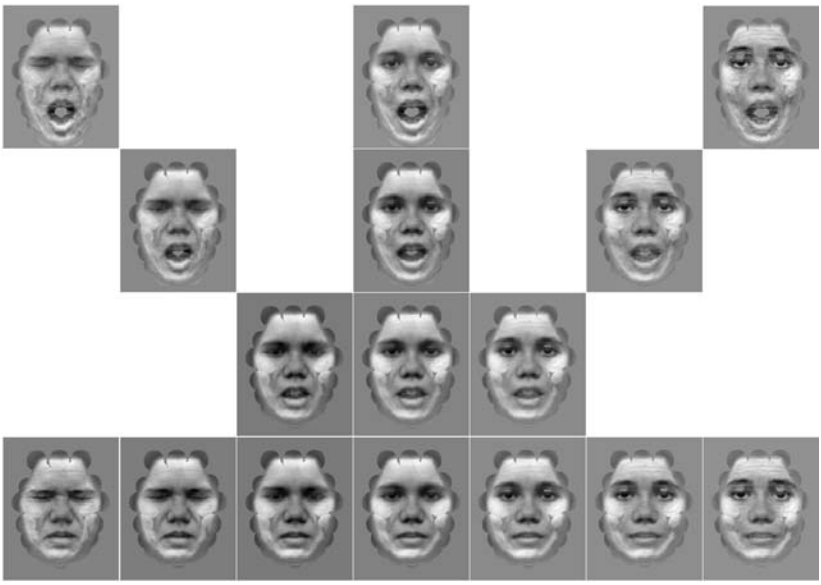


Fig. 4. Synthesised facial expressions using the first (shown vertically) and fourth (shown horizontally) Principal Deformation as well as superpositions

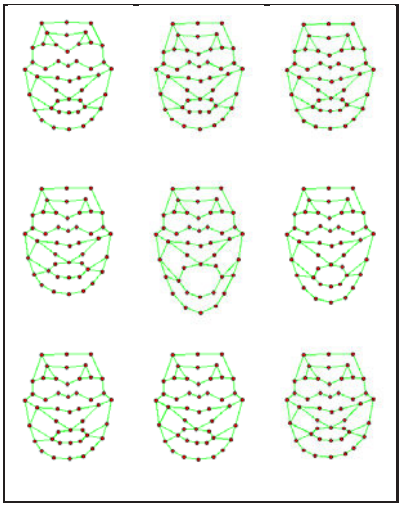


Fig. 5. Shape deformations as per Neural Gas. Expressions are shown corresponding to figure 2

have trained it on four different persons (where we used the total number of sequences collected from all persons while each person contributes a data amount as described in section 3).

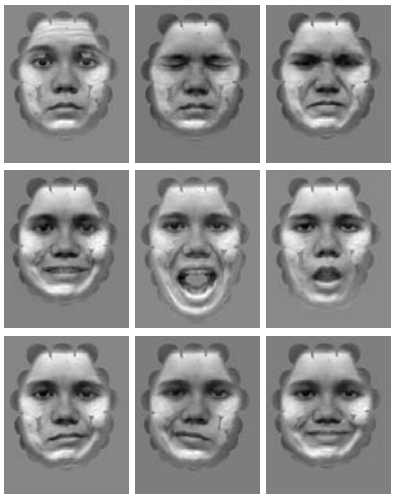


Fig. 6. Synthesised images using shape deformations as shown in figure 5

Our test images display facial gestures of persons not contained in the data set used for training the FOM. We first find the face to some precision with the help of a bunch graph [1] constructed out of frontal neutral-expression images for six persons (again different from the test persons). After suppressing the background of the image outside the convex hull of the bunch graph by Gaussian smoothing we replace the bunch graph by the graph of the FOM and improve the match by separate scale moves in vertical and horizontal directions using the reference shape. Starting from this reference graph, we now apply the nine "canonical" gesture deformations trained by the methods of the last section on four persons (each with the amplitude represented by the trained neurons) and pick the best-matching gesture. Figure 7 shows examples for six different facial expressions. The first and third column show test images with suppressed background and superimposed best-matching graph, each image accompanied on its right by a reconstruction from the 4-person FOM in the best-matching expression.

In addition to accurate landmark finding in spite of image deformation the system can be used to identify the gesture displayed in the image. Using several persons to construct the FOM increased the robustness of the model for person-independent matching (just as the bunch graph increases the robustness of face finding), and in addition handled personal differences in the reference persons' performance of gestures (although the gestures were originally selected for ease of performance [3]).

5 Correction of Facial Expression

An important application of our FOM will be face recognition. Even for collaborating subjects, variation in facial expression cannot be totally avoided and passive methods of dealing with it are compromising accuracy. What is required

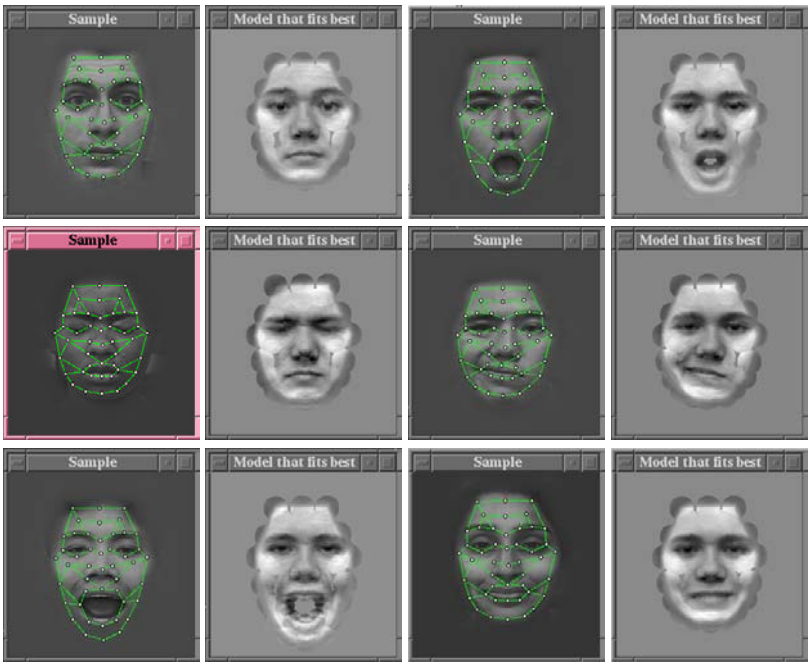


Fig. 7. FOM Matching for six different facial expressions. The sample images (first and third column) are shown with suppressed background and superimposed final graph position, while the correspondent image to the right is a reconstruction from the 4-person flexible object model

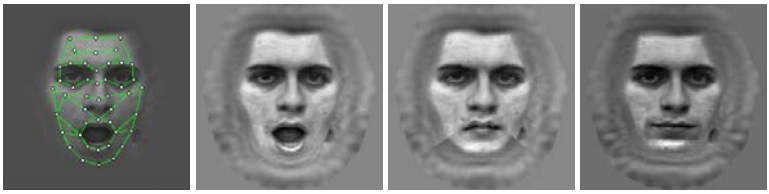


Fig. 8. Estimation of neutral expression using the FOM. From left to right are shown the original image with the best-matching graph, the image reconstructed from that graph, the estimated neutral expression using a person-indpendent FOM, and finally a reconstruction of the neutral-expression gallery image from its graph representation

is active modeling of the effect of expression change so that the test image's expression can be adjusted to that of the gallery entry (or vice versa). After that step, standard recognition tools can be used. We here show in exploratory experiments that our FOM is a viable basis for this operation.

Without loss of generality we assume that images in the gallery are of neutral expression. Using a FOM, trained as described in the previous two sections on data for several (4) persons, we first recognise the expression in the test image by selecting the best-matching canonical expression (including neutral). After

landmark finding, feature vectors are extracted from the test image and the face is transformed with the help of the FOM into neutral expression by applying the reference shape and replacing only those jets which are *significantly* deformed with the corresponding jets of the neutralised FOM. By keeping the jets which belong to landmarks hardly deformed as much as possible of the subject's identity should be preserved. An example of this approach is shown in figure 8. The thus synthesised model is compared with the one stored in the database. A similar approach can be applied to changing head pose.

6 Conclusions

We have presented an extension of the established concept of Elastic Graph Matching. Instead of synthetically constructing a model for shape variation we empirically learn it from sample image sequences requiring only minimal assistance. The model describes flexible objects in terms of deformation in shape and in texture as well as a linear mapping between the two. Applications to facial gestures are investigated in exploratory experiments. As the model is based on the data format of EGM it is immediately applicable to image matching operations, as demonstrated. More extensive experiments like recognition tasks using a larger database and further applications are in progress.

Acknowledgements

Funding by the European Commission in the Research and Training Network MUHCI (HPRN-CT-2000-00111) and by the Deutsche Forschungsgemeinschaft (WU 314/2-1) is gratefully acknowledged.

References

1. Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **19** (1997) 775–779
2. Okada, K.: Analysis, Synthesis and Recognition of Human Faces with Pose Variations. PhD thesis, University of Southern California (2001)
3. Hong, H.: Analysis, Recognition and Synthesis of Facial Gestures. PhD thesis, University of Southern California (2000)
4. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60** (2004) 135 – 164
5. Maurer, T., von der Malsburg, C.: Tracking and learning graphs and pose on image sequences of faces. In: Proceedings of the 2nd INternational Conference on Automatic Face and Gesture Recognition (FG '96), IEEE Computer Society (1996) 76
6. Bartlett, M.S.: Face Image Analysis by Unsupervised Learning. Kluwer Academic Publishers (2001)
7. Fritzke, B.: A growing neural gas network learnes topologies. In: Advances in Neural Information Processing Systems 7. MIT Press (1995) 625–632

8. Kegl, B., Krzyzak, A., Linder, T., Zeger, K.: Learning and design of principal curves. IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 281–297
9. Campbell, S.L., C.D. Meyer, J.: Generalized Inverses of Linear Transformations. Dover Publications (1991)
10. Pötzsch, M., Maurer, T., Wiskott, L., von der Malsburg, C.: Reconstruction from graphs labeled with responses of gabor filters. In: Proceedings of the ICANN 1996. (1996) 845–850

Face Reconstruction Across Different Poses and Arbitrary Illumination Conditions

Sen Wang, Lei Zhang, and Dimitris Samaras

Department of Computer Science,
SUNY at Stony Brook, NY, 11794
{swang,lzhang,samaras}@cs.sunysb.edu

Abstract. In this paper, we present a novel method for face reconstruction from multi-posed face images taken under arbitrary unknown illumination conditions. Previous work shows that any face image can be represented by a set of low dimensional parameters: shape parameters, spherical harmonic basis (SHB) parameters, pose parameters and illumination coefficients. Thus, face reconstruction can be performed by recovering the set of parameters from the input images. In this paper, we demonstrate that the shape and SHB parameters can be estimated by minimizing the silhouettes errors and image intensity errors in a fast and robust manner. We propose a new algorithm to detect the corresponding points between the 3D face model and the input images by using silhouettes. We also apply a model-based bundle adjustment technique to perform this minimization. We provide a series of experiments on both synthetic and real data and experimental results show that our method can have an accurate face shape and texture reconstruction[1].

1 Introduction

Face recognition from images has received significant attention in the past few decades. Although rapid progress has been made in this area during the last few years, the general task of recognition remains unsolved. In general, face appearance does not depend solely on identity. It is also influenced by illumination and viewpoint. Thus, recovery of 3D shape and texture from face images is an important task for an accurate face recognition system. In this paper, we propose a novel method to extract accurate 3D shape and texture from multi-pose face images taken under arbitrary unknown lighting.

Previous work[19][20] has shown that any face image taken under arbitrary unknown lighting and pose can be represented by a set of low dimensional parameters: shape parameters, spherical harmonic basis parameters, pose parameters and illumination parameters. Thus, given input images, 3D face reconstruction can be performed by estimating the shape and spherical harmonic basis parameters of the face. In this paper, we demonstrate that, given a set of multi-posed

[1] We would like to thank Sudeep Sarkar and Simon Baker for providing databases and Thomas Vetter and Sami Romdhani for helpful discussions. This research was supported by grants from U.S. Department of Justice(2004-DD-BX-1224) and National Science Foundation(ACI-0313184)

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 91–101, 2005.
© Springer-Verlag Berlin Heidelberg 2005

face images, the shape and texture parameters can be recovered by minimizing the silhouette errors and image intensity errors respectively.

We recover shape by using silhouette images because the silhouette images depend only on the shape and pose of the objects and thus are illumination independent. This reconstruction technique is also called visual hull[10][8] and the accuracy of shape reconstruction depends on the number and location of cameras used to capture images. In general, such methods cannot perform shape recovery accurately for complex objects such as human faces when the visual hull is constructed from a small number of cameras. However, prior knowledge of the object to be reconstructed can help shape recovery by providing an important constraint. In our method, the 3D face model we constructed with separate shape and texture parts provides such prior knowledge and thus facilitates accurate shape recovery.

Our method can be described by the following steps: 1) From a set of 3D faces[2] obtained by laser-based cylindrical scanners, we construct a 3D face Model with separate shape and texture parts; 2) Given a set of multi-pose input images of a human face under unknown lighting, we estimate the pose parameters and shape parameters by minimizing the difference between the silhouette of the face model and the input images. 3) Using the correspondences provided by the recovered 3D shape, we recover the illumination parameters and the spherical harmonic basis parameters by minimizing the image intensity errors. Thus, the texture of the face can be computed from the recovered spherical harmonic basis.

The main contributions of our paper are the following:

- We propose a new and efficient method to recover 3D shape and appearance from multi-pose face images under arbitrary unknown lighting.
- We present a novel algorithm to detect the corresponding points between the 3D face model and the input images by using silhouettes and use model-based bundle adjustment[16] to minimize errors and recover shape and pose parameters.
- We reconstruct appearance by recovering the spherical harmonics basis parameters from multiple input face images under unknown light while texture and illumination information are recovered in tandem.

This paper is organized as follows. In the next section, we will discuss the related work on face reconstruction. In Section 3, we will introduce shape recovery by using silhouette face images. In Section 4, we will explain appearance recovery by using our 3D face model. Experimental results on both synthetic and real data are presented in Section 5. The final Section presents the conclusions and future work directions.

2 Related Work

In recent years, there is extensive research on face reconstruction both from a single image and from image sequences. The main approaches are shape from stereo[4], shape from shading[15], shape from structured light[12] and shape from silhouettes[18].

Blanz and Vetter's face recognition system is the closest in spirit to our work. They are the first to reconstruct the shape and texture by using a face morphable model. They also apply the 3D morphable model successfully in both face recognition and synthesis applications [3][2]. In their method, they acquire the necessary point to point correspondences by using a gradient-based optical flow algorithm[2][14]. This method might suffers in situations where the illumination information is general and unknown. Compared with their method, our method determines the correspondences from silhouette which is less sensitive to illumination and texture variations.

Lee et al.[9] proposed a method of silhouette-based 3D face shape recovery by using a morphable model. They used a boundary weight XOR method to optimize the procedure and used a downhill simplex method to solve the minimization problem which is time consuming. Since they fitted a generic face model to silhouette images by marking several feature points by hand, the accuracy of their method depends on the accuracy of these feature points which can not be updated after manually marked in the generic face model. Compared with their work, we apply a model-based bundle adjustment technique to solve the optimization problem and during the optimization, the pose information is also updated thus providing better shape recovery.

Fua[6] used a generic face model to derive shape constrains and used a model-driven bundle adjustment algorithm to compute camera motions. However, the 3D face model by recovered this model-driven bundle adjustment method needs to be refined through an additional step of mesh-based deformable model optimization. In [5], Dimitrijevic et al. also used a 3D morphable model to recovery shape from face image sequences. A simple correlation-based algorithm is used to find feature points whose performance might depend on the accuracy of the correspondences detected by the cross correlation algorithm.

3 Shape Recovery

In this section we introduce our new approach to the recovery shape from multi-pose face images by using silhouette images as input to extract correspondence and recover shape parameters.

3.1 Shape Part of 3D Face Model

Let $S(\alpha)$ be the shape of an arbitrary face model parameterized by a vector $\alpha = \alpha_1, \alpha_2, ..., \alpha_n$. We want to use the silhouette images to recover this vector α. In our method, we used a collection of 3D faces supplied by USF as the bootstrap data set and we applied PCA[2] to register and align the database of 3D faces to get the statistical shape model. This model can be used to reconstruct both a new and existing faces through the linear combination of a bootstrap set of 3D face shapes.

$$s(\alpha) = \overline{s} + \sum_{i}^{M} S_i \alpha_i. \tag{1}$$

where S_i is the ith eigen-vector of the variation shape matrix and $\overline{s}$ is the mean shape of the bootstrap faces.

3.2 Silhouette Extraction

We extract face silhouettes from each input image. At the beginning we initialize a 3D face model from the input images and project the face model onto the image plane in order to extract the silhouettes of this model. Because the face model we use is not the whole head model, we do not need the complete head silhouette but only the silhouette of the facial area (in Fig. 1: example of silhouette extraction).

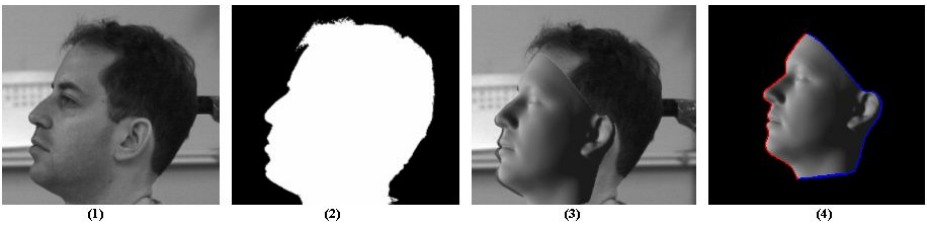


Fig. 1. Example of silhouette extraction. (1) is one of the input images, (2) is the face silhouette of this input image, (3) shows the fitting of the generic face model (shaded surface rendering) to input image, (4) is the silhouette of the fitted model (we just use the silhouette of the facial area, the red curve in left. The right blue curve is the silhouette of the omitted head boundary.

3.3 Correspondence Detection

Once we have extracted the silhouettes from the input face images and the 3D face model after fitting, we need to find the correspondences between them to update the shape and pose parameters. First, we detect the points with high curvature in the silhouettes of the face model and the input images and match them as initial correspondences by using a shortest Euclidean distance metric. Using these initial correspondences, we detect the correspondences of the remaining points in silhouettes by using a smooth matching function. Given a set of known distance vectors of feature points $u_i = p_i - \hat{p}_i$ at every matched high-curvature point i, we construct a function that gives the distance vectors u_j for every unmatched vertex j. We attempt to find a smooth vector-valued function $f(p)$ fitted to the known data $u_i = f(p_i)$, from which we can compute $u_j = f(p_j)$. There are several choices for constructing this function [7][11]. Similar to [11], we use a method based on radial basis functions $f(p) = \sum_i w_i \phi(\|p - p_i\|)$ where $\phi(r)$ is radial symmetric basis function. We also use an affine basis as part of our algorithm, so the function has the form: $f(p) = \sum_i w_i \phi(\|p - p_i\|) + Tp + m$.

To determine the coefficients w_i and the affine components T and m, we solve a linear equation $u_i = f(p_i)$, with the constraints $\sum_i w_i = 0$ and $\sum_i w_i p_i^T = 0$, which remove the effects of any affine transformation from the radial basis

function. Here we have chosen to use $\phi(r) = e^{-r/c}$, where c is a pre-selected constant ($c = 64$ in our experiment).

After we construct the function $u = f(p)$, we can use $\hat{p}_j = p_j - f(p_j)$ and a shortest Euclidean distance metric to find the remaining correspondences in silhouettes of both the face model and the input images.

3.4 Shape and Pose Parameters Update

Given a set of multi-pose input images, the shape and pose parameters can be recovered as following:

1) Initialize the shape parameters α as 0 and initialize a number of feature points in the input images. In our experiments, we manually mark 7 feature points in both the first image and the 3D face model. By matching these features across images using the point matching technique in [21], we can acquire the corresponding feature points in the other input images and thus get the initial fitting information.

2) Extract the face contour c_i (image (2) in Fig. 1) in each input image and using the current fitting information, project the face model to the image plane and extract the face model contours s_i (red line in (4) of Fig. 1) as described in section 3.2.

3) From the contours of the face model $\{s_i, i = 1...N\}$, find the corresponding points in the silhouettes of the input images $\{c_i, i = 1...N\}$ by using the methods presented in section 3.3.

4) The contour of the model s_i can be represented as $s_i = C_i^m[P \times M^p(\bar{s} + S\alpha)]$ where $C_i^m(x)$ is the contour extraction operator. M^p is the transformation matrix from the original face model coordinate system to the camera coordinate system. P is the camera projection matrix to project the 3D model to the 2D image.

Thus, the minimization can he written as follows:

$$min \sum_{i=1}^{n} \|c_i - s_i\|^2 = min \sum_{i=1}^{n} \|c_i - C_i^m[P \times M^p(\bar{s} + S\alpha)]\|^2 \quad (2)$$

For such an expression, we update the shape and pose parameters by using model-based bundle adjustment techniques[16] to solve this minimization problem.

5) After we get the new face model and new fitting parameter values, we reproject the new 3D face model to the input images and perform 2)- 4) iteratively until the change of shape and pose parameters are smaller than ξ_s and ξ_p, which are pre-selected thresholds.

4 Texture Recovery

In this section we describe a method that recovers texture from multi-pose face images under arbitrary unknown lighting. We use a spherical harmonics illumination representation to recover the spherical harmonic basis which contains texture information.

4.1 Texture Component of the 3D Face Model

As described in [1][13], any image under arbitrary illumination conditions can be approximately represented by the linear combination of the spherical harmonic basis as:

$$I \approx b\ell \tag{3}$$

where b is the spherical harmonic basis and ℓ is the vector of the illumination coefficients.

The set of images of a convex Lambertian object obtained under a wide variety of lighting conditions can be approximated accurately by a 9 dimensional linear subspace. Since human faces can be treated approximately as Lambertian, we compute a set of 9 spherical harmonic basis images by using a collection of 3D faces similar to [1] as follows:

$$\begin{aligned}
&b_{00} = \tfrac{1}{\sqrt{4\pi}}\lambda, && b_{10} = \sqrt{\tfrac{3}{4\pi}}\lambda. * n_z, && b_{20} = \tfrac{1}{2}\sqrt{\tfrac{3}{4\pi}}\lambda. * (2n_{z^2} - n_{x^2} - n_{y^2}),\\
&b^o_{11} = \sqrt{\tfrac{3}{4\pi}}\lambda. * n_y, && b^e_{11} = \sqrt{\tfrac{3}{4\pi}}\lambda. * n_x, && b^o_{22} = 3\sqrt{\tfrac{5}{12\pi}}\lambda. * n_{xy},\\
&b^o_{21} = 3\sqrt{\tfrac{5}{12\pi}}\lambda. * n_{yz}, && b^e_{21} = 3\sqrt{\tfrac{5}{12\pi}}\lambda. * n_{xz}, && b^e_{22} = \tfrac{3}{2}\sqrt{\tfrac{5}{12\pi}}\lambda. * (n_{x^2} - n_{y^2}).
\end{aligned} \tag{4}$$

where the superscripts o and e denote the odd and the even components of the harmonics respectively, λ denote the vector of the object's albedos, n_x, n_y, n_z denote three vectors of the same length that contain the x, y and z components of the surface normals. Further, n_{xy} denote a vector such that the ith element $n_{xy,i} = n_{x,i}n_{y,i}$.

In recent work [20], the set of spherical harmonic basis images of a new face can be represented by a linear combination of a set of spherical harmonic basis computed from a bootstrap data set of 3D faces.

$$b(\beta) = \overline{b} + \sum_{i}^{M} B_i\beta_i. \tag{5}$$

where $\overline{b}$ is the mean of the spherical harmonic basis and B_i is the ith eigen-vector of the variance matrix.

4.2 Texture and Illumination Parameters Update

According to Eq. 3 and 5, using the recovered shape and pose information, a realistic face image can be generated by:

$$I = (\bar{b} + B\beta)\ell \tag{6}$$

where β is the spherical harmonic basis parameter to be recovered and ℓ is the vector of illumination coefficients. Thus, given a set of n input images $I^i_{input}, i = 1...n$ of a face, the spherical harmonic basis parameters β of the face and the illumination coefficients $\ell = (\ell_1, \ell_2, ...\ell_n)$ can be estimated by minimizing the difference between the input images and the rendered images from Eq.6:

$$min_{\beta,\ell} \sum_{i=1}^{n} \| I^i_{input} - (\bar{b} + B\beta)\ell_i \|^2 \tag{7}$$

Eq. 7 is similar to Eq. 2, thus, we can solve Eq. 7 similarly. Given input images $I : I_1, I_2, ..., I_n$, we initialize the set of spherical harmonic basis parameters $\beta = 0$ and thus, $b = \bar{b} + B\beta = \bar{b}$. Hence, the set of illumination coefficients ℓ_i of each input image I_i can be initially estimated by solving a linear equation: $b\ell_i = I_i$. With the initial illumination coefficients ℓ_i, we can solve Eq. 7 using the same technique applied to Eq. 2.

The core of the recovery process is the minimization of the image errors as shown in Eq. 7. Thus, the recovery results depend on the initial values of the illumination coefficients. Our experiments on synthetic data showed that the illumination coefficients ℓ computed by using the mean spherical harmonic basis ($\bar{b}$) were close to the actual values, which made the whole recovery fast and accurate.

After we estimate the spherical harmonic basis from input images, the texture of a face can be computed as $\lambda = b_{00}\sqrt{4\pi}$ according to Eq. 4.

5 Experiments

In this section, we provide experimental results of our method on both synthetic data and real data for face reconstruction.

5.1 Synthetic Data

We use synthetic data as ground truth to show the accuracy and robustness of our method. In our experiments, we synthesize 30 face models by randomly assigning different shape and spherical harmonic basis parameters to our 3D face model. For each model we also generate 14 images with different poses and different illuminations (image sequence of one face in Figure 2). We recover the shape and texture from these images and compare them with shape and texture of the original face models.

To quantify the accuracy of our method we compute the errors between recovered models and original synthesized face models. At first, we compute the errors of shape and texture in each vertex between the reconstructed face model and the ground truth face model by: $err_s(i) = \frac{\sqrt{(\tilde{x}_i - x_i)^2 + (\tilde{y}_i - y_i)^2 + (\tilde{z}_i - z_i)^2}}{\sqrt{x_i^2 + y_i^2 + z_i^2}}$ and $err_t(i) = \frac{\|\tilde{I}_i - I_i\|}{I_i}$ where (x_i, y_i, z_i) and I_i are the coordinate and texture of ith vertex of the ground truth face model, and $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$ and $\tilde{I}_i$ are the coordinate and texture of ith vertex of the reconstructed face model. Then, we compute the maximum, minimum, mean and standard deviation of the shape and texture errors by comparing all 30 reconstructed 3D face models to the original faces as shown in Table 1. From these experimental results we can see that our method achieves accurate shape and texture recovery from synthetic data. Figure 3 shows the relationship between the reconstructed shape and the number of input images

Table 1. Statistical errors of shape and texture recovery all these 30 synthetic faces

	Max	Min	Mean	Std. dev.
Shape	12.35%	0.97%	3.53%	3.237%
Texture	23.83%	1.87%	4.78%	4.659%

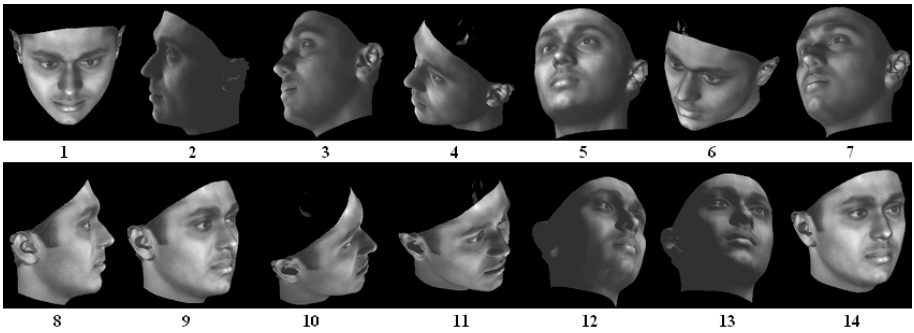


Fig. 2. 14 input images synthesized for the same face in different pose and different illumination

as a subset of the input image sequence in Figure 2 and Figure 4 shows the errors between the recovered shape from different numbers input images and the original face shape. With the increase of the number of input images, we get more accurate results of shape recovery and if the input images are more than 6, the improvement of shape reconstruction will be less influenced by the number of input images. Figure 5 shows 2 examples of shape and texture reconstruction results.

5.2 Real Data

We use the CMU PIE database [17] for our real data experiments. In the PIE data set, there are 13 different poses and 22 illumination conditions per pose for each subject. The silhouettes of face images can be detected by subtracting the background image from the input images. Figure 6 shows two accurate shape and texture recovery results of our method. The experimental results on the real data demonstrate that our method can recover good shape and texture from multi-pose face images under unknown illumination conditions.

6 Conclusions and Future Work

In this paper, we proposed a novel method for face modeling from multi-pose face images taken under arbitrary unknown illumination conditions. We demonstrated that the shape and spherical harmonic basis parameters can be estimated by minimizing the silhouette errors and image intensity errors. We proposed a new algorithm to detect the corresponding points between the model and the

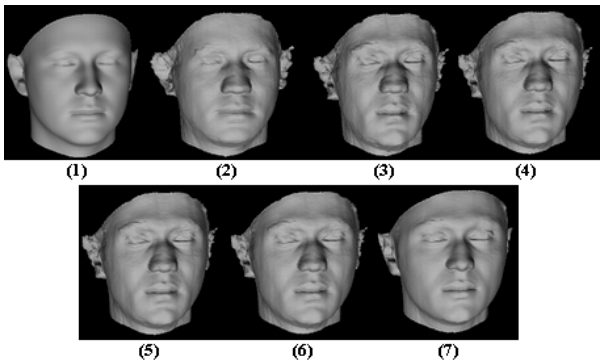


Fig. 3. Shape reconstruction using a varying number of input images in Fig 2. (1) is the mean face model, (2) is the reconstructed shape from image 1 to 3, (3) is the reconstructed shape from image 1 to 6, (4) is the reconstructed shape from image from 1 to 9, (5) is the reconstructed shape from image from 1 to 12, (6) is the reconstructed shape from image from 1 to 14 and (7) is the original shape of the face in Fig 2

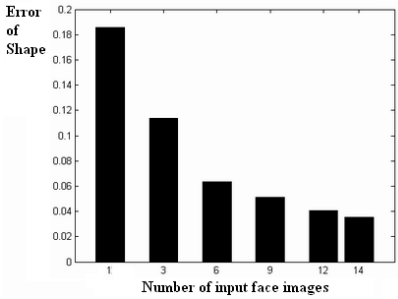


Fig. 4. The errors between the reconstructed face shape and the original face shape in Fig. 3



Fig. 5. Some reconstruction results from synthetic faces. In each row, the first image is original shape (shaded surface rendering) followed by original texture. The third image is the mean face model which is initially fitted to the input images. The last 2 images are the reconstructed face shape and texture



Fig. 6. Reconstruction results for 2 subjects from real images. Original images are in the first row, reconstructed face shapes are in the second row and recovered textures are in the last row

input images by using silhouettes. We also applied a model-based bundle adjustment technique to solve the minimization problems. We provide a series of experiments on both synthetic and real data and experimental results show that our method can reconstruct accurate face shape and texture from multi-pose face images under unknown lighting. In future, in order to extract more robust correspondences for shape recovery, we plan to use both silhouette information and image intensity information after delighting the input face images. At this time, there exist few publicly available sets of face images under arbitrary illumination conditions, so we plan to continue validation of our method on databases with greater variability of light sources as they become available.

References

1. R. Basri and D.W. Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, February 2003.
2. V. Blanz and T. Vetter. A morphable model for the synthesis of 3d-faces. In *SIGGRAPH*, 1999.
3. V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *PAMI*, 25(9):1063–1074, Sept. 2003.
4. F. Devernay and O. D. Faugeras. Computing differential properties of 3d shape from stereoscopic images without 3d models. In *CVPR*, pages 208–213, 1994.
5. M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face model from uncalibrated and ill-lit video sequences. In *CVPR*, pages 1034–1041, 2004.
6. P. Fua. Regularized bundle adjustment to model heads from image sequences without calibration data. *IJCV*, 38(2):153–171, 2000.

7. M. Gregory and Nielson. Scattered data modeling. In *Computer Graphics and Application*, 1993.
8. A. Laurentini. The visual hull concept for silhouette based image understanding. *PAMI*, 16(2):150–162, 1994.
9. J. Lee, B. Moghaddam, H. Pfister, and R. Machiraju. Silhouette-based 3d face shape recovery. In *Graphics Interface*, 2003.
10. W. Matusik, C. Buehler, R. Raskar, L. McMillan, and S. J. Gortle. Image-based visual hulls. In *SIGGRAPH*, 2003.
11. F. Pighin, Hecker J., Lischinski D., Szeliski R., and D. Salesin. Synthesizing realistic facial expressions from photographs. In *SIGGRAPH*, pages 75–84, 1998.
12. M. Proesmans, L. Vangool, and A. Osterlinck. Active acquisition fo 3d shape for moving objects. In *ICIP*, 1996.
13. R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, 2001.
14. S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error function. In *ECCV*, volume 4, 2002.
15. D. Samaras and D. Metaxas. Incorporating illumination constraints in deformable models. In *CVPR*, pages 322–329, 1998.
16. Y. Shan, Z. Liu, and Z. Zhang. Model-based bundle adjustment with application to face modeling. In *ICCV*, 2001.
17. T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination and expression database. *PAMI*, page 1615–1618, 2003.
18. L. Tang and T.S. Huang. Analysis-based facial expression synthesis. In *ICIP*, volume 94, pages 98–102, 1996.
19. L. Zhang and D. Samaras. Face recognition under variable lighting using harmonic image exemplars. In *CVPR*, volume I, pages 19–25, 2003.
20. L. Zhang, S. Wang, and D. Samaras. Face synthessis and recognition from a single image under arbitrary unknow lighting using a spherival harmonic basis morphable model. In *CVPR*, 2005.
21. Z. Zhang, R. Deriche, O. Faugeras, and Q.T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence Journal*, 78:87–119, 1995.

Stepwise Reconstruction of High-Resolution Facial Image Based on Interpolated Morphable Face Model

Jeong-Seon Park and Seong-Whan Lee*

Center for Artificial Vision Research,
Department of Computer Science and Engineering, Korea University
Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
{jspark,swlee}@image.korea.ac.kr

Abstract. This paper proposes a new method for reconstructing a high-resolution facial image from a low-resolution facial image using stepwise reconstruction based on the interpolated morphable face model. First, we defined an interpolated morphable face model that an interpolated face is composed of a low-resolution face, its interpolated high-resolution face from a low-resolution one, and its original high-resolution one. We also proposed a stepwise reconstruction method for preventing over reconstruction caused by direct reconstruction of a high-resolution image from a low-resolution facial image. The encouraging results show that our proposed method can be used to improve the performance of face recognition systems, specifically in resolution enhancement of facial images captured on visual surveillance systems.

1 Introduction

There is a growing interest in the visual surveillance systems for security areas such as international airports, borders, sports grounds, and safety areas. Though various research on face recognition have been carried out for some time now, there still exists a number of difficult problems. These include such things as estimating facial pose, facial expression variations, resolving object occlusion, changes of lighting conditions, and in particular, the low-resolution (LR) images captured on visual surveillance systems.

Handling LR images is one of the most difficult and commonly occurring problems in various image processing applications such as analysis of scientific, medical, astronomical, and weather images, archiving, retrieval and transmission of those images as well as video surveillance or monitoring[1]. Numerous methods have been reported in the area of estimating or reconstructing high-resolution (HR) images from a series of LR images or single LR image. Super-resolution is a typical example of techniques used in reconstructing a HR image from a series of LR images[2], whereas interpolation enlarges a LR image to a HR image.

* To whom all correspondence should be addressed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 102–111, 2005.
© Springer-Verlag Berlin Heidelberg 2005

We are concerned with building a HR facial image from a LR facial image for visual surveillance systems. Our reconstruction method is example-based, object-class-specific or top-down approach. The example-based approach to interpreting images of deformable objects is now attracting considerable interest among many researchers[3][4][5] because of its potential of deriving high-level knowledge from a set of prototypical examples.

In this paper, we define an interpolated morphable face model by adding interpolated image to the extended image and present new reconstruction methods for obtaining a HR facial image from a LR facial image using stepwise reconstruction of example-based learning.

2 Definition of the Interpolated Morphable Face Model

In this section, we present an overview of our reconstruction methods using example-based learning based on the interpolated morphable face model. Suppose that sufficiently large amount of facial images are available for off-line training, we could then represent any input face by a linear combination of a number of facial prototypes[7][8].

Moreover, if we have a pair of LR facial image and its corresponding HR image for the same person, we can obtain an approximation to the deformation required for the given LR facial image by using the coefficients of examples. We can then obtain a HR facial image by applying the estimated coefficients to the corresponding HR example faces as shown in Fig. 1.

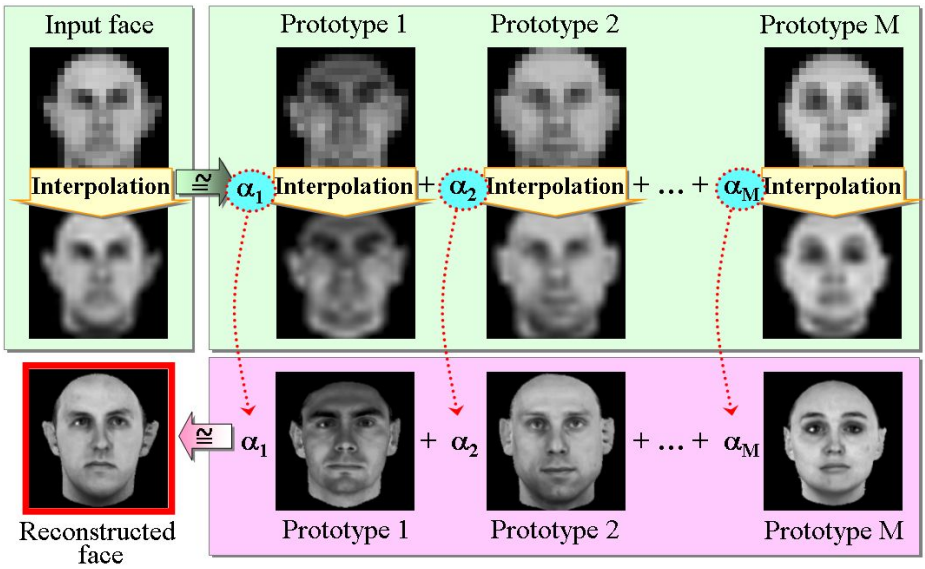


Fig. 1. Basic idea of the HR reconstruction using example-based learning

Consequently, our goal is to find an optimal parameter set $\boldsymbol{\alpha}$ which can best estimates the HR facial image from a given LR facial image.

2.1 Reconstruction Procedure of High-Resolution Facial Image

Based on the morphable face model[6][7], our reconstruction method is consists of following 4 steps, starting from a LR facial image to a HR facial image. Here the displacement of the pixels in an input LR face which correspond to those in the LR reference face is known.

Step 1. *Obtain the texture by warping an input LR face onto the reference face with its given LR shape.*
Step 2. *Reconstruct a HR shape from a given LR shape.*
Step 3. *Reconstruct a HR texture from the obtained LR texture at Step 1.*
Step 4. *Synthesize a HR face by warping the reconstructed HR texture with the reconstructed HR shape.*

Step 1(backward warping) and Step 4(forward warping) are explained from the previous studies of morphable face models in many studies[4][7]. Step 2 and Step 3 are carried out by similar mathematical procedure except that the shape about a pixel is 2D vector and the texture is 1D(or 3D for RGB color image) vector.

2.2 Definition of Interpolated Morphable Face Model

In order to reconstruct a HR facial image from a LR one, we defined an extended morphable face model in which an extended face is composed of a pair of LR face and its corresponding HR one, and we separated an extended face by an extended shape and an extended texture according to the definition of morphable face model[9].

In addition to we applied interpolation techniques to the extended shape and the extended texture under the assumption that we can enlarge the amount of information from LR input image by applying interpolation techniques such as bilinear, bicubic, and so on except nearest method. Fig. 2 shows an example of the facial image defined by the interpolated morphable face model, where bicubic method is used for enlarging LR shape and LR texture.

Then we can define $S^+ = (d_1^x, d_1^y, \cdots, d_L^x, d_L^y, d_{L+1}^x, d_{L+1}^y \cdots, d_{L+I}^x, d_{L+I}^y, d_{L+I+1}^x, d_{L+I+1}^y \cdots, d_{L+I+H}^x, d_{L+I+H}^y)^T$ to be an interpolated shape vector by simply concatenating a LR shape, the interpolated HR shape, and original HR shape, where L, I and H is the number of pixels in input LR facial image,in the interpolated HR one, and in the original HR one, respectively. Similarly, let us define $T^+ = (i_1, \cdots, i_L, i_{L+1}, \cdots, i_{L+I}, i_{L+I+1}, \cdots, i_{L+I+H})^T$ to be an interpolated texture vector.

Next, we transform the orthogonal coordinate system by principal component analysis(PCA) into a system defined by eigenvectors s_p^+ and t_p^+ of the covariance matrices C_S^+ and C_T^+ computed over the differences of the interpolated shape

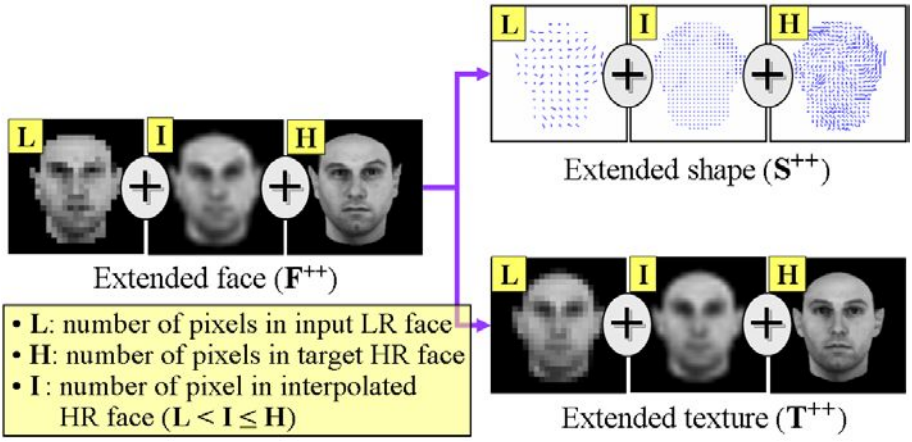


Fig. 2. An example facial image defined by the interpolated morphable face model

and texture, $\tilde{S^+} = S^+ - \bar{S^+}$ and $\tilde{T^+} = T^+ - \bar{T^+}$. Where $\bar{S^+}$ and $\bar{T^+}$ represent the mean of interpolated shape and that of texture, respectively. Then, an interpolated facial image can be represented by the following equation.

$$S^+ = \bar{S^+} + \sum_{p=1}^{M-1} \alpha_p {s_p}^+, \quad T^+ = \bar{T^+} + \sum_{p=1}^{M-1} \beta_p {t_p}^+ \tag{1}$$

where $\boldsymbol{\alpha},\ \boldsymbol{\beta} \in \Re^{M-1}$.

3 High-Resolution Shape Reconstruction Methods

In order to reconstruct a HR facial image from an input LR one, we need to reconstruct both HR shape and texture from a LR shape and texture, respectively. As described before, both reconstructions can be carried out by similar mathematical procedure, therefore we will describe only the Step 2 of reconstructing HR shape from LR one.

3.1 Mathematical Solution for High-Resolution Reconstruction

We can use both LR shape and interpolated HR one from input LR facial image, according to the previous definition of interpolated shape. We need an approximation to the deformation required for both shapes by using the coefficients of the bases(see Fig. 1). The goal is to find an optimal parameter set α_p that satisfies

$$\tilde{S}^+(x_j) = \sum_{p=1}^{M-1} \alpha_p s_p^+(x_j), \quad j = 1,\ 2, \cdots,\ L + I, \tag{2}$$

where x_j is a pixel in the LR facial image, $M - 1$ the number of bases and L and I the number of pixels in input LR image and interpolated HR image.

We assume that the number of observations, $L+I$, is larger than the number of unknowns, $M-1$. Generally there may not exist a set of α_p that perfectly fits the $\tilde{S}^+$. So, the problem is to choose $\hat{\boldsymbol{\alpha}}$ so as to minimize the reconstruction error. For this, we define following error function, $E(\boldsymbol{\alpha})$ the sum of square of errors which measures the difference between the known displacements of pixels in the LR input image and its represented ones.

$$E(\boldsymbol{\alpha}) = \sum_{j=1}^{L+I} (\tilde{S}^+(x_j) - \sum_{p=1}^{M-1} \alpha_p s_p^+(x_j))^2. \tag{3}$$

Then the problem of reconstruction is formulated as finding $\hat{\boldsymbol{\alpha}}$ which minimizes the error function

$$\hat{\boldsymbol{\alpha}} = arg \min_{\boldsymbol{\alpha}} E(\boldsymbol{\alpha}). \tag{4}$$

The solution to Eqs. (3) - (4) is nothing more than least square solution. Eq. (2) is equivalent to the following equation.

$$\begin{pmatrix} s_1^+(x_1) & \cdots & s_{M-1}^+(x_1) \\ \vdots & \ddots & \vdots \\ s_1^+(x_{L+I}) & \cdots & s_{M-1}^+(x_{L+I}) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{M-1} \end{pmatrix} = \begin{pmatrix} \tilde{S}^+(x_1) \\ \vdots \\ \tilde{S}^+(x_{L+I}) \end{pmatrix} \implies \mathbf{S}^+ \boldsymbol{\alpha} = \tilde{\mathbf{S}}^+. \tag{5}$$

By applying our mathematical solution for solving the least square minimization problem[9], we can obtain an optimal parameter as follows.

$$\boldsymbol{\alpha}^* = (\mathbf{S}^{+T}\mathbf{S}^+)^{-1}\mathbf{S}^{+T}\tilde{\mathbf{S}}^+. \tag{6}$$

Finally, we can obtain a HR shape by using the solution $\boldsymbol{\alpha}^*$

$$S(x_{L+I+j}) = \bar{S}^+(x_{L+I+j}) + \sum_{p=1}^{M-1} \alpha_p^* s_p^+(x_{L+I+j}), \quad j = 1, 2, \ldots, H, \tag{7}$$

where $x_{L+I+1}, \cdots, x_{L+I+H}$ are pixels in the HR facial image, H is the number of pixels in the HR facial image.

3.2 Iterative Error Back-Projection Method

According to our example-based learning methods, we approximate a given LR shape or texture with some errors, that is defined by Eq.(3) of the estimated α^*. So, we can easily guess that the estimated HR shape also has some error if we know the original HR shape,

$$E(\boldsymbol{\alpha}) = \sum_{j=1}^{H} (\tilde{S}^+(x_{L+I+j}) - \sum_{p=1}^{M-1} \alpha_p^* s_p^+(x_{L+I+j}))^2 \tag{8}$$

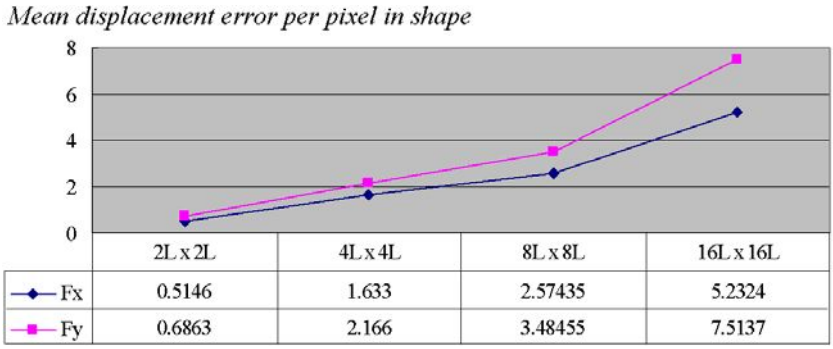
where $x_1, , x_2, \cdots, x_H$ are pixels in the HR facial image.

In our previous works, iterative error back-projection is applied to reduce the above reconstruction error by iteratively compensating for the HR error, which is estimated by similar error reconstruction from a simulated LR error. The previous iterative error back-projection is composed or three stages: estimation of HR data, simulation of LR data and error compensation of estimated HR data with the reconstructed HR error. The detailed procedure was described our previous report[9].

3.3 Stepwise Reconstruction Method

In our previous reconstruction method, we tried to directly reconstruct a target HR image from an input LR image based on the interpolated morphable face model. But, the initially reconstructed facial images are somewhat fluctuated on the texture and unnatural caused by over-sized reconstruction.

Fig. 3 shows the changes of mean displacement errors according to the size of reconstructed HR shape from input $L \times L$ LR shape. As shown in this figure, the reconstruction errors are increased according to the ratio of input LR image and target HR one, as we can easily guess.



	2L x 2L	4L x 4L	8L x 8L	16L x 16L
Fx	0.5146	1.633	2.57435	5.2324
Fy	0.6863	2.166	3.48455	7.5137

Fig. 3. Changes of mean displacement errors according to the size of reconstructed image

In order to reduce the reconstruction errors caused by over-sized reconstruction such as reconstructing 256×256 HR images from 16×16 LR image, we proposed stepwise reconstruction method. The proposed stepwise method sequentially reconstructs next upper HR image starting from an input $L \times L$ LR image as shown in the Fig. 4. We applied the stepwise reconstruction method for reconstructing HR shape and texture, respectively.

4 Experimental Results and Analysis

4.1 Face Database

For testing the performance of our reconstruction methods, we used 200 facial images of Caucasian faces that were rendered from a database of 3D head models

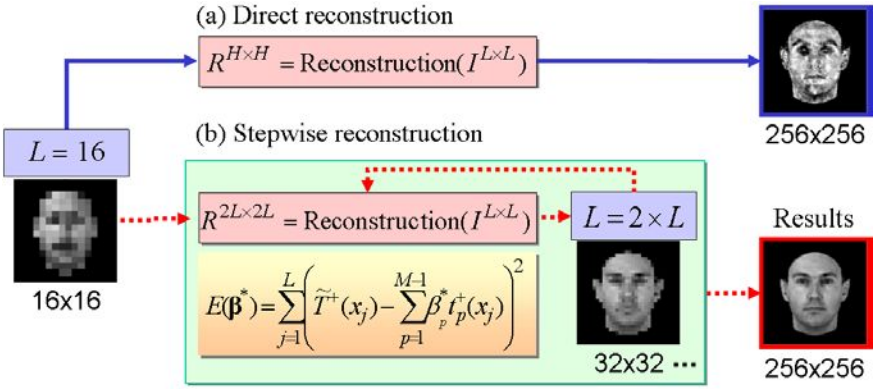


Fig. 4. Comparison of direct reconstruction and stepwise reconstruction

recorded by a laser scanner[7]. The original images were color images, set to the size of 256×256 pixels. They were converted to an 8-bit gray level and resized to 16×16 and 32×32 for LR facial images. PCA was applied to a random subset of 100 facial images for constructing bases of the defined face model. The other 100 images were used for testing our reconstruction methods.

4.2 Reconstruction Results and Analysis

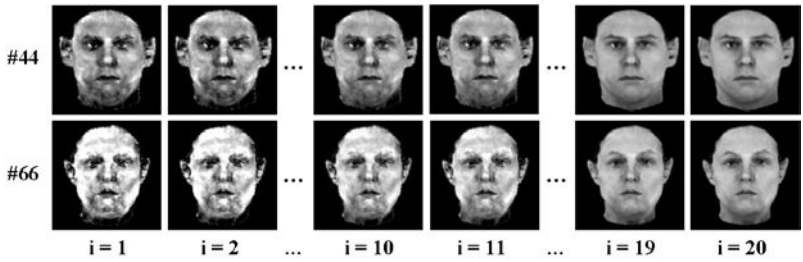
As mentioned before, 2D-shape and texture of facial images are treated separately. Therefore, a HR facial image is reconstructed by synthesizing the estimated HR shape and the estimated HR texture.

Fig. 5 shows the effects of the iterative error back-projection, where (a) examples of iteratively updated HR facial images and (b) the changes in the mean and the standard deviation of the intensity errors per pixel between the original HR image and the iteratively updated HR images. From this gradually decreasing distance trend, we can conclude that the similarity between the original HR facial images and the compensated one increased as the number of iterations increased.

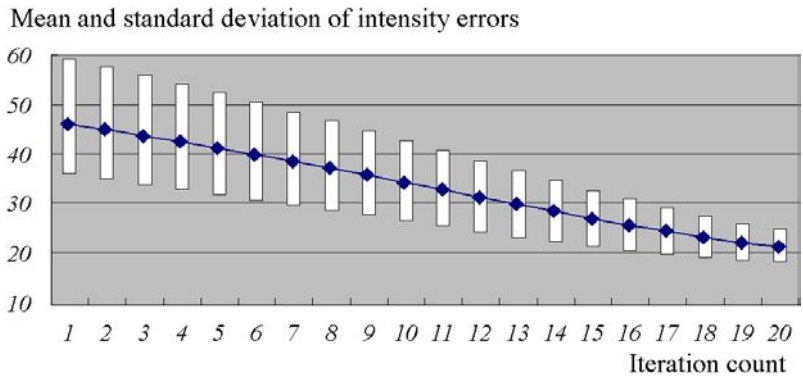
Figs. 6 shows the examples of the 256×256 HR facial image and their edge images reconstructed from 16×16 LR image. In this figure, (a) shows the input LR images, (b) and (c) the interpolated HR images using Bilinear and Bicubic interpolation, respectively. And (d) and (e) the reconstructed HR images by proposed method 1(direct reconstruction method) and by proposed method 2 (stepwise reconstruction method), respectively. Finally, (f) shows the original HR facial images.

As shown in Fig. 6, classifying the input LR faces is almost impossible, even with the use of Bilinear or Bicubic interpolations. On the other hand, reconstructed HR facial images by the proposed reconstruction methods, especially the reconstructed images by our stepwise method are more similar to the original faces than others.

From the encouraging results of the proposed method as shown in Fig. 6, we can expect that it can be used to improve the performance of face recognition



(a) Examples of iteratively updated HR facial image



(b) Changes of the mean and standard deviation of intensity errors per pixel

Fig. 5. Effects of the iterative error back-projection

systems by reconstructing a HR facial image from a LR facial image captured on visual surveillance systems.

In order to verify the effect of HR reconstruction, we carried out simple face recognition experiments under the following configurations. The original 256×256 facial images were registered, and the reconstructed HR facial images from 16×16 facial images were used as recognition data. Figure 7 shows us the correct recognition rates of face recognition experiments. As we can see, the recognition performance has improved by employing the proposed reconstruction methods.

5 Conclusions

In this paper, we provided efficient methods of reconstructing a high-resolution facial image using stepwise reconstruction based on the interpolated morphable face model. Our reconstruction method consists of the following steps: computing linear coefficients minimizing the error or difference between the given shape or texture and the linear combination of the shape or texture prototypes in the low-resolution image, and applying the coefficient estimates to the shape and texture prototypes in the high-resolution facial image, respectively. Finally

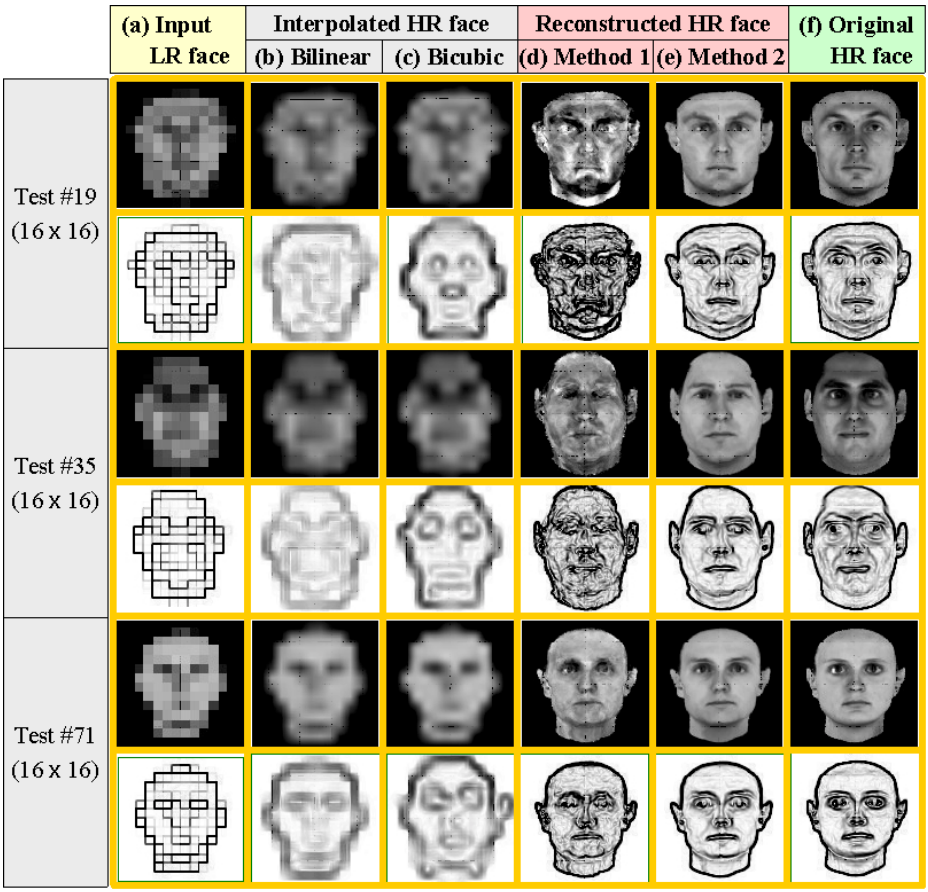


Fig. 6. Examples of 256 × 256 HR facial images reconstructed from 16 × 16 LR facial images

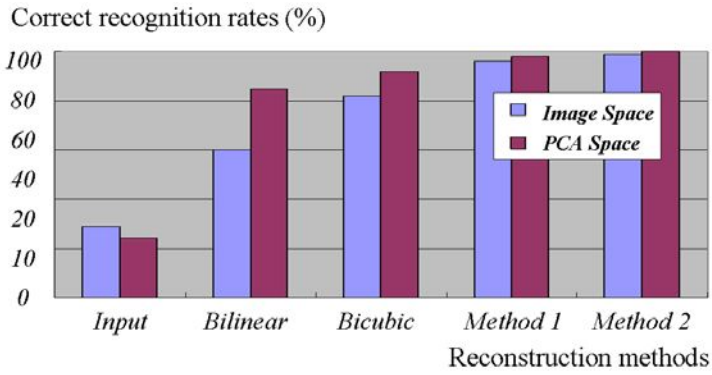


Fig. 7. Comparisons of recognition performance

applying iterative error back-projection or stepwise reconstruction for reducing the measured reconstruction errors.

The experimental results appear very natural and plausible similar to original high-resolution facial images. This was achieved when displacement among the pixels in an input face which correspond to those in the reference face, were known. It is a challenge for researchers to obtain the correspondence between the reference face and a given facial image under low-resolution vision tasks.

Acknowledgments

This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea. We would like to thank the Max-Planck Institute for providing the MPI Face Database.

References

1. Tom, B., Katsaggelos, A.K.: Resolution Enhancement of Monochrome and Color Video Using Motion Comensation. IEEE Trans. on Image Processing, Vol. 10, No. 2 (Feb. 2001) 278–287
2. Baker, S., Kanade, T.: Limit on Super-Resolution and How to Break Them. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 24 No. 9 (Sep. 2002) 1167-1183
3. Windyga, P. S.: Fast impulsive noise removal. IEEE Trans. on Image Processing, Vol. 10, No. 1 (2001) 173–178
4. Jones, M. J., Sinha, P., Vetter, T., Poggio, T.: Top-down learning of low-level vision tasks[brief communication]. Current Biology, Vol. 7 (1997) 991–994
5. Hwang, B.-W., Lee, S.-W.: Reconstruction of Partially Damaged Face Images Based on a Morphable Face Model. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25, No. 3 (2003) 365–372
6. Beymer, D., Shashua, A., Poggio, T.: Example-Based Image Analysis and Synthesis. AI Memo 1431/CBCL Paper 80, Massachusetts Institute of Technology, Cambridge, MA (Nov. 1993)
7. Vetter, T., Troje, N. E.: Separation of Texture and Shape in Images of Faces for Image Coding and Synthesis. Journal of the Optical Society of America A. Vol. 14, No. 9 (1997) 2152–2161
8. Blanz, V.,Vetter, T.: Face Recognition Based on Fitting a 3D Morphable Model. IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol. 25, No. 9 (Sep. 2003) 1063-1074
9. Park, J.-S., Lee, S.-W.: Resolution Enhancement of Facial Image Using an Error Back-Projection of Example-based Learning. Proc. of the 6th Int'l Conf. on Automatic Face and Gesture Recognition, Seoul, Korea (May 2004) 831–836

Illumination Invariant Face Recognition Using Linear Combination of Face Exemplars

Song-Hyang Moon, Sang-Woong Lee, and Seong-Whan Lee*

Center for Artificial Vision Research/
Department of Computer Science and Engineering, Korea University,
Anam-dong, Seongbuk-ku, Seoul 136-713, Korea
{shmoon,sangwlee,swlee}@image.korea.ac.kr

Abstract. Facial appearance changes induced by lighting variation cause serious performance degradation in face recognition. Current face recognition systems encounter the difficulty to recognize faces under arbitrary illuminations. In this paper, we propose a new face recognition method under arbitrary lighting conditions, given only a single registered image and training data under unknown illuminations. Our proposed method is based on the exemplars which are synthesized from photometric stereo images of training data and the linear combination of those exemplars are used to represent the new face. We make experiments for verifying our approach and compare it with two traditional approaches. As a result, higher recognition rates are reported in these experiments using the illumination subset of Max-Planck Institute Face Database.

1 Introduction

Changes in a person's appearance induced by illumination are sometimes larger than differences in appearance of individuals, illumination changes are the most challenging problem in face recognition. In the past few years, many methods have been proposed to solve this problem with improvements in recognition being reported. Early works in illumination invariant face recognition focused on image representations that are mostly insensitive to changes under various lighting [4]. Various images representations are compared by measuring distances on a controlled face database. Edge map, second derivatives and 2D Gabor filters are examples of the image representations used. However, these kind of approaches have some drawbacks. First, the different image representations can be only extracted once they overcome some degree of illumination variations. Second, features for the person's identity are weakened whereas the illumination-invariant features are extracted.

The different approaches, called the photometric-stereo method, are based on the low dimensionality of the image space [1]. The images of one object with a Lambertian surface, taken from a fixed viewpoint and varying illuminations lie in a linear subspace. We can classify the new probe image by checking to

* To whom all correspondence should be addressed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 112–121, 2005.
© Springer-Verlag Berlin Heidelberg 2005

see if it lies in the linear span of the registered gallery images. These gallery images are composed of at least three images of the same person under different illuminations. Since it recognizes the new image by checking that it is spanned in a linear subspace of the multiple gallery images, it cannot handle the new illuminated images of a different person.

To avoid the necessity of multiple gallery images, the bilinear analysis approach is proposed [2]. It applies SVD(Singular Value Decomposition) to a variety of vision problems including identity and lighting. The main limitation of these bilinear analysis methods is that prior knowledge of the images, like the lighting direction of training data are required.

Unlike the methods described above, Blanz and Vetter use 3D morphable models of a human head [5]. The 3D model is created using a database collected by Cyberware laser scans. Both geometry and texture are linearly spanned by the training ensemble. This approach enables us to handle illumination, pose and expression variations. But it requires the external 3D model and high computational cost.

For illumination-robust face recognition, we have to solve the following problem: *Given a single image of a face under the arbitrary illumination, how can the same face under the different illumination be recognized?* In this paper, we propose a new approach for solving this problem based on the synthesized exemplars. The illuminated-exemplars are synthesized from photometric stereo images of each object and the new probe image can be represented by a linear combination of these synthesized exemplars. The weight coefficients are estimated in this representation and can be used as the illumination invariant identity signature.

For face recognition, our proposed method has several distinct advantages over the previously proposed methods. First, the information regarding the lighting condition of training data is not required. We can synthesize the arbitrary illuminated-exemplars from the photometric stereo images of training data. Second, we can perform recognition with only one gallery image by using linear analysis of exemplars in the same class. Third, the coefficients of exemplars are the illumination invariant identity signature for face recognition, which results in high recognition rates.

2 Background

We begin with a brief review of the photometric stereo method with Lambertian lighting model and bilinear analysis of illuminated training images. We will explain what is the Lambertian reflectance and how it can be used in the photometric stereo images for face recognition [1]. We will also explain recognition methods using the bilinear analysis of the training images [2],[3].

2.1 Photometric Stereo

We assume the face has the Lambertian surface, the illuminated image I can represented by

$$I = \rho N^T L = T^T L \tag{1}$$

where n is the surface normal and ρ is the albedo, a material dependant coefficient. The object-specific matrix, T includes albedo and surface normal information of object. We have n images, $(I_1, I_2, ..., I_n)$ of one object under varying illumination. These images, called photometric stereo images, were observed at a fixed pose and different lighting sources. Assuming that they are from the same object a with single viewpoint and various illuminations, the following can be expressed

$$\mathbf{I} = \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_n \end{pmatrix} = \begin{pmatrix} T^T L_1 \\ T^T L_2 \\ \vdots \\ T^T L_n \end{pmatrix} = T^T \begin{pmatrix} L_1 \\ L_2 \\ \vdots \\ L_n \end{pmatrix} = T^T \mathbf{L} \tag{2}$$

where $\mathbf{I}$, the collection images $\{I_1, I_2, ..., I_n\}$ of the same object under different lighting condition, is the observation matrix. $\mathbf{L} = \{L_1, L_2, ..., L_n\}$ is the light source matrix. If the lighting parameters are known, we can extract the surface normal orientation for objects. To solve T, the least squares estimation of $\mathbf{I}$ using SVD. We can classify a new probe image by computing the minimum distance between the probe image and n-dimensional linear subspace. Photometric stereo method requires at least 3 gallery images for one object. Multiple gallery images are large restrictions for a real face recognition system.

2.2 Bilinear Models

Bilinear models offer a powerful framework for extracting the two-factor structure, identity and lighting. Bilinear analysis approaches had applied SVD to a variety of vision problems including identity and lighting [2],[3]. For bilinear analysis, training images of different objects under the same set of illuminations are required. Theses approaches also assume the Lambertian surface and the image space $T^T L$, where both T and L vary. Let $L_1, L_2, ..., L_n$ be a basis of linearly independent vectors, $\mathbf{L} = \sum_{j=1}^{n} \beta_j L_j$ for some coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_n)$. Let $\{T_1, ..., T_m\}$ be a basis for spanning all the possible products between albedo and surface normal of the class of objects, thus $T = \sum_{i=1}^{m} \alpha_i T_i$ for some coefficients $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$. Let $\mathbf{A} = \{A_1, ..., A_m\}$ be the matrix whose columns are the images of one object, i.e., $A_k = \alpha_k T_k \sum_{j=1}^{n} \beta_j L_j$. A_k are n images of k-th object and the column of A_k, A_{k_j} is the image of k-th object under j-th illumination. Therefore we can represent the new probe image H by linear combination of $\{A_1, ..., A_m\}$ with the bilinear coefficients, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

$$H = \rho_H N^T L = T_H^T L = (\sum_{i=1}^{m} \alpha_k T_k)(\sum_{j=1}^{n} \beta_j L_j) = \boldsymbol{\alpha}\boldsymbol{\beta}\mathbf{A} \tag{3}$$

The bilinear problem in the $m + 3$ unknowns is finding $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Clearly, we solve these unknowns, we can generate the image space of object H from any desired illumination condition simply by keeping $\boldsymbol{\alpha}$ fixed and varying $\boldsymbol{\beta}$. But these approaches require the same set of illuminations per object, so that we have to know about the lighting condition of training data in advance.

3 Linear Analysis of the Synthesized Exemplars

We propose an illumination invariant face recognition method based on the synthesized exemplars. We synthesize the illuminated-exemplars from photometric stereo images and represent a new probe image by linear combination of those exemplars. The procedure has two phases: training and testing. Images in the database are separated into two groups for either training or testing. In the training procedure, we construct the training data to consist of at least three illuminated images per object. However we do not know the lighting conditions of training data and the training data can be constructed using different objects and different sets of illuminations unlike bilinear analysis method. In our experiments, we construct the train matrix as m people under n different illuminated images. This is followed by computing the orthogonal basis images by the PCA for inverting the observation matrix per person. The orthogonal basis images of one person are used to synthesize the exemplars. We can then reconstruct a novel illuminated image using these basis images of the same face. In the testing procedure, we synthesize the exemplars under the same illumination as the input image. The lighting conditions of these m synthesized exemplars and input images are same. The input image can be represented by the linear combination of the exemplars, the weight coefficients are used as those signature identities for face recognition. In the registration, those gallery images are already saved for the recognition, we find the facial image that has the nearest coefficient by computing the correlation.

3.1 Synthesis of the Exemplars

We assume that the face has a Lambertian surface and the light source, whose locations are not precisely known, emits light equally in all directions from a single point. Then, an image I is represented by $T^T L$ as shown Eq. 1 and the matrix $\mathbf{I}$ that made n images can be represented by $T^T\mathbf{L}$ as shown Eq. 2. The photometric stereo images are from the same object, we can assume that they have the same object-specific matrix T and different illumination vector $\mathbf{L}$. If the light source matrix $\mathbf{L}$ is non-singular($|\mathbf{L}| \neq 0$) and $\{L_1, L_2, ..., L_n\}$ are linearly independent, the matrix $\mathbf{L}$ is invertible and then T can be expressed by the product of matrix $\mathbf{I}$ and the pseudo-inverse of $\mathbf{L}$, $\mathbf{L}^+$.

$$T = \mathbf{I}\mathbf{L}^+ \tag{4}$$

The light source matrix $\mathbf{L}$ can be invertible when $\{L_1, L_2, ..., L_n\}$ are linearly independent of each other. To make the images independent from each other, we transform the photometric stereo images into the orthogonal basis images, $\{B_1, B_2, ..., B_{n-1}\}$, by principal component analysis (PCA). By applying PCA to photometric stereo images, we can express a new illuminated image of the same object using the orthogonal basis images by changing the coefficients $\boldsymbol{\alpha}$ and the orthogonal basis images can be obtained in off-line training. Our method for synthesizing the image, called '***exemplar***', proposes that we use the input image

as a reference. Since photometric stereo images have the same object-specific matrix and the input image is used as a reference, the synthesized exemplar's lighting condition is similar to that of input image. The input image H can be represented using a linear combination of orthogonal basis images.

$$H = \bar{B} + \boldsymbol{\alpha}\, \mathbf{B} \tag{5}$$

where $\bar{B}$ represents the mean of orthogonal basis images per object and $\boldsymbol{\alpha} \in \Re^{n-1}$. We can find the coefficient $\boldsymbol{\alpha}$ as follows. The columns of matrix are orthogonal to each other, the transpose is the inverse and we can now easily find the coefficient vector $\boldsymbol{\alpha}^*$ by transpose instead inverse.

$$\boldsymbol{\alpha}^* = \mathbf{B}^{-1}(H - \bar{B}) = \mathbf{B}^T(H - \bar{B}) \tag{6}$$

In the photometric stereo images, we choose three images of random lighting directions, $\{\tilde{I}_1, \tilde{I}_2, \tilde{I}_3\}$ and we transform those images into the orthogonal coordinate system by PCA by eigenvectors $\{\tilde{B}_1, \tilde{B}_2\}$. Where $\bar{\tilde{B}}$ is the mean of $\{\tilde{B}_1, \tilde{B}_2\}$ and $\tilde{\boldsymbol{\alpha}}^* = \{\tilde{\alpha}_1, \tilde{\alpha}_2\}$ is the coefficient for synthesizing the exemplar $\tilde{E}$, an exemplar using three images is as follows.

$$\tilde{E} = \bar{\tilde{B}} + \sum_{j=1}^{2} \tilde{\alpha}_j \tilde{B}_j = \bar{\tilde{B}} + \tilde{\boldsymbol{\alpha}}^* \tilde{\mathbf{B}} \tag{7}$$

Fig. 1 shows examples of the synthesized exemplars from the training data. We choose three images under random illumination of each person and those chosen images for each person are different set. The top row represents three different illuminated images of the same person from the training data. The middle row shows examples of the synthesized exemplars using the images from the top row. While bottom row shows examples of the different illuminated input images. Each synthesized exemplar image (middle row) references the illumination of input image found directly below it. As shown, the synthesized exemplars have very similar lighting conditions to that of the input image. One exemplar image is synthesized per object, so there are m exemplar images under the same lighting condition of the input image where the training data is collected by the images of m objects.

3.2 Linear Combination of Synthesized Exemplars

In the previous section, we described that how the exemplar is synthesized. Using both the photometric stereo images and input image as illumination reference, m exemplars are synthesized per person. The exemplar $\tilde{E}_k$ of k-th person can be represent as

$$\tilde{E}_k = \bar{\tilde{B}}_k + \sum_{j=1}^{2} \tilde{\alpha}_{k_j} \tilde{B}_{k_j} = \bar{\tilde{B}}_k + \tilde{\boldsymbol{\alpha}}_k^* \tilde{\mathbf{B}}_k \tag{8}$$

where $\bar{\tilde{B}}$ is the mean of orthogonal basis images $\{\tilde{B}_1, \tilde{B}_2\}$ from three photometric stereo images, $\{\tilde{I}_{k_1}, \tilde{I}_{k_2}, \tilde{I}_{k_3}\}$. The column of $\mathbf{I}_k$, I_{k_i} is the image under i-th

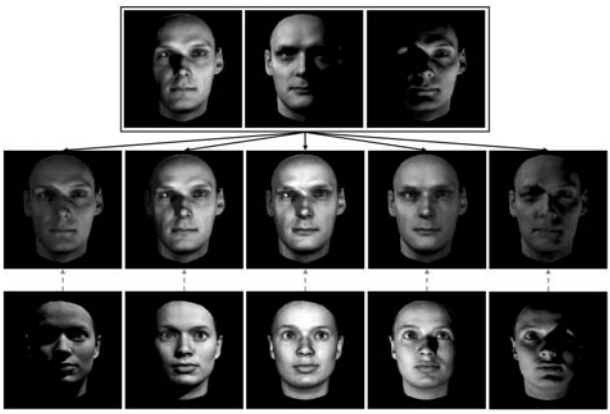


Fig. 1. Example of the synthesized exemplars

illumination of k-th person. The input image is represented well by the linear combination of the exemplars. At this time, the linear coefficients are estimated from the exemplars under the same illumination. That means, the coefficients depend on the m exemplars but not on the lighting conditions. Because the exemplars are for the object class only, the coefficients provide a signature identity that is invariant to illumination. The coefficient vector $\boldsymbol{f}$ is computed by the following equation.

$$H = \sum_{k=1}^{m} f_k \tilde{E}_k = \boldsymbol{f}\tilde{\mathbf{E}} \tag{9}$$

where $\boldsymbol{f} = \{f_1, f_2, ..., f_m\}$ is the coefficient vector from the m exemplars and used for recognition. f_k is the weight coefficient for the k-th exemplar object. $\tilde{\mathbf{E}} = \{\tilde{E}_1, \tilde{E}_2, ..., \tilde{E}_m\}$ is the matrix of the synthesized exemplars. The problem is to choose $\boldsymbol{f}$ so as to minimize the cost function, $\mathcal{C}(\boldsymbol{f})$. We define the cost function as the sum of square errors which measures the difference between the input image and the linear sum of the exemplars. We can find the optimal coefficient $\boldsymbol{f}$, which minimizes the cost function, $\mathcal{C}(\boldsymbol{f})$.

$$\boldsymbol{f}^{\,*} = \arg\min_{\boldsymbol{f}} \mathcal{C}(\boldsymbol{f}) \tag{10}$$

with the cost function,

$$\mathcal{C}(\boldsymbol{f}) = \sum_{i=1}^{d} (H(x_i) - \sum_{k=1}^{m} f_k \tilde{E}_k(x_i))^2 \tag{11}$$

To represent the input image H using exemplars, we have to find $\boldsymbol{f}$ by the equation of $H = \tilde{\mathbf{E}}\,\boldsymbol{f}$, where $\tilde{\mathbf{E}} = \{\tilde{E}_1, \tilde{E}_2, ..., \tilde{E}_m\}$. The least square solution satisfies $\tilde{\mathbf{E}}^T H = \tilde{\mathbf{E}}^T \tilde{\mathbf{E}} \boldsymbol{f}$. If the columns of $\tilde{\mathbf{E}}$ are linearly independent, then $\tilde{\mathbf{E}}^T \tilde{\mathbf{E}}$ is non-singular and has an inverse.

$$\boldsymbol{f}^{\,*} = (\tilde{\mathbf{E}}^T \tilde{\mathbf{E}})^{-1} \tilde{\mathbf{E}}^T H \tag{12}$$

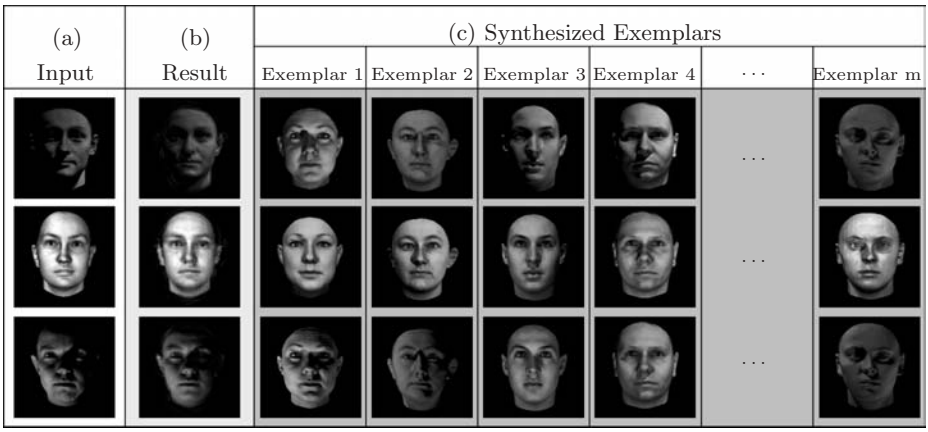


Fig. 2. Example of the reconstructed input images

We can express the input image H using the computed $\boldsymbol{f}^*$ on the assumption that the columns of matrix $\tilde{\mathbf{E}}$ are linearly independent. If they are not independent, the solution $\boldsymbol{f}^*$ will not be unique, in this case, the solution can be solved by the pseudo-inverse of $\tilde{\mathbf{E}}$, $\tilde{\mathbf{E}}^+$. But, that is unlikely to the happen for proposed method. The reconstructed image H^R of the input image H is represented as follows.

$$H^R = \sum_{k=1}^{m} f_k^* \tilde{E}_k = \boldsymbol{f}^* \tilde{\mathbf{E}} \tag{13}$$

By using Eq.(13), we can get the optimal weight coefficient vector to represent the input image. To verify the coefficients as the signature identity, we reconstruct the input image using the computed coefficients. Fig. 2 shows the example of reconstructed images using coefficients $\boldsymbol{f}^*$. In this figure, (a) shows the input image under the arbitrary lighting condition, (b) shows the reconstructed images using the linear combination of exemplars and (c) is the synthesized exemplars with the input image as illumination reference.

3.3 Recognition

In this section, we describe what kind of signature is used for recognizing the face. We use the linear coefficients of the synthesized exemplars for face recognition. When the gallery or probe image is taken, we synthesize exemplars from the photometric stereo images with each gallery or probe image. We analyze the input image, gallery and probe image, by the synthesized exemplars. We can then obtain the linear coefficients of both the gallery image and probe image, those coefficients are used the signatures for face recognition. Suppose that a gallery image G has its signature $\boldsymbol{f}_g^*$ and a probe image P has its signature $\boldsymbol{f}_p^*$.

$$\boldsymbol{f}_g^* = (\tilde{\mathbf{E}}_g^T \tilde{\mathbf{E}}_g)^{-1} \tilde{\mathbf{E}}_{\mathbf{g}}^T G, \qquad \boldsymbol{f}_p^* = (\tilde{\mathbf{E}}_p^T \tilde{\mathbf{E}}_p)^{-1} \tilde{\mathbf{E}}_{\mathbf{p}}^T P, \tag{14}$$

where $\tilde{\mathbf{E}}_g$ and $\tilde{\mathbf{E}}_p$ are the matrices of synthesized exemplars using G and P as illumination reference image. The normalized correlation between a gallery and probe image is

$$corr(G,P) = \frac{Cov(\boldsymbol{f}_g^*, \boldsymbol{f}_p^*)}{sd(\boldsymbol{f}_g^*)sd(\boldsymbol{f}_p^*)} \tag{15}$$

where $sd(a)$ is the standard deviation of a and $Cov(a,b)$ means the covariance of a and b.

4 Experiments

We have conducted a number of experiments with our approach using the MPI (Max-Planck Institute) Face Database [5]. In these experiments, we compared the proposed method with 'Eigenface/WO3 [7]' and 'Bilinear analysis'[2] method. To solve the illumination problem, this method is applied without three principal components, the most influential factor in degradation of performance. We also implemented the bilinear analysis method for comparison.

4.1 Face Database

The MPI Face Database is used to demonstrate our proposed approach. We use 200 two-dimensional images of Caucasian faces that were rendered from a database of three-dimensional head models recorded with a laser scanner ($Cyberware^{TM}$) [6]. The images were rendered from a viewpoint 120cm in front of each face with ambient light only. For training, we use the images of 100 people. We use 25 face images in different illumination conditions, from $-60°$ to $+60°$ in the yaw axis and from $-60°$ to $+60°$ in the pitch axis, per person.

4.2 Experimental Results and Analysis

We present the recognition results when the images of training and testing sets are taken from the same database. We have conducted two experiments by changing the lighting directions of the gallery and probe set.

Gallery Set of Fixed Direction and Probe Set of All Lighting Directions: Graph in Fig. 3 shows the position configuration of the lights and the recognition rates for the fixed gallery set of lighting conditions(100 images) with the probe sets of varying lighting conditions (100 images under each illumination). We use the gallery set under the first lighting condition, $L11$ and the probe sets under the other 24 lighting conditions, from $L12$ to $L55$ in the testing set. In this experiment, we obtain good recognition results although the illumination changes are rapidly. As shown Fig. 3, when the distance between the light sources of the gallery and probe sets are small, the recognition performance is high, conversely when the distance between the two are large, the recognition results are of lower quality, especially when using the eigenface/WO3 and the

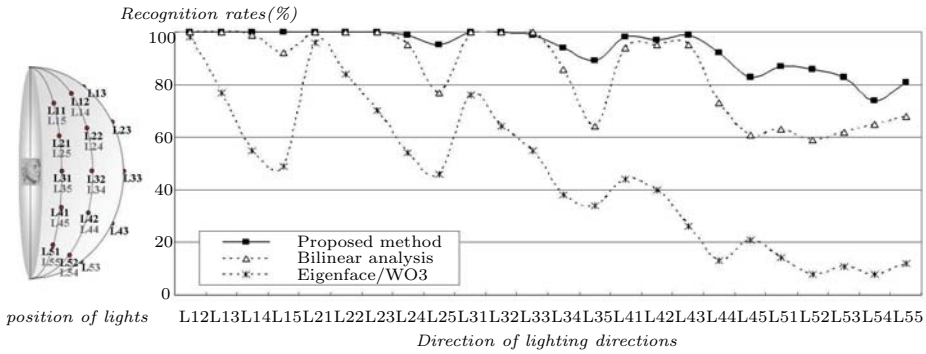


Fig. 3. Recognition rates for all the probe sets with a fixed gallery set

bilinear methods. The bilinear analysis method allows higher recognition rates than the eigenface/WO3 method, though neither method results in as high a performance as our proposed method.

Gallery Set and Probe Set of Varying Lighting Directions: The next experiment is designed for the gallery and probe sets both under different lighting conditions. Table 1 represents the comparison results between our approach and bilinear analysis approach for the gallery and probe sets under the different directions of lighting, $\{L11, L22, L33, L44, L55\}$. P means the probe sets and G means the gallery sets. The right number is for bilinear analysis approach and the left one for our approach. The average rates obtained by bilinear analysis are 88.9%, while our approach outperforms it at an average of 95.1%.

Table 1. Recognition rates comparison

G \ P	L11	L22	L33	L44	L55	Avg.
L11	-	100/100	99/100	92/73	81/68	94.4/88.2
L22	100/100	-	100/100	100/79	86/51	97.2/86.0
L33	99/100	100/100	-	100/100	99/100	99.6/100.0
L44	85/77	99/87	100/100	-	100/100	96.8/92.8
L55	63/43	79/47	95/97	100/100	-	87.4/77.4
Avg.	89.4/84.0	95.6/86.8	98.8/99.4	98.4/90.4	93.2/71.0	95.1/88.9

5 Conclusions and Future Work

We have addressed a new approach for illumination invariant face recognition. The idea here is to synthesize exemplars using photometric stereo images and apply them to represent the new input image under the arbitrary illumination, while only one input image and one registered image per person are required for recognition. The weight coefficients are used as the signature identity, so that

a new image can be represented as a linear combination of a small number of exemplars of training data. Experimental results on various face images have shown a good performance when compared with the previous approaches and our approach also shows a stable recognition performance even under the large changes of illumination. In the future, we need to make more experiments with the other face database. Furthermore, it can become particularly difficult when illumination is coupled with pose variation. Because there are the extreme lighting changes which are caused by pose variation, we are also trying to treat not only lighting changes but also pose changes.

Acknowledgments

This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea and we would like to thank the Max-Planck-Institute for providing the MPI Face Database.

References

1. Basriand, R., Jacobs, D.: Photometric Stereo with General, Unknown Lighting. Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. (2001) 374–381
2. Freeman, W.T., Tenenbaum, J.B.: Learning bilinear models for two-factor problems in vision. Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 17 (1997) 554–560
3. Shashua, A., Raviv, T.R.: The Quotient Image: Class Based Re-rendering and Recognition with Varying Illuminations. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 2 (2001) 129–139
4. Adini, Y., Moses, Y., Ullman, S.: Face Recognition: the Problem of Compensating for Changes in Illumination Direction. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7 (1997) 721–732
5. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 9 (2003) 1063–1074
6. Blanz, V., Romdhani, S., Vetter, T.: Face Identification across Different Poses and Illuminations with a 3D Morphable Model. Proc. of the 5th International Conference on Automatic Face and Gesture Recognition (2002) 202–207
7. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Congnitive Neutoscience, Vol. 3 (1991) 72–86

Video-Based Face Recognition Using Bayesian Inference Model

Wei Fan[1], Yunhong Wang[1,2], and Tieniu Tan[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, P.R. China
[2] School of Computer Science and Engineering, Bei Hang University, Beijing, 100083, P.R. China

Abstract. There has been a flurry of works on video sequence-based face recognition in recent years. One of the hard problems in this area is how to effectively combine the facial configuration and temporal dynamics for the recognition task. The proposed method treats this problem in two steps. We first construct several view specific appearance sub-manifolds learned from the training video frames using locally linear embedding (LLE). A general Bayesian inference model is then fit on the recognition task, transforming the complicated maximum likelihood estimation to some elegant distance measures in the learned sub-manifolds. Experimental results on a middle-scale video database demonstrate the effectiveness and flexibility of our proposed method.

1 Introduction

A majority of state-of-the-art face recognition algorithms [1] put emphasis on still image-based scenarios either by holistic template matching [2,3] or geometric feature-based methods [4]. Although these dominating approaches have achieved a certain level of success in restricted conditions such as mug-shot matching, they often fail to yield satisfactory performance when confronted with large pose, illumination and expression variations.

Recently, there is a significant trend in performing video-based face analysis [5,6,7,8], aiming to overcome the above limitations by utilizing visual dynamics or temporal consistence to facilitate performance of the recognition task. These approaches take root in relevant psychological and neural studies [9] which indicate that information for identifying a human face can be found both in the invariant structure of features and in idiosyncratic movements and gestures. As illustrated in Fig. 1, the *dynamic information* in terms of human face recognition can be typically divided into three categories: rigid head motions, no-rigid facial movements and the combination of both. Several researchers in this area have conjectured that if expressive dynamic information can be properly extracted, they will surely give a favorable improvement to video-based face recognition.

With this motivation, a new phase of recognition strategies that use both spacial and temporal information simultaneously has started. In [5], an identity surface for each subject is constructed in a discriminant feature space from

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 122–130, 2005.
© Springer-Verlag Berlin Heidelberg 2005

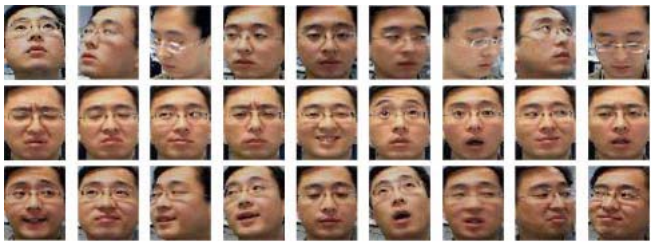


Fig. 1. Facial dynamics can be typically divided into three categories: rigid head motions (Top), no-rigid facial movements (Middle), and the combination of both (Bottom)

one or more learning sequences, and recognition is performed by computing distances between object trajectory and a set of model trajectories which encode the spatio-temporal information of a moving face. Zhou *et al* [6] simultaneously characterize the kinematics and identity using a motion vector and an identity variable respectively in a probabilistic framework. Sequential importance sampling (SIS) algorithm is developed to estimate the joint posterior distribution, and marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable. Recently, Hidden MarKov Models (HMM) [7] and probabilistic appearance manifolds [8] are both used to learn the transition probabilities among several viewing states embedded in the observation space. Hadid *et al* [10] compared the joint spatio-temporal representation (e.g. the HMM) with classical ones based on static images (e.g. PCA/LDA) for performing dynamic face recognition, and pointed out that the former model outperforms its counterparts in most experiments.

Although facial dynamics, if properly modelled, are tolerate to appearance variations induced by changes in head pose orientation and expressions (see Fig. 1 as an example), the most essential features for recognition still lie on those *static facial configurations*. Thus dynamic information, which provides us with some unstable behavioral characteristics, should be only treated as an assistant cue to the recognition task under non-optimal viewing conditions. The proposed approach in this paper is an attempt to somewhat balance the attention to static facial configurations for video-based recognition scenario.

To this end, we view sets of face images as high dimensional points whose underlying degrees of freedom is far less than the actual number of pixels per image. A well-known manifold learning algorithm, locally linear embedding (LLE) [11,12], is used to detect low dimensional structure in the image sequences for different individuals. As all human faces are similar patterns, we may anticipate under identical viewing conditions, e.g. rotation from left to right profiles [13], the manifolds of different individuals are often fairly close and parallel. Thus view specific sub-manifolds can be well constructed using classic clustering techniques on an individual's low dimensional embedding, assuming there is sufficient data (such that the manifold is well-sampled). Face images extracted from other training videos are sequentially assigned to its corresponding submanifolds under the nearest "distance-from-feature-space" (DFFS) criteria [14].

To exploit the temporal coherence among successive video frames, we fit a general Bayesian inference model on the recognition task, transforming the complicated maximum likelihood estimation to some elegant distance measures in the learned view specific sub-manifolds. Experimental results conducted on a middle-scale video database strongly support our assumption and show high superiority of the newly developed method to its traditional still image-based counterparts.

2 View Specific Sub-manifolds Construction

2.1 Dimensionality Reduction Using LLE

In typical appearance-based methods, $m \times n$ face images are often represented by points in the mn-dimensional space. However, coherent structure in the facial appearance leads to strong correlations between them, generating observations that lie on or close to a low-dimensional manifold. When the face images are extracted from video sequences, it is reasonable to assume that the manifold is smooth and well-sampled. Unlike traditional linear techniques, PCA and LDA, which often over-estimate the true degrees of freedom of the face data set, recent nonlinear dimensionality reduction methods, Isomap [15] and LLE [11,12], can effectively discover an underlying low dimensional embedding of the manifold. In this section, we use LLE to map the high-dimensional data to a single global coordinate system in a manner that preserves the neighboring relationships. An overview of the LLE algorithm is given in Table 1.

Table 1. An overview of the LLE algorithm [11,12]

INPUT:		$X = \{ x_1, x_2, \cdots, x_N \}$, where $x_i \in \mathbb{R}^D$.
OUTPUT:		$Y = \{ y_1, y_2, \cdots, y_N \}$, where $y_i \in \mathbb{R}^d, d \ll D$.
METHOD:		Repeat for each data point x_i:
	1	Find K nearest neighbors.
	2	Reconstruct x_i from its neighbors, minimizing the cost function $\varepsilon(W) = \left\| x_i - \sum_j W_{ij} x_j \right\|^2$ subjected to the additional constraints that $\sum_j W_{ij} = 1$ and $W_{ij} = 0$ if x_i and x_j are not neighbors.
	3	Define the embedding cost function $\varepsilon(y) = \left\| y_i - \sum_j \hat{W}_{ij} y_j \right\|^2$ where $\hat{W}_{ij}$ is the optimal result from step 2. Find the reconstructed vectors $\hat{y}_i = \arg\min_y \varepsilon(y), y_i \in \mathbb{R}^d$, with the additional constraints that $\sum_i y_i = 0$ and $\sum_i y_i y_i^T / N = I$.

2.2 View Specific Sub-manifolds

To illustrate the effectiveness of LLE, we applied it to a sequence of face images corresponding to a single person arbitrarily rotating his head. This data set contained $N = 788$ grayscale images at 23×28 resolution ($D = 644$). Fig. 2 shows the first three components of these images discovered by LLE (using $K = 12$ nearest neighbors) together with some representative frames. As we can see, the algorithm successfully revealed the meaningful hidden structure of the nonlinear face manifold.

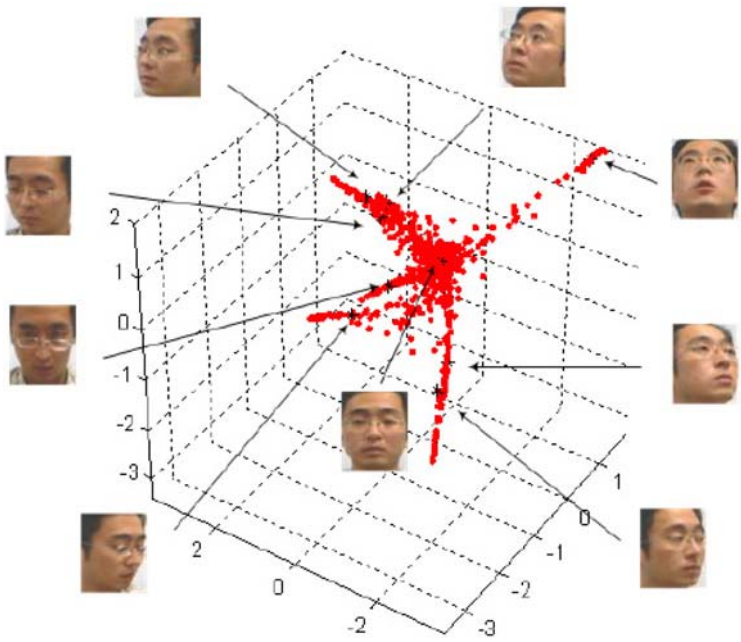


Fig. 2. LLE applied to a sequence of face images corresponding to a single person arbitrarily rotating his head

To construct view specific sub-manifolds, we performed K-means clustering to points in the low-dimensional feature space given by LLE. The initial k cluster seeds were selected as some frames bearing distinct pose variations in the sequence (just like those shown in Fig. 2). Given large distances between the initial seeds and a moderate k, the output clusters could act as the expected view specific sub-manifolds in our method, which were further approximated by a set of linear subspaces ($P_i, i = 1, 2, \cdots, k$). Face images extracted from other training videos of different persons were sequentially assigned to its corresponding sub-manifolds under the nearest "distance-from-feature-space" (DFFS) criteria [14]. Thus in the training process, successive video frames will continuously update P_i, and provide an enhanced model of the view specific sub-manifolds.

3 Bayesian Inference Model for Recognition

In statistical pattern recognition [16], Bayesian inference model offers an efficient and principled approach for integrating prior knowledge and observed data to improve the classification performance. It is also an effective tool to characterize abundant temporal information in the video-based face recognition scenario [17]. Although these facial dynamics are by no means stable features as mentioned before, the temporal coherence among successive video frames still provides a significant aid in the recognition process.

Suppose w is the identity signature for a c class problem, i.e. $w \in \{1, 2, \cdots, c\}$. Given a sequence of face images $F = \{f_1, f_2, \cdots, f_N\}$ containing the appearances of the same but unknown person, Bayesian inference model aims to find the solution with the maximum a posterior probability (MAP)

$$\hat{w} = \arg \max_{\{1,2,\cdots,c\}} P(w|f_{1:N}) \tag{1}$$

According to the Bayesian theory

$$P(w|f_{1:N}) = \frac{P(w)P(f_{1:N}|w)}{P(f_{1:N})} \tag{2}$$

We further assume the prior probability $P(w)$ to be non-informative and neglect the normalization factor $P(f_{1:N})$ which is independent to the final decision. Thus the MAP solution is converted to an equivalent maximum likelihood (ML) estimation

$$\hat{w} = \arg \max_{\{1,2,\cdots,c\}} P(f_{1:N}|w) \tag{3}$$

As the face images tend to lie on or close to a non-convex low-dimensional manifold, it is hard to analytically capture its complexity in a universal or parametric solution. One possible way to tackle this problem is to build a view-based formulation with a set of subspaces $(P_i, i = 1, 2, \cdots, k)$ covering the whole manifold, as introduced in Section 2. Here we simply associate each image with a hidden view parameter θ, where $\theta \in \{P_1, \cdots, P_k\}$, and decompose (3) as follows:

$$\begin{aligned} P(f_{1:N}|w) &= \sum_{\theta_{1:N}} P(f_{1:N}|\theta_{1:N}, w)P(\theta_{1:N}) \\ &= \sum_{\theta_{1:N}} \prod_{t=1}^{N} P(f_t|\theta_t, w)P(\theta_t|\theta_{1:t-1}) \\ &= \sum_{\theta_{1:N}} \prod_{t=1}^{N} P(f_t|\theta_t, w)P(\theta_t|\theta_{t-1}) \end{aligned} \tag{4}$$

In the above derivation, we use two intuitive rules which are appropriate for video-based face recognition, namely (a) observational conditional independence: $P(f_{1:N}|\theta_{1:N}, w) = \prod_{t=1}^{N} P(f_t|\theta_t, w)$ and (b) the first-order Markov chain rule: $P(\theta_{1:N}) = \prod_{t=1}^{N} P(\theta_t|\theta_{1:t-1}) = \prod_{t=1}^{N} P(\theta_t|\theta_{t-1})$, $P(\theta_1|\theta_0) \doteq P(\theta_1)$. The following subsections show how to compute the two probabilities involved in (4).

3.1 Computation for $P(f_t|\theta_t, w)$

The term $P(f_t|\theta_t, w)$ denotes the probability of observing face image f_t at time t, given its corresponding identity w in the view sub-manifold θ_t. Typically a multivariate Gaussian density is fitted for this distribution when a large training set is available. Here we avoid directly estimating the particular density function for the limited training data, and convert it to some elegant distance measures related to the learned sub-manifolds.

From the definition of the *law of total probability* and the conditional probability we have

$$P(f_t|\theta_t, w) = P(\theta_t|f_t)P(w|f_t, \theta_t)\frac{P(f_t)}{P(\theta_t|w)P(w)} \tag{5}$$

As before, the prior $P(w)$ and evidence $P(f_t)$ are assumed non-informative. And $P(\theta_t|w)$ represents the likelihood of w being in sub-manifold θ_t at time t, which is related to the behavioral characteristic of subject w. To simplify the computational setting in our case, all three terms are treated as constants, thus

$$P(f_t|\theta_t, w) \propto P(\theta_t|f_t)P(w|f_t, \theta_t) \tag{6}$$

For a k sub-manifold problem, let $d_m(P_i, f_t)$ be the Euclidean distance between the ith sub-manifold and the test sample f_t (approximated by the DFFS measure [14]), an estimation of the probability $P(\theta_t|f_t)$ can be approximated as

$$P(\theta_t|f_t) = \frac{1/d_m(\theta_t, f_t)^2}{\sum_{i=1}^{k} 1/d_m(P_i, f_t)^2} \tag{7}$$

Similarly for a c class problem, let $d_c(j, f_t)$ be the distance between the jth class center and the test image f_t with all the related training data belonging to the sub-manifold θ_t. Thus the term $P(w|f_t, \theta_t)$ can be approximated as

$$P(w|f_t, \theta_t) = \frac{1/d_c(w, f_t)^2}{\sum_{j=1}^{c} 1/d_c(j, f_t)^2} \tag{8}$$

Here $d_c(j, f_t)$ is measured by the "distance-in-feature-space" (DIFS)criteria [14], and the feature space is constructed using null space-based linear discriminant analysis (NLDA) [18].

3.2 Computation for $P(\theta_t|\theta_{t-1})$

Motivated by the similar work of [8], the transition probability $P(\theta_t|\theta_{t-1})$ is defined by counting the actual transitions between different sub-manifolds P_i observed in all the training sequences:

$$P(\theta_t|\theta_{t-1}) = \frac{1}{\lambda}\sum_{t=2}^{k} \delta(f_{t-1} \in \theta_{t-1})\delta(f_t \in \theta_t) \tag{9}$$

where $\delta(f_t \in \theta_t) = 1$ if $f_t \in \theta_t$ and otherwise is 0. The normalization factor λ ensures $P(\theta_t|\theta_{t-1})$ to be a probability measure.

4 Experiments

To demonstrate the effectiveness of the proposed method, extensive experiments were performed on a 25-subject video dataset which bears large pose variation and moderate differences in expression and illumination. Each person is represented by one training clip and one testing clip both captured in our lab with a CCD camera at 30 fps for about 15 seconds. The faces were manually cropped from all frames and resized to 23×28 pixel gray level images, followed by a histogram equalization step to eliminate lighting effects. The examples shown in Fig. 3 are representative of the amount of variation in the data.



(a) training video clips (b) testing video clips

Fig. 3. Representative examples for two subjects from the training and testing data used in the experiments. Note the significant pose variation in both sets

Nine view specific sub-manifolds ($P_i, i = 1, 2, \cdots, 9$) are learned from all the training videos using strategies proposed in Section 2.2. The baseline image sequence for LLE modelling (Fig. 2) contains a person with abundant pose variations. Subsequent video frames of other subjects are sequentially absorbed by the relevant sub-manifolds. This step inevitably produces a few false assignments which are automatically detected by certain thresholds and manually corrected.

The MAP estimation for each testing sequence was evaluated by (2). Fig. 4 shows the computed posterior probabilities of the two persons in Fig. 3 as a function of time t. From the figure, it is obvious that the true signature (red line) always gives the largest posterior probability.

To illustrate the superiority of this newly developed method to its traditional still image-based counterparts, we implemented the LLE+clustering algorithm [10] which chose the cluster centers of each training sequence as extracted exemplars for template matching and took a vote to give the final decision. PCA and LDA were used as the classification methods in [10]. Here we also provide experimental result given by NLDA classifier [18]. Table 2 summarizes the recognition rates on our databet averaged among various sequence length (like the testing strategy in [7]) using different approaches mentioned above. The results clearly show that the proposed method outperforms all its still image-based counterparts, as it greatly profits from the Baysian inference model while other approaches use dynamic information only in its most crude form through voting.

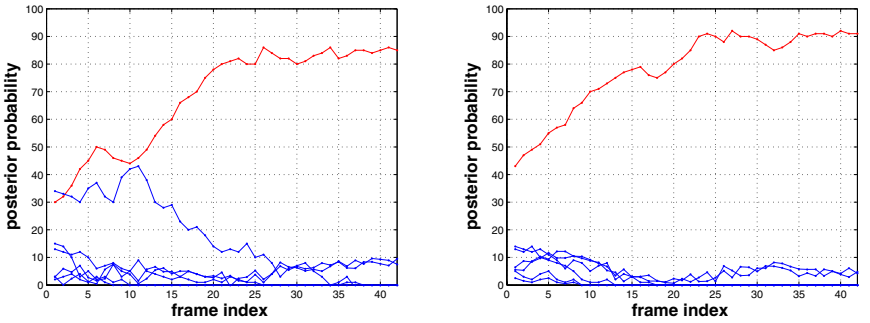


Fig. 4. Posterior probability $P(w|f_{1:N})$ of the two persons in Fig. 3 against time t. Only the seven most likely candidates are shown in this figure. From the figure, it is obvious that the true signature (red line) always gives the largest posterior probability

Table 2. Recognition rate (%) for different methods using $k = 9$ clusters

Method	LLE+PCA	LLE+LDA	LLE+NLDA	Our method
Recogition rate	82.62	87.21	91.62	95.24

5 Discussion and Conclusions

This paper presents a novel video-based face recognition method using both spacial and temporal information simultaneously. Unlike most other joint spatio-temporal representations which excessively rely on unstable facial dynamics for recognition, we exploit dynamic information in a moderate fashion, i.e. only those constraints of common transitions along the face manifold are modelled by the Baysian inference framework. More emphases are put on the construction of view specific sub-manifolds, which essentially convey relevant discriminating information, i.e. the static facial configurations, for the recognition task. As our work combines the major analytic features of the manifold learning algorithm LLE − precise preservation of the neighboring relationships in a single global coordinate system − with the flexibility to learn a moderate model of facial dynamics, it is especially suitable to the video-based face recognition scenario and exhibited satisfactory performance in a middle-scale video dataset.

Acknowledgements

This work is funded by research grants from the National Basic Research Program of China (No. 2004CB318110) and the National Natural Science Foundation of China (No. 60332010).

References

1. R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey", Proceedings of the IEEE, Vol. 83, pp. 705-741, May 1995.
2. M. Turk and A. Pentland, "Eigenfaces for recognition", J. of Cognitive Neuroscience, Vol. 3, No. 1, pp. 71-86, 1991.
3. V. Belhumeur, J. Hespanda, and D. Kiregeman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection", IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 711-720, July 1997.
4. L. Wiskott, J.M. Fellous, N. Kruger and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No.7, pp. 775-779, July, 1997.
5. Y. Li, S. Gong, and H. Liddell, "Video-Based Online Face Recognition Using Identity Surfaces", Proceedings of IEEE ICCV Workshop on Recogntion, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 40-46, 2001
6. S. Zhou and R. Chellappa. "Probabilistic human recognition from video". European Conference on Computer Vision (ECCV), May 2002.
7. X.Liu, and T.Chen, "Video-Based Face Recognition Using Adaptive Hidden Markov Models", Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 340-345, 2003.
8. K.C.Lee, J.Ho, M.H.Yang, and D.Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds", In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 313-320, 2003.
9. Alice J. O'Toole*, Dana A. Roark, Herv Abdi, "Recognizing moving faces: A psychological and neural synthesis", Trends in Cognitive Sciences, 6, 261-266. Reed, CL, Stone, VE, Bozova, S., Tanaka, J. (2003).
10. A.Hadid, and M.Pietikainen, "From Still Image to Video-Based Face Recognition: An Experimental Analysis", Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 813-818, 2004.
11. S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," Science 290, 2323-2326, 2000.
12. L. K. Saul and S. T. Roweis, "Think Globally, Fit Locally : Unsupervised Learning of Nonlinear Manifolds," Technical Report MS CIS-02-18, University of Pennsylvania, 2003.
13. Gong S, McKenna S J and Collins J J, "An Investigation into Face Pose Distributions", Second International Conference on Automated Face and Gesture Recognition, Vermont, USA, October 1996.
14. A. Pentland, B. Moghaddam, T. Starner, "View-based and modular eigenspaces for face recognition", Proceedings of IEEE, CVPR, 1994.
15. J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," Science, 290(5500):2319–2323, 2000.
16. R. Duda, P. Hart, and D. Stork. Pattern Classification. WileyInterscience, 2001.
17. S. Zhou and R. Chellappa. "Probabilistic Identity Characterization for Face Recognition". In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 805-812, 2004.
18. Wei Fan, Yunhong Wang, Wei Liu, Tieniu Tan, "Combining Null Space-based Gabor Features for Face Recognition", In Proc. of the 17th International Conference on Pattern Recognition. pp. 330-333, Cambridge, UK, 2004.

A Hybrid Swipe Fingerprint Mosaicing Scheme

Yong-liang Zhang[1], Jie Yang[1], and Hong-tao Wu[2]

[1] Inst. of Image Processing & Pattern Recognition, Jiaotong University, Shanghai, 200030, P.R. China
yongliangzhang@sjtu.edu.cn
[2] School of Computer Science & Software, Hebei University of Technology, Tianjin, 300130, P.R. China

Abstract. Due to their very small contact areas and low cost, swipe fingerprint sensors provide the very convenient and reliable fingerprint security solutions and are being increasingly used for mobile phones, PDAs, portable computers and security applications. In this paper, the minimum mean absolute error as the registration criterion is used to find an integer translation shift while the extension of the phase correlation method with singular value decomposition is applied to evaluate a non-integer translation shift. Based on the merits and faults of these two methods, we develop a hybrid swipe fingerprint mosaicing scheme. The proposed scheme succeeds in registering swipe fingerprint frames with small overlap down to 5% of the frame size and improves the mosaicing precision, that is, non-integer translation shift can be directly determined without spatial domain interpolation. Experimental data indicate that our scheme has high reliability and precision and less time consumption, therefore, it is very suitable for the real-time applications.

1 Introduction

Fingerprints are today the most widely used biometric features due to their uniqueness and immutability[14]. Using current technology, fingerprint identification is in fact much more reliable than other possible personal identification methods based on signature, face, or speech alone[1].

Swipe fingerprint sensors[15][16] provide the very convenient and reliable fingerprint security solutions and are being increasingly used for mobile phones, PDAs, portable computers and security applications due to their very small contact areas and low cost. For example, the active sensing area of FPC1031B is 152×32 pixels and that of AES2510 is 196×16 pixels. Swipe sensor captures a fingerprint by swiping the finger past it and its captured swipe fingerprint(Fig.1) is a stream of swipe fingerprint frames (SFFs), which are contiguous and share some mutual support. The frame data are then "stitched" or "registered" together to form a fingerprint image.

Fingerprint registration has already become an important issue for the success of reliable fingerprint verification using small solid state fingerprint sensors. Current major registration techniques include template synthesis [7][12]and fingerprint image mosaicing[2][9]. Template synthesis merges the fingerprint features while image mosaicing merges the fingerprint images to generate a composite fingerprint image from features of the images. However all above methods

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 131–140, 2005.
© Springer-Verlag Berlin Heidelberg 2005

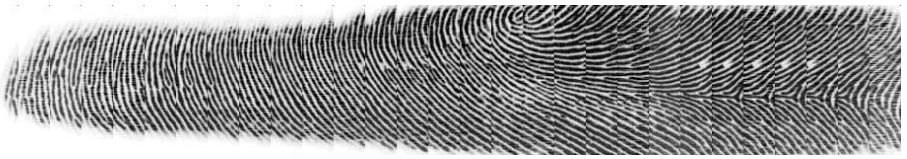


Fig. 1. A stream of swipe fingerprint frames

base on minutiae alignment and the nature of swipe design in swipe sensor leads to very few minutiae presented in each captured swipe fingerprint frame. These facts make it unavailable to align two adjacent SFFs using the minutiae. However, two adjacent SFFs can be mosaiced by measuring the overlap between subsequent partial images of the finger. In other words, we can preform two adjacent SFFs registration by aligning them based on the similarity measurement. The important process of swipe fingerprint mosaicing can be modelled with the following simplifications derived from[9]: (i) a large number of SFFs are available covering the fingerprint (the frames of a swipe fingerprint stream are being acquired at the real-time frame rate of 30 frames a second); (ii) the swipe time period is relatively short, so there are no changes in value of the intrinsic and extrinsic imaging parameters. Thus we can assume that two adjacent SFFs represent the same scene sampled on identical grids but offset from each other by an unknown translation shift (TSS). Extensive experiments show the assumption is feasible.

Image mosaicing involves automatic alignment of two or more images into an integrated image without any visible seam or distortion in the overlapping areas. The phase correlation method (PCM)[4][5] is known to provide straightforward estimation of TSS between two images. And the extension of the popular PCM with singular value decomposition (SVD)[10] leads to non-integer TSS without interpolation, robustness to noise, and limited computational complexity. However, these methods based on phase correlation have a fatal limitation: the corresponding overlap between two images to be registered must be 30% bigger than the smaller image size[13]. Fortunately, the mean absolute error (MAE) as a similarity measurement to compute the optimal integer TSS between two images overcomes the limitation.

In this paper, the MAE is used as the similarity measure to evaluate the optimal integer TSS between two adjacent SFFs while the extension of the PCM with SVD is applied to evaluate a non-integer TSS. On the basis of the merits and faults of these two methods, we develop a hybrid swipe fingerprint mosaicing scheme. The paper is organized as follows: the pixel level mosaicing algorithm based on MAE is described in Section 2; In Section 3 the phase correlation and SVD are applied to get subpixel level mosaicing; In Section 4 we develop a hybrid scheme to mosaic the swipe fingerprint; Implementation issues including the elliptoid masking and Kaiser window are presented in Section 5; Experimental results are presented in Section 6. Finally some concluding remarks are provided in Section 7.

2 The Pixel Level Mosaicing

Typical similarity measures for image are cross-correlation with or without pre-filtering, sum of absolute differences, and Fourier invariance properties such as phase correlation[8]. The minimum mean square error (MSE) has already become one of registration criteria. Compared to the minimum MSE, the minimum MAE criterion is preferred for real-time applications because it requires no multiplication while yielding the performance. So we choose the minimum MAE as our registration criterion to find an integer TSS that registers two adjacent SFFs to the nearest integral pixel coordinates.

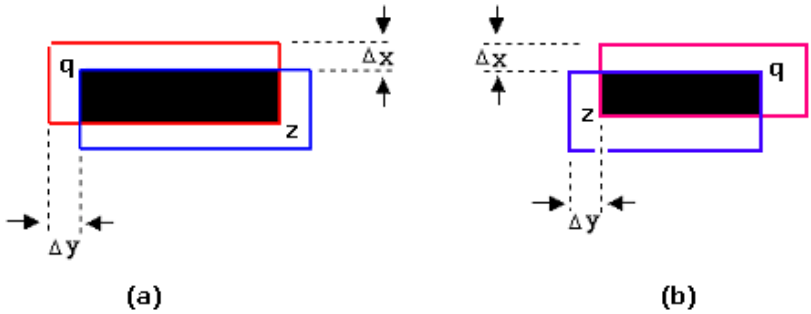


Fig. 2. Two translation shift types, (a) $\Delta x \geq 0, \Delta y \geq 0$; (b) $\Delta x \geq 0, \Delta y < 0$

Let $I(x, y)$ and $I^{'}(x, y)$ are two adjacent SFFs that differ only by an integer TSS $(\Delta x, \Delta y)$ (Fig.2), that is, their mutual support (the black region in Fig.2) satisfies:

$$I^{'}(x, y) = I(x + \Delta x, y + \Delta y) \tag{1}$$

where

$$1 \leq x \leq H - \Delta x, \quad \max\{1, 1 - \Delta y\} \leq y \leq \min\{W, W - \Delta y\}$$

In a general way, there are only two ways to swipe a finger through the sensor area: from the top down or, conversely, from the bottom up. For the sake of argument, we define Ω as a search space:

$$\Omega = \{(\Delta x, \Delta y) | 0 \leq \Delta x < H, 0 \leq |\Delta y| \leq W_p < W\} \tag{2}$$

where W and H are the height and width of each frame, W_p is a given upper bound of $|\Delta y|$. Thus, the MAE of the overlap between I and $I^{'}$ is:

$$\mathcal{M}(\Delta x, \Delta y) = \frac{\sum_{k=1}^{H-\Delta x} \sum_{l=\max\{1,1-\Delta y\}}^{\min\{W,W-\Delta y\}} |I(k + \Delta x, l + \Delta y) - I^{'}(k, l)|}{(H - \Delta x)(W - |\Delta y|)} \tag{3}$$

Then the optimal integer TSS $(\Delta x_1, \Delta y_1)$ is computed by minimizing the $\mathcal{M}(\Delta x, \Delta y)$, that is, let

$$\mathcal{M}_m = \min_{(\Delta x, \Delta y) \in \Omega} \{\mathcal{M}(\Delta x, \Delta y)\} \tag{4}$$

and we can get

$$(\Delta x_1, \Delta y_1) = \{(\Delta x, \Delta y) \big| \mathcal{M}(\Delta x, \Delta y) = \mathcal{M}_m, (\Delta x, \Delta y) \in \Omega\} \tag{5}$$

Finding the minimum MAE is a two-dimensional optimization problem. The only method yielding global extreme solution is an exhaustive search over the entire search space Ω. Although it is computationally demanding, it is often used if only translations are to be estimated[3]. However, for the real-time applications, the mosaicing speed should be fast enough to satisfy the on-line processing. Because the SFFs are contiguous, the integer TSS in horizontal direction Δy is often very small. Extensive experiments show that $W_p = 8$ is enough for the great majority applications. In this case, the search is just over the small window of $H \times W_p$ size. Moreover, we can optimize the search process as follows: (i) compute the $\mathcal{M}(\triangle x, \triangle y)$ only at these points where the $\triangle x$ and $\triangle y$ are both even. Let the optimal integer TSS in this case is $(\Delta x_2, \Delta y_2)$; (ii) search the minimum MAE in the small window Ω_W of size 3×3, and the center of Ω_W is $(\Delta x_2, \Delta y_2)$; (iii) select $(\Delta x_1, \Delta y_1)$ whose corresponding MAE is the minimum as the optimal integer TSS evaluation.

3 The Subpixel Level Mosaicing

Many methods have been developed to estimate the TSS between similar images. The PCM is a popular choice due to its robust performance and computational simplicity. The idea behind the PCM is quite simple and is based on the Fourier shift property, which states that a shift in the coordinate frames of two functions is transformed in the Fourier domain as linear phase differences.

Let the corresponding Fourier transforms of $I(x, y)$ and $I^{'}(x, y)$, denoted $\mathcal{I}$ and $\mathcal{I}^{'}$ respectively, are related by

$$\mathcal{I}^{'}(u, v) = \mathcal{I}(u, v) exp\{-j(u\psi_x + v\psi_y\} \tag{6}$$

where (ψ_x, ψ_y) are the TSS that occur between $\mathcal{I}$ and $\mathcal{I}^{'}$. The normalized cross power spectrum of $I(x, y)$ and $I^{'}(x, y)$ is given by

$$\mathcal{Q}(u, v) = \frac{\mathcal{I}^{'}(u, v)\mathcal{I}(u, v)^*}{|\mathcal{I}(u, v)\mathcal{I}(u, v)^*|} = exp\{-j(u\psi_x + v\psi_y)\} \tag{7}$$

where $\mathcal{Q}$ is also called normalized phase correlation matrix. Traditionally, the approach to evaluate the TSS, which is more practical and also more robust to noise, is to first inverse Fourier transform of $\mathcal{Q}$ [4]. It is then a simple matter to determine (ψ_x, ψ_y), since from (7) the result is $\delta(x - \psi_x, y - \psi_y)$ which is a Dirac delta function centered at (ψ_x, ψ_y).

A close inspection of (7) reveals that the noise-free model for $\mathcal{Q}$ is in fact a rank one matrix[10]. Each element in $\mathcal{Q}$ can be separated as

$$\mathcal{Q}(u, v) = exp\{-ju\psi_x\}\{-jv\psi_y\} \tag{8}$$

This allows the definition of two vectors: $q_x(u) = exp\{-ju\psi_x\}, q_y(v) = exp\{jv\psi_y\}$, then (7) can be rewritten as $\mathcal{Q} = q_x q_y^H$. This allows one to rewrite (6) as $\mathcal{I}^{'} = (q_x q_y^H) \circ \mathcal{I}$, where $\{\cdot\}^H$ denotes a complex-conjugate transpose, and $\circ$ indicates an element-by-element product. Therefore, the problem of finding the exact TSS between two images is recast as finding the rank one approximation of the normalized phase correlation matrix, $\mathcal{Q}$.

For a given singular vector, $\mathcal{V}$, we find the coefficients of a polynomial $P(x)$ of degree n to fit the vector, $P(x(i))$ to $\mathcal{V}(i)$, in a least squares sense. The result P is a row vector of length $n+1$ containing the polynomial coefficients in descending powers, that is, $P(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1}$. In our experiment, $n = 1$, that is, we use a line $P(x) = p_1 x + p_2$ to fit $\mathcal{V}$, p_1 and p_2 are the slope and abscissa of the fitted line, respectively. Let $\mathcal{P} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$, we construct the set of normal equations, $\mathcal{R}\mathcal{P} = unwrap\{\angle\mathcal{V}\}$, where the rows of $\mathcal{R}$ are equal to $[r, 1]$ for $r = \{0, 1, 2, \cdots, (s-1)\}$, s is equal to the length of $\mathcal{V}$. Let $\mathcal{B} = unwrap\{\angle\mathcal{V}\}$, then this system is solved to give $\mathcal{P} = (\mathcal{R}^T\mathcal{R})^{-1}\mathcal{R}^T\mathcal{B}$.

The slope of the fitted line, p_1 maps to the non-integer TSS. Specifically, $\psi_x = p_1 H/(2\pi)$ for the case $\mathcal{V} = q_x$, and $\psi_y = p_1 W/(2\pi)$ for the case $\mathcal{V} = q_y$.

4 The Hybrid Mosaicing Scheme

The most remarkable property of the PCM is the accuracy of the evaluated TSS. A second important property is its robustness to noise, therefore, the PCM is suitable for registration across different spectral bands. Using the convolution theorem, the method can also handle blurred images[4]. However, these methods based on phase correlation have a fatal limitation: the corresponding overlap between two images to be registered must be 30% bigger than the smaller image size[13]. Fortunately, the method based on MAE overcomes the limitation.

Based on the merits and faults of these two methods, we develop a hybrid scheme to mosaic the stream of SFFs so that *i*)the overlap between two adjacent frames can be very small down to 5% of the frame size, and *ii*)the registration precision can be subpixel level. The full scheme consists of these major steps:
1)Use the algorithm described in Section 2 to evaluate an integer TSS $(\Delta x_1, \Delta y_1)$;
2)Use the algorithm described in Section 3 to compute a non-integer TSS (ψ_x, ψ_y);
3)Select (x_0, y_0) as the last optimal TSS:

$$x_0 = \begin{cases} \psi_x & if|\psi_x - \Delta x_1| < 1 \\ \Delta x_1 & otherwise \end{cases}, \quad y_0 = \begin{cases} \psi_y & if|\psi_y - \Delta y_1| < 1 \\ \Delta y_1 & otherwise \end{cases} \tag{9}$$

4)Register two adjacent SFFs using the optimal TSS (x_0, y_0).
5)Mosaic the stream of SFFs to be an integrated fingerprint by iterating two adjacent SFFs registration.

5 Implementation Issues

The quality of the linear fit depends on the linearity of the unwrapped phase vector. In practice, the implicit eigen-filtering nature of identifying the dominant singular vectors of $\mathcal{Q}$ provides the unwrapping algorithm with less noisy data[10]. However, two dominant spectrum corruption sources remain: aliasing and edge effects. The ability of our scheme to handle both is detailed below.

5.1 Elliptoid Masking

Stone, et.al.,[5], recommend *masking* the phase correlation matrix, $\mathcal{Q}$, to restrict the spectrum components corrupted by aliasing from the shift estimation. This mask captures the components of $\mathcal{I}$ with magnitude larger than a given threshold α that are present within a radius $r = 0.6 \min\{W/2, H/2\}$ of the spectrum origin. However, it isn't suitable to SFFs because the height of each frame is much smaller than its width. So, we use an elliptoid *masking* to restrict the phase correlation matrix, that is, given two weight coefficients, κ_1 and κ_2,

$$\tau(i,j) = \frac{(i - H/2)^2}{(\kappa_1 H/2)^2} + \frac{(j - W/2)^2}{(\kappa_2 W/2)^2}, \quad 0 \le \kappa_1, \kappa_2 \le 1 \tag{10}$$

where (i, j) are the spatial domain coordinates. In our experiment, $\kappa_1 = 0.9, \kappa_2 = 0.8$. Let the mask matrix is $\mathcal{G}$, then

$$\mathcal{G}(i,j) = \begin{cases} 0 \text{ if } \tau(i,j) > 1 \\ 1 \quad \text{otherwise} \end{cases} \tag{11}$$

This masking is applied to the matrix $\mathcal{Q}$: $\mathcal{Q} = \mathcal{Q} \circ \mathcal{G}$, that is, only those components within the ellipse are utilized in the linear phase angle determination.

5.2 Kaiser Window

Image features close to the image edge can have a negative effect on the ability to evaluate TSS between two frames. For images acquired via optical methods, Stone,et.al.,[5], recommend applying a 2D spatial Blackman or Blackman-Harris window to the image before transforming the image to the Fourier domain. Unfortunately, this spatial window removes a significant amount of the signal energy, and isn't suitable for SFFs. Here, we apply a 2D spatial Kaiser window to the frames, which is a relatively simple approximation of the prolate spheroidal functions. For discrete time the Kaiser window $\mathcal{K}(N, \beta)$ is expressed as[6]:

$$\mathcal{K}(N,\beta) = \mathfrak{B}_0\left(\beta\sqrt{1 - \frac{4n^2}{(N-1)^2}}\right) \Big/ \mathfrak{B}_0(\beta), \quad -\frac{N-1}{2} \le n \le \frac{N-1}{2} \tag{12}$$

where N is the window length which controls the main lobe width of the window; β is a parameter which controls the amplitude for the sidelobes; $\mathfrak{B}_0(x)$ is the modified zeroth-order Bessel function.

For a 2D spatial Kaiser window: $\mathcal{W}_k = \mathcal{K}(\lambda_1 H, \beta)^T \circ \mathcal{K}(\lambda_2 W, \beta)$, $0 \le \lambda_1, \lambda_2 \le 1$, where λ_1 and λ_2 are two weight coefficients. In our experiment, $\lambda_1 = 0.9$, $\lambda_2 = 0.8$ and $\beta = 2$.

6 Experimental Result

In order to verify the scheme experimentally, some simulations are performed by shifting images using a procedure similar to the one used in[4]: $g_m = \mathcal{H} * f_m$, where m is the frame number, f_m are shifted versions of a high-resolution image convolved by a blurring kernel $\mathcal{H}$ which characterizes image degradations (In our simulations, we choose the white noise with gaussian distribution as $\mathcal{H}$). Each frame g_m is then downsampled at a predetermined rate so that the correspondence between different frames is reduced to subpixel level. We use three different mosaicing methods: $a)\Gamma_m$: based on the MAE described in Section 2, $b)\Gamma_p$: based on the popular PCM[11], $c)\Gamma_h$: our hybrid scheme.

Table.1 shows the evaluated TSS using the above three different mosaicing methods: $(\Delta x, \Delta y)$, (ψ_x, ψ_y) and (x_0, y_0) are evaluated by the Γ_m, Γ_p and Γ_h methods respectively. (Υ_x, Υ_y) are the predetermined TSS and γ denotes the ratio of the overlap to the each frame size. From the experimental data, we can see that $(i)$the evaluated TSS (ψ_x, ψ_y) using the Γ_p method are not accurate when $\gamma < 0.10$ while the other two methods haven't this limitation; $(ii)$the evaluated non-integer TSS (x_0, y_0) using our hybrid scheme are much more precise than $(\Delta x, \Delta y)$ and (ψ_x, ψ_y), that is, our hybrid scheme has better accuracy compared to the other two methods.

Table 1. Performance comparison with simulations: Γ_m, Γ_p and Γ_h

(Υ_x, Υ_y)	$(\Delta x, \Delta y)$	(ψ_x, ψ_y)	(x_0, y_0)	γ
(0.0000,0.0000)	(0,0)	(0,0)	(0,-1.2673e-016)	1
(0.0000,0.5000)	(0,0)	(0,0)	(0.0000,0.5156)	0.9967
(0.2500,0.2500)	(0,0)	(0,0)	(0.2502,0.2496)	0.9906
(0.5000,0.2500)	(1,0)	(1,0)	(0.4952,0.2398)	0.9828
(0.5000,0.5000)	(1,0)	(1,1)	(0.5166,0.4782)	0.9811
(5.0000,5.0000)	(5,5)	(5,5)	(5.0213,5.0090)	0.8160
(10.0000,5.0000)	(10,5)	(10,5)	(10.0127,4.9882)	0.6649
(14.0000,5.0000)	(14,5)	(14,5)	(14.0116,4.9665)	0.5440
(18.0000,5.0000)	(18,5)	(18,5)	(18.0193,5.0101)	0.4231
(20.0000,5.0000)	(20,5)	(20,5)	(20.0236,5.0106)	0.3627
(22.0000,5.0000)	(22,5)	(22,5)	(22.0281,5.0073)	0.3022
(24.0000,5.0000)	(24,5)	(24,5)	(24.0275,5.0078)	0.2418
(26.0000,5.0000)	(26,5)	(26,5)	(26.0000,5.0113)	0.1813
(28.0000,5.0000)	(28,5)	(28,5)	(28.0000,5.0186)	0.1209
(29.0000,5.0000)	(29,5)	(0,0)	(29.0000,5.0143)	0.0907
(30.0000,5.0000)	(30,5)	(0,2)	(30.0000,5.0163)	0.0604
(31.0000,5.0000)	(31,5)	(0,8)	(31.0000,5.0133)	0.0302

The above simulations are ideal so that the evaluated TSS (ψ_x, ψ_y) using the Γ_p method still works well when $0.10 \leq \gamma < 0.30$. However, many types of noise exist in real data, e.g., uniform variations of illumination, offsets in average

Table 2. Performance comparison with real data: Γ_p, Γ_p and Γ_h

$(\Delta x, \Delta y)$	(ψ_x, ψ_y)	(x_0, y_0)	γ
(10,-1)	(10,-1)	(10.0144,-1.0002)	0.6916
(11,-1)	(11,-1)	(11.0133,-0.9966)	0.6601
(11,-1)	(11,-1)	(11.0070,-1.0018)	0.6604
(9,-1)	(9,-1)	(9.0129,-0.9991)	0.7231
(8,0)	(8,0)	(8.0115,0.0051)	0.7496
(9,-1)	(8,-1)	(8.5236,-1.0144)	0.7385
(10,0)	(10,0)	(10.0152,0.0028)	0.6870
(10,0)	(10,0)	(10.0155,0.0134)	0.6870
(9,0)	(9,0)	(9.0130,0.0037)	0.7183
(10,-1)	(10,-1)	(10.0102,-0.9919)	0.6917
(13,0)	(13,0)	(13.0128,0.0143)	0.5933
(13,0)	(13,0)	(13.0118,0.0109)	0.5933
(18,-1)	(18,-1)	(18.0232,-0.9666)	0.4396
(9,0)	(9,0)	(9.0130,0.0037)	0.7183
(7,0)	(7,0)	(7.0107,0.0081)	0.7809
(23,0)	(0,0)	(23.0000,0.0096)	0.2805
(24,-1)	(0,-1)	(24.0000,-1.0012)	0.2509

intensity, and fixed gain errors due to calibration. So, further experimentations are performed on the real streams of SFFs captured by the FPC1031B swipe fingerprint sensor of FINGERPRINT CARDS[16]. For example, Fig.1 is a real stream of SFFs captured by this sensor. Table.2 shows the evaluated TSS parameters using three different methods: Γ_m, Γ_p and Γ_h. Since the actual TSS parameters were unknown for these real data, the performance was evaluated by the visible seams in the mosaiced swipe fingerprint image, that is, the less seams the better performance. From Fig.3, we can see that there are visible seams in the mosaiced fingerprint image using the Γ_m and Γ_p methods while the mosaiced fingerprint image using our hybrid scheme Γ_h hasn't.

Our scheme has been implemented and applied to actual SFFs. The scheme is fast enough for real-time application: it can mosaic a series of 40 pre-captured frames into a high quality fingerprint image in less than one second which is much faster than the real-time requirement of 30 frames per second. Fig.4 shows six mosaiced fingerprint images using our hybrid scheme.

7 Conclusions

In this paper we have described a hybrid mosaicing scheme for the stream of swipe fingerprint frames, which uses the minimum mean absolute error as the registration criterion to find an integer translation shift that registers two adjacent swipe fingerprint frames to the nearest integral pixel coordinates while the extension phase correlation method with singular value decomposition is used to get subpixel precision. Experimental data show that our scheme is reliable and less time consumption. So it is suitable for the real-time applications.

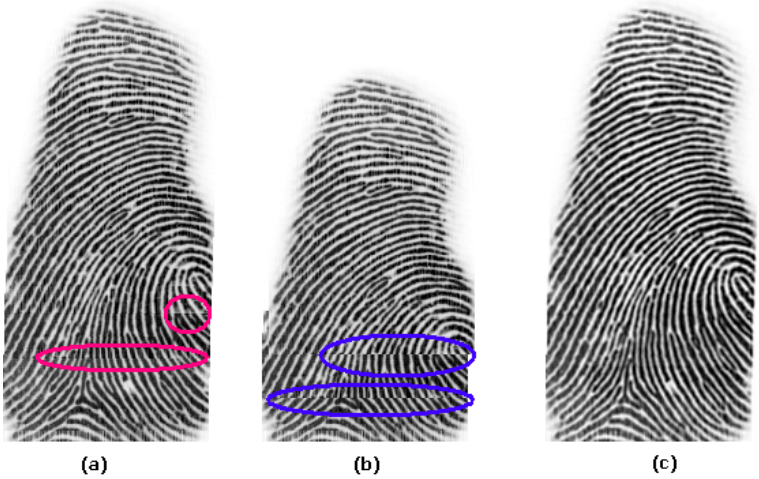


Fig. 3. The mosaiced fingerprint images: Γ_m, Γ_p, Γ_h



Fig. 4. The mosaiced fingerprint images using our hybrid scheme

However, in our mosaicing scheme it is assumed that no rotation, scaling and shear effects are present in individual frames of the swipe fingerprint images. Due to the nature of the skin, some amount of elastic deformation is expected. This has not been taken into account and this is a very serious drawback of the proposed scheme. In future work, we will discuss the swipe fingerprint mosaicing when the variations in swiping speeds and elastic deformation are taken into account.

References

1. Jain.A, Hong.L, and Bolle.R: *On-Line Fingerprint Verification*. IEEE Trans.PAMI, Vol.19, No.4, Apr.1997, pp.302-314.
2. Arun Ross: *Information Fusion in Fingerprint Authentication*. Ph.D thesis, Michigan State University 2003.
3. Zitova.B, Flusser.J: *Image Registration Methods: A Survey*. Image and Vision Computing, Vol.21, No. 11, 2003, pp.977-1000.
4. Foroosh.H, Zerubia.J.B, and Berthod.M: *Extension of phase correlation to subpixel registration*. IEEE Trans. Image Processing, Vol.11, No.3, Mar.2002, pp. 188-200.
5. Stone.H.S, Orchard.M.T, Chang.E.-C, and Martucci.S.A: *A fast direct Fourier-based algorithm for subpixel registration of images*. IEEE Trans. Geosci. Remote Sensing, Vol.39, No.10, Oct.2001, pp.2235-2243. Oct.2001, pp.2235-2243.
6. Kaiser.J.F: *Nonrecursive Digital Filter Design Using the I_0- sinh Window Function*. Proc.1974 IEEE Symp.Circuits and Systems, (April 1974), pp.20-23.
7. Toh.K.A, Yau.W.Y, Jiang.X.D, Chen.T.P, Lu.J and Lim.E: *Minutiae data synthesis for fingerprint identification applications*. in Proceedings of International Conference on Image Processing, Vol.III, pp. 262-265, 2001.
8. Lisa Gottesfeld Brown: *A Survey of Image Registration Techniques*. ACM Comput. Surv. Vol.24, No. 4, 1992, pp.325-376.
9. Ratha.N.K, Connell.J.H, and Bolle.R.M: *Image Mosaicing for Rolled Fingerprint Construction*. in Proc. of 14th International Conference on Pattern Recognition, Vol.2,1998, pp.1651-1653.
10. Hoge.W.S: *Subspace identification extension to the phase correlation method*. IEEE Trans. Medical Imaging, Vol.22, No.2, Feb.2003, pp.277-280.
11. Reddy.B.S, Chatterji.B.N: *An FFT-based technique for translation, rotation, and scale-invariant image registration*. IEEE Trans. Image Processing, Vol.5, No.8, Aug.1996, pp.1266-1271.
12. Yau.W.Y, Toh.K.A, Jiang.X.D, Chen.T.P and Lu.J: *On Fingerprint Template Synthesis*. in Proceedings of Sixth International Conference on Control, Automation, Robotics and Vision (ICARCV 2000). Singapore. 5-8 December 2000.
13. Keller.Y, Averbuch.A, Israeli.M: *pseudo-polar based estimation of large translations rotations and scalings in images*. to appear in IEEE Trans. On Image Processing, January 2005.
14. Zsolt Miklos and Kovacs-Vajna: *A Fingerprint Verification System Based on Triangular Matching and Dynamic Time Warping*. IEEE Trans.PAMI, Vol.22, No.11, November 2000.
15. *http://www.authentec.com*
16. *http://www.fingerprints.com*

A Study on Multi-unit Fingerprint Verification

Kangrok Lee[1], Kang Ryoung Park[2], Jain Jang[3],
Sanghoon Lee[1], and Jaihie Kim[1]

[1] Biometrics Engineering Research Center(BERC),
Department of Electrical and Electronic Engineering,
Yonsei University,
134, Sinchon-dong Seodaemun-gu, Seoul 120-749, Korea,
{plusu,hoony,jhkim}@yonsei.ac.kr
[2] Division of Media Technology, Sangmyung University,
7, Hongji-dong, Chongro-gu, Seoul 110-743, Korea,
Biometrics Engineering Research Center(BERC)
parkgr@smu.ac.kr
[3] Biometrics Engineering Research Center(BERC),
Division of Computer and Information Engineering,
Yonsei University,
134, Sinchon-dong Seodaemun-gu, Seoul 120-749, Korea,
jijang@cs.yonsei.ac.kr

Abstract. Previous fingerprint verification systems have achieved good results, but these have been affected by the quality of input data. Fingerprint verification systems that use many fingers or multiple impressions of the same finger are more efficient and reliable than systems that use a single finger. However, multiple impressions give inconvenience to the user and increase the overall verification time. Therefore, we use only two fingers (multi-unit fusion) to improve performance of the fingerprint verification system. We show that performance can be improved by selecting a better quality fingerprint image of two fingerprints. Also, we propose a new quality checking algorithm composed of three stages. Our experimental results show that when the quality checking algorithm is performed by selecting a better quality fingerprint image of two fingerprints, there is a significance improvement in performance of the fingerprint verification system.

1 Introduction

A reliable automatic fingerprint identification system is critical in forensic, civilian, and commercial applications such as criminal investigation, issuing driver's licenses, welfare disbursement, resident registration, credit cards, PDA usage, and access control[1]. Fingerprint verification is much more reliable than other kinds of personal identification methods such as signature, face, and speech[2]. With recent advances in solid-state sensor technology, fingerprint sensors can now be miniaturized and made cheaper. However, due to the small size of solid-state sensors, only a part of the finger tip is captured in the image, and as a

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 141–150, 2005.
© Springer-Verlag Berlin Heidelberg 2005

result, system performance is diminished. In addition, the various sources of noise to image acquisition and the vulnerability of the feature extractor to noise and distortion in fingerprint images make very difficult to achieve a desirable false rejection rate(FRR),when the specified false acceptance rate(FAR) is very low. To solve those problems, it is necessary to combine more than two fingers or multiple impressions of the same finger.

Jain et al.[3] proposed a classifier combination at the decision level and emphasized the importance of classifier selection in the classifier combination. Hong et al.[4] combined fingerprints and faces as a means of identification. Bigun et al.[5] proposed a Bayesian framework scheme to combine different pieces of evidence. On the contrary, J. Daugman[6] insisted that a biometric system using only a strong biometric in the classifier level is superior to using biometrics with respect to both the FAR and the FRR. However, this may not be applied when the quality of the input data is poor. Therefore we propose that the performance of the verification system can be improved by selecting a better quality fingerprint image of two fingerprints. Also, we prove theoretically and empirically that the proposed method is more accurate than methods that use an AND rule or an OR rule and a single finger in the respect of both the FAR and the FRR.

2 Performance Improvement of Combined Biometrics in Fingerprint Verification

First, we explain the theory of the proposed method and prove the validity of the method with respect to both the FAR and the FRR. As shown in Fig. 1, we use two fingerprint images that were sequentially acquired by a sensor. In real-time fingerprint verification systems, it does not take much time to obtain two fingerprint images. Once the images have been obtained, we operate the histogram stretching and the median filtering to improve the contrast of the fingerprint images in the preprocessing stage. Feature extraction is followed by matching in the fingerprint verification system. At the feature extraction stage, we find the minutiae as feature.[7]. With true minutiae, we estimate the quality of the fingerprint images to select a better quality fingerprint image between two fingerprints. At the matching stage, we compare the minutiae points extracted from the good fingerprint image with those in the fingerprint template, and calculate the matching score[8]. Supposing there are two biometrics, F1 and F2. F1 is a fingerprint verification system that uses good quality left fingers and F2 is also a fingerprint verification system that uses right fingers of good and bad qualities. Bad quality means that the input fingerprint image is not good enough to be identified due to excessive humidity of the sensor surface or incomplete fingerprint input.So, we suppose that the FAR and the FRR of F1 and F2 are followed by,

$$\begin{aligned} &\mathrm{F1}: P_{\mathrm{F1}}(\mathrm{FA}) = P_{\mathrm{F1}}(\mathrm{FR}) = \mathrm{P} \\ &\mathrm{F2}: \begin{cases} P_{\mathrm{F2}}(\mathrm{FA}) = P_{\mathrm{F2}}(\mathrm{FR}) = \mathrm{P}\ (\text{with good quality fingerprints}) \\ P_{\mathrm{F2}}(\mathrm{FR}) = \mathrm{Q},\ P_{\mathrm{F2}}(\mathrm{FA}) = \mathrm{P}\ (\text{with bad quality fingerprints}) \\ (\mathrm{P} < \mathrm{Q}\ ,\ \mathrm{P} < 1\ ,\ \mathrm{Q} \leq 1\ ,\ \mathrm{Q} = \mathrm{tP}\ (\mathrm{t} > 1)) \end{cases} \end{aligned} \tag{1}$$

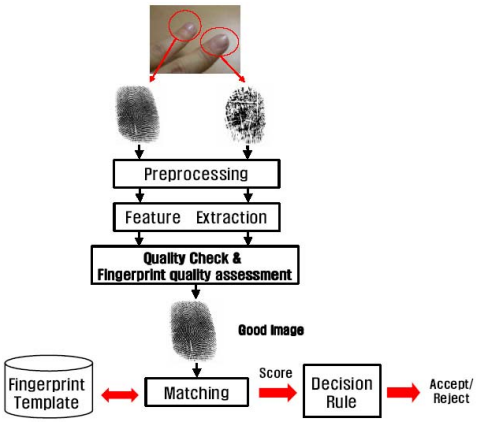


Fig. 1. Fingerprint verification system using multiple fingerprint units

If there are genuine tests of numbers (the number of good data: X, the number of bad data: M − X) and imposter tests of numbers (the number of good data: Y, the number of bad data: M − Y), then we can calculate the Total Error Counts (TEC), as shown below.

Case (1) Using the F1 at the first trial:

$$\mathrm{TEC} = \mathrm{M} \times \mathrm{P} + \mathrm{M} \times \mathrm{P} = 2\mathrm{MP}$$

Case (2) Using the F2 at the second trial:

$$\mathrm{TEC} = \mathrm{X} \times \mathrm{P} + (\mathrm{M} - \mathrm{X}) \times \mathrm{Q} + \mathrm{M} \times \mathrm{P} = (\mathrm{M} + \mathrm{X})\mathrm{P} + (\mathrm{M} - \mathrm{X})\mathrm{Q}$$

Case (3) Using an OR rule:

$$\mathrm{TEC} = \mathrm{XP}^2 + (\mathrm{M} - \mathrm{X})\,\mathrm{PQ} + \mathrm{M}\,(2\mathrm{P} - \mathrm{P}^2)$$

Case (4) Using an AND rule:

$$\mathrm{TEC} = \mathrm{X}(2\mathrm{P} - \mathrm{P}^2) + (\mathrm{M} - \mathrm{X})(\mathrm{P} + \mathrm{Q} - \mathrm{PQ}) + \mathrm{MP}^2$$

Case (5) Using the proposed method:

Considering the quality check error(P_{E1},P_{E2},P_{E3}), we can analyze four cases. We can calculate the TEC by the case 1) case 4).

(1) When the F1 and F2 are actually good quality fingerprint data and they are identified as the good quality data by the quality checking algorithm. Our system select the F1 or F2. In this case, both the F1 and F2 are good quality data and the FAR/FRR are P as shown in Eq.(1). So, the TEC is followed by

$$\mathrm{TEC} = (\mathrm{X} + \mathrm{Y})\mathrm{P}$$

(2) When the F1 is identified as a good quality fingerprint and the F2 is misidentified as a bad quality fingerprint(actually, F2 is a good quality fingerprint having the number of X and Y as shown in Eq.(1)) by the quality checking algorithm(having the error of P_{E1}), our system will select the F1 because it is

good quality data and the FAR/FRR are P as shown in Eq.(1). So, the TEC is followed by

$$\mathrm{TEC} = ((\mathrm{X}+\mathrm{Y})\mathrm{P})\mathrm{P}_{\mathrm{E1}}$$

(3) When the F1 is identified as a good quality fingerprint, the F2 is misidentified as a good quality fingerprint(actually, F2 is a bad quality fingerprint having the number of $\mathrm{M}-\mathrm{X}$ and $\mathrm{M}-\mathrm{Y}$ as shown in Eq.(1)) and our system erroneously selects the F2 as better quality fingerprint than the F1 by the quality checking algorithm (having the error of P_{E2}), the FRR is increased to Q, while the FAR is P as shown in Eq.(1). So, the TEC is followed by

$$\mathrm{TEC} = \{(\mathrm{M}-\mathrm{X})\mathrm{Q}+(\mathrm{M}-\mathrm{Y})\mathrm{P}\}\mathrm{P}_{\mathrm{E2}}$$

(4) When the F1 is identified as a good quality fingerprint, the F2 is misidentified as a good quality fingerprint(actually, F2 is the bad quality fingerprint having the number of $\mathrm{M}-\mathrm{X}$ and $\mathrm{M}-\mathrm{Y}$ as shown in Eq.(1)) and our system correctly selects the F1 as a better quality fingerprint than the F2 by the quality checking algorithm (having the error of P_{E3}), the FAR and FRR are P as shown in Eq.(1). So, the TEC is followed by

$$\mathrm{TEC} = \{(\mathrm{M}-\mathrm{X})\mathrm{P}+(\mathrm{M}-\mathrm{Y})\mathrm{P}\}\mathrm{P}_{\mathrm{E3}}$$

We prove that the TEC of the proposed method (case(5)-**(1)**∼**(4)**) is smaller than the TEC of the other method (case(1)∼(4)). If we compare case(1)∼(4) with case(5)-**(1)**, we get the difference function (f(P, Q, M, X)) between the TEC of case(1)∼(4) and that of case(5)-**(1)** as shown in Eq.(2)∼(5).

$$\begin{aligned}\mathrm{f(P,\ Q,\ M,\ X)} &= \mathrm{TEC\ of\ case(1)} - \mathrm{TEC\ of\ case(5)\text{-}\mathbf{(1)}} \\ &= 2\mathrm{MP} - (\mathrm{X}+\mathrm{Y})\mathrm{P} \\ &= (\mathrm{M}-\mathrm{X})\mathrm{P} + (\mathrm{M}-\mathrm{Y})\mathrm{P}\end{aligned} \tag{2}$$

$$\begin{aligned}\mathrm{f(P,Q,M,X)} &= \mathrm{TEC\ of\ case\,(2)} - \mathrm{TEC\ of\ case(5)\text{-}\mathbf{(1)}} \\ &= (\mathrm{M}+\mathrm{X})\mathrm{P} + (\mathrm{M}-\mathrm{X})\mathrm{Q} - (\mathrm{X}+\mathrm{Y})\mathrm{P} \\ &= (\mathrm{M}-\mathrm{Y})\mathrm{P} + (\mathrm{M}-\mathrm{X})\mathrm{Q}\end{aligned} \tag{3}$$

$$\begin{aligned}\mathrm{f(P,Q,M,X)} &= \mathrm{TEC\ of\ case\,(3)} - \mathrm{TEC\ of\ case(5)\text{-}\mathbf{(1)}} \\ &= \mathrm{XP}^2 + (\mathrm{M}-\mathrm{X})\mathrm{PQ} + \mathrm{M}(2\mathrm{P}-\mathrm{P}^2) - (\mathrm{X}+\mathrm{Y})\mathrm{P} \\ &= (\mathrm{M}-\mathrm{X})\mathrm{P}(\mathrm{Q}-\mathrm{P}) + (\mathrm{M}-\mathrm{X})\mathrm{P} + (\mathrm{M}-\mathrm{Y})\mathrm{P}\end{aligned} \tag{4}$$

$$\begin{aligned}\mathrm{f(P,Q,M,X)} &= \mathrm{TEC\ of\ case\,(4)} - \mathrm{TEC\ of\ case(5)\text{-}\mathbf{(1)}} \\ &= \mathrm{X}(2\mathrm{P}-\mathrm{P}^2) + (\mathrm{P}+\mathrm{Q}-\mathrm{PQ})(\mathrm{M}-\mathrm{X}) + \mathrm{MP}^2 - (\mathrm{X}+\mathrm{Y})\mathrm{P} \\ &= (\mathrm{M}-\mathrm{X})\mathrm{P}^2 + \mathrm{P}(\mathrm{M}-\mathrm{Y}) + \mathrm{Q}(1-\mathrm{P})(\mathrm{M}-\mathrm{X})\end{aligned} \tag{5}$$

In the Eq.(2)∼(5), all the $(\mathrm{M}-\mathrm{X})$, $(\mathrm{M}-\mathrm{Y})$, Q, P, $(\mathrm{Q}-\mathrm{P})$, $(\mathrm{M}+\mathrm{X})$ and $(1-\mathrm{P})$ are greater than 0. So, all the $\mathrm{f(P,Q,M,X)}s$ of Eq.(2)∼(5) are greater than 0 and the TEC of case(5)-**(1)** is smaller than those of case(1)∼(4), consequently. Only when all the fingerprint images are composed of good quality data at genuine and imposter tests ($\mathrm{X}=\mathrm{M},\ \mathrm{Y}=\mathrm{M}$), the TEC of case(5)-**(1)** is same

to those of case(1)∼(4). However, in real fingerprint verification systems, this is impossible. Due to excessive humidity, dust or sweat on the sensor surface, bad quality images are frequently obtained($M - X > 0$,$M - Y > 0$). Also, in case we compare the TECs of case(1)∼(4) to that of case(5)-**(2)** by same method as shown in Eq.(2) ∼(5), we can know the latter (that of case(5)-**(2)**) is smaller than the formers (those of case(1)∼(4)). Therefore, we can conclude that the proposed method shows better performance than case(1)∼(4). Furthermore, if we compare the TECs of case(1)∼(4) to that of case(5)-**(3)**, we get the difference functions as shown in Eq.(6)∼(9).

$$\begin{aligned} f(P,Q,M,X) &= \text{TEC of case}\,(1) - \text{TEC of case(5)-}\mathbf{(3)} \\ &= 2MP - \{(M-X)Q + (M-Y)P\}P_{E2} \\ &= MP(1-P_{E2}) + MP(1-tP_{E2}) + XQP_{E2} + YPP_{E2} \\ &= MP(1-P_{E2}) + MPt(\tfrac{1}{t} - P_{E2}) + XQP_{E2} + YPP_{E2} \end{aligned} \tag{6}$$

$$\begin{aligned} f(P,Q,M,X) &= \text{TEC of case}\,(2) - \text{TEC of case(5)-}\mathbf{(3)} \\ &= (M+X)P + (M-X)Q - \{(M-X)Q + (M-Y)P\}P_{E2} \\ &= (M-X)Q(1-P_{E2}) + M(1-P_{E2})P + XP + YPP_{E2} \end{aligned} \tag{7}$$

$$\begin{aligned} f(P,Q,M,X) &= \text{TEC of case}\,(3) - \text{TEC of case(5)-}\mathbf{(3)} \\ &= XP^2 + (M-X)PQ + M(2P-P^2) - \{(M-X)Q + (M-Y)P\}P_{E2} \\ &= (M-X)P(Q-P) + MP(1-P_{E2}) + MPt(\tfrac{1}{t} - P_{E2}) + (XQ+YP)P_{E2} \end{aligned} \tag{8}$$

$$\begin{aligned} f(P,Q,M,X) &= \text{TEC of case}\,(4) - \text{TEC of case(5)-}\mathbf{(3)} \\ &= X(2P-P^2) + (P+Q-PQ)(M-X) + MP^2 - \{(M-X)Q + (M-Y)P\}P_{E2} \\ &= Q(1-P)(M-X) + MPt(\tfrac{1}{t} - P_{E2}) + P(X - MP_{E2}) + P^2(M-X) + (XQ+YP)P_{E2} \end{aligned} \tag{9}$$

The proposed method is superior to methods of case(1)∼(4), because the conditions of the difference function(Eq.(6)∼(9)) are bigger than 0 on condition that $\frac{P}{Q} > P_{E2} (= \frac{1}{t} > P_{E2}$, because $Q = tP$ as shown in Eq.(1)). So, in case we can make the quality check algorithm, of which the quality checking error(P_{E2}) is smaller than $1/t$(our experimental results show that P_{E2} is smaller than $1/t$), the proposed method shows the best performance compared to those of case(1)∼(4). Also, when we compare case(1)∼(4) with case(5)-**(4)**, we can know the proposed method produces better results than methods of case(1)∼(4). By conclusion, we can know that the proposed method of selecting better quality data can show the better performance than those of case(1)∼(4) on the condition of $\frac{1}{t} > P_{E2}$ theoretically.

3 Quality Checking Algorithm

In this section, we explain how to select the best quality fingerprint image between two fingerprints. The quality checking algorithm is composed of three stages as shown in Fig. 5 and section 3.1∼3.4. At the first stage, it examines the size of the foreground and background areas in the input fingerprint image as the number of foreground blocks. The second stage examines the number of good blocks by using the classifying method. Finally, we examine the number of true minutiae.



Fig. 2. Foreground/Background in the input fingerprint image: (a) Input fingerprint image, (b) Segmented image

3.1 Quality Checking of Foreground Areas

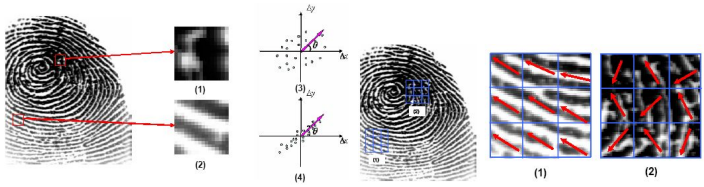
This section is corresponding to the first stage of Fig. 5. A fingerprint image can be separated into blocks of pixels. Although each foreground block can be calculated as a block direction, each background block cannot be calculated as a block direction. We decide how many foreground areas are obtained by the input fingerprint image using the number of foreground blocks. Fingerprint image quality is determined by foreground areas of the input fingerprint image by a sensor. If the foreground area in the input fingerprint image is large, probability for the existence of minutiae is increased. Fig. 2 shows how the large foreground area is included in the input fingerprint image. The input fingerprint image is divided into foreground and background areas using only variance of gray values. The variance($VAR(I)$) of gray values in the image block is calculated by Eq.(10).

$$\begin{aligned} M(I) &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} I(n,m) \\ VAR(I) &= \frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} (I(n,m) - M(I)) \end{aligned} \tag{10}$$

where, $I(n,m)$: the pixel gray value of n , m position , N : the block size

3.2 Quality Checking of a Local Fingerprint Image and a Global Fingerprint Image

In the section 3.2 and 3.3, it is corresponding to the second stage of Fig. 5. We account for the quality checking algorithm in the context of local and global fingerprint images. In fingerprint images, the difference in quality of the images is shown in Fig. 3(a). In Fig. 3(a)-(2), a block of good quality has uniform direction of the ridge structure. However a block of bad quality has irregular direction of the ridge structure in Fig. 3(a)-(1)and high curvature area. Minutiae detected in high curvature areas are not reliable. This is especially true of the core and delta regions of a fingerprint. The coherence is a measure of the local strength of the directional field [9]. This measurement, which is called the coherence (Coh as shown in Eq.(11)), presents how well all squared gradient vectors share the same orientation. Using those characteristics, we examine the quality of the fingerprint image by calculating the coherence of a pixel gradient within a block of image pixels(16 * 16 pixels). Coherence is the quality measurement of a local fingerprint



(a) (b)

Fig. 3. Quality checking method using local and global characteristics: (a) Quality checking method using local characteristics, (b) Quality checking method using global characteristics

image (Q_C) [9] ,which is estimated by variances and crosscovariances of G_x and G_y, averaged over the window W.

$$Coh = \frac{\sqrt{(G_{xx} - G_{yy})^2 + 4G_{xy}^2}}{G_{xx} + G_{yy}} \tag{11}$$

where, $G_{xx} = \sum_{W} G_x^2$, $G_{yy} = \sum_{W} G_y^2$ and $G_{xy} = \sum_{W} G_x G_y$
Generally, fingerprints have the characteristic that flows of ridges vary slowly. As shown in Fig. 3(b)-(1), in the good quality image, there is no variance of orientation among adjacent blocks. On the contrary, in bad quality image, there is no coherence ($Coh = 0$) among adjacent blocks as shown in Fig. 3(b)-(2)). Using these characteristics, we calculate the quality of a global fingerprint image by circular variance among adjacent blocks [10]. The method of calculating the quality measurement (Q_O) of a global fingerprint image is shown in Eq.(12).

$$Q_O = 1 - V \tag{12}$$

where, $V = 1 - \sqrt{\bar{C}^2 + \bar{S}^2}$: *Circular variance*, $\bar{C} = \frac{1}{W} \sum_{j=1}^{W} \cos\theta_j$, $\bar{S} = \frac{1}{W} \sum_{j=1}^{W} \sin\theta_j$
W: the number of block, θ_j: jth block direction of the adjacent block

3.3 Classifying Method of Local and Global Fingerprint Images

To consider local and global fingerprint characteristics, we calculate the mean and the covariance matrix using supervised-learning method. The quality of the region is defined by minutiae-based method. Fig. 4 shows the true minutiae and false minutiae. Good quality regions mean small blocks including true minutiae, and bad quality regions mean small blocks including false minutiae. It is assumed that both a local quality measurement and a global quality measurement have two-dimensional Gaussian random distributions. Q_C is local quality measurement and Q_O is global quality measurement. μ_C is local mean of quality and μ_O is global mean of quality.



Fig. 4. Minutiae points of manually-defined quality:(a) Original image (b) Enhanced binary image (Blue circle:False minutiae, Red rectangle:True Minutiae)

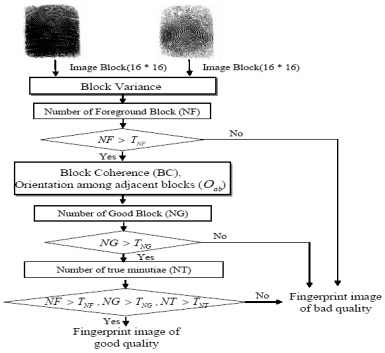


Fig. 5. Flowchart of the proposed quality checking algorithm

$$p(X|W_i) = \frac{1}{2\pi \times |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^t \Sigma^{-1}(X-\mu)\right] \tag{13}$$

where, $X = \begin{bmatrix} Q_C \; Q_O \end{bmatrix}^T$, $\mu = \begin{bmatrix} \mu_C \; \mu_O \end{bmatrix}^T$
W_0 is the class of good quality fingerprint images and W_1 is the class of bad quality fingerprint images. We use the Bayesian Theorem as a classification method in Eq.(14) to minimize errors, when the quality of a fingerprint image is tested by using each quality measurement out of 352 blocks.

$$\text{If } p(W_i|X) > p(W_j|X) \quad \text{then} \quad \text{select } W_i \tag{14}$$

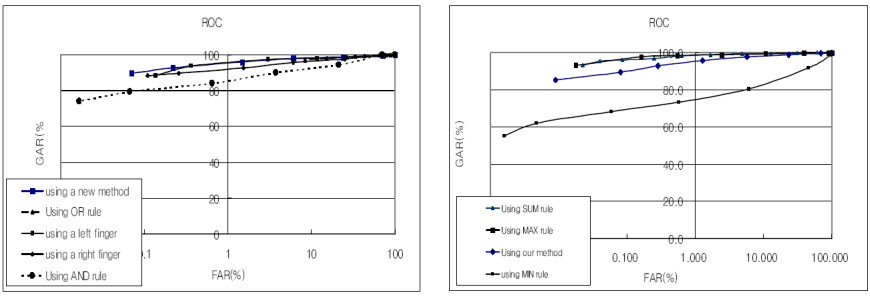
If the number of selected window is higher than the threshold value (Th = 304), the input image is classified as a good quality image.

3.4 Quality Checking of True Minutiae

At the final stage of quality checking method, we calculate the number of true minutiae. The more true minutiae, the more matching minutiae in the fingerprint image compared with the fingerprint template. To find a true minutiae, we extract the features (ridge ending, ridge bifurcation) and eliminate false minutiae in the fingerprint images[7]. The overall flowchart of the quality checking algorithm is shown in Fig. 5.

4 Experimental Results

A set of fingerprint images was acquired through the fingerprint sensor manufactured by Testech, Inc. We obtained 1,260 (20 fingerprints for 63 people) fingerprint images of varying quality. We used two fingerprint images as enrollment and rest of them used the test. We performed 1,008 authentic tests and 70,938 impostor tests. In Fig. 6(a), the proposed method shows better performance than others that use a single finger, an AND rule and an OR rule. As we see the Fig. 6(a), when we use the AND rule as a combination method, the GAR is lower than when we use the single fingerprint. It is due to characteristic of proposed system, which needs to be high security. A criterion of system error rate is selected that FRR is more than FAR instead of EER(Equal Error Rate). AND rule affects that FAR becomes less and FRR becomes more, so whole error rate is increased. Also, the fingerprint verification system using the proposed method showed better performance in comparison with others in terms of both the FRR and the FAR. In Fig. 6(b), the proposed method is inferior to methods that use an SUM and MAX rule[11]. This is because the overlapping area between the input fingerprint image and the fingerprint template is critical. Though the input fingerprint image has good quality at the verification stage, the matching score in the input fingerprint image is lower in case that the overlapping areas between the input fingerprint image and the fingerprint template are small. To solve this problem, we will consider the quality of the fingerprint template at the enrollment stage and the input fingerprint image at the verification stage. Although the proposed method is poorer than that using an SUM and MAX rule, it has many profits. Our proposed method has the advantage that it can be used even if we do not know the internal system of each multimodal biometrics different from that using SUM,MAX and MIN rule. However, in case of using SUM, MAX and MIN rule, we should know the internal output matching score from each multimodal biometrics.



(a) (b)

Fig. 6. ROC curves for the proposed method, a single finger, AND, OR, SUM, MAX, MIN rule[11] : (a) Using the proposed method vs. a single finger, AND, OR rule at the decision level (b) Using the proposed method vs. SUM, MAX, MIN rule at the matching score level

5 Conclusions

We have designed and implemented a fingerprint verification system which uses two fingerprint images. We show that considerable improvement in fingerprint verification performance can be achieved by selecting a better quality fingerprint image of two fingerprints. Also, we used fusion methods at the matching score level (SUM, MAX, MIN rule[11]). We proved theoretically and empirically that the proposed method is better than methods that use an AND rule or an OR rule and a single finger.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center(BERC) at Yonsei University.

References

1. A. Jain, L. Hong, and R. Bolle: On-line fingerprint verification. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, no. 4, pp.302-314, Apr. 1997.
2. H. C. Lee and R. E. Gaensslen: Advances in Fingerprint Technology. Elsevier, New York, 1991.
3. S. Prabhakar and A. K. Jain: Decision-level Fusion in Fingerprint Verification. Pattern Recognition, Vol. 35, no. 4, pp.861-874, 2002.
4. Lin Hong and Anil Jain: Integrating Faces and Fingerprint for Personal Identification. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, no. 12, pp.1295-1307, Dec. 1998.
5. E.S. Bigun, J. Bigun, B.Duc, and S. Fisher: Expert Conciliation for Multimodal Person Authentication Systems using Bayesian Statistics. in Proceedings of First International Conference on AVBPA, Crans-Montana, Switzerland, pp. 291-300, 1997.
6. J. Daugman: Biometric decision landscapes. Technical Report No. TR482, University of Cambridge Computer Laboratory, 2000.
7. S. Kim, D. Lee, and Jaihie Kim: Algorithm for Detection and Elimination of False Minutiae in Fingerprint Images. Lecture Notes in Computer Science on The Third Conference Audio- and Video-Based Biometric Person Authentication, Halmstad, Sweden, pp. 235-240, Jun. 2001.
8. D. Lee, K. Choi, and Jaihie Kim: A Robust Fingerprint Matching Algorithm Using Local Alignment. Proceedings. 16th International Conference on Pattern Recognition, Vol. 3, pp. 803-806, Aug. 2002.
9. J. Bigun and G. H. Granlund: Optimal Orientation Detection of Linear Symmetry. First International Conference on Computer Vision, IEEE Computer Society Press, Washington, DC, pp.433-438, June 1987.
10. K.V.Mardia and P.E.Jupp: Directional Statistics. John Wiley Sons Ltd, 2000.
11. M. Indovina, U. Uludag, R. Snelick, A. Mink and A. Jain: Multimodal Biometric Authentication Methods: A COTS Approach. Proc. MMUA 2003, Workshop on Multimodal User Authentication, Santa Barbara, CA, pp. 99-106, Dec. 2003.

A Fingerprint Authentication System Based on Mobile Phone*

Qi Su, Jie Tian**, Xinjian Chen, and Xin Yang

Center for Biometrics and Security Research,
Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation,
Chinese Academy of Sciences, Graduate School of the Chinese Academy of Sciences,
P.O. Box 2728, Beijing 100080, China
tian@doctor.com, jie.tian@mail.ia.ac.cn
http://www.fingerpass.net

Abstract. With the increasing volume of sensitive and private information stored in the mobile phone, the security issue of mobile phone becomes an important field to investigate. This paper proposes a fingerprint authentication system for mobile phone security application. A prototype of our system is developed from the platform of BIRD E868 mobile phone with external fingerprint capture module. It is composed of two parts. One is the front-end fingerprint capture sub-system, and the other is back-end fingerprint recognition system. A thermal sweep fingerprint sensor is used in the fingerprint capture sub-system to fit the limitations of size, cost, and power consumption. In the fingerprint recognition sub-system, an optimized algorithm is developed from the one participated in the FVC2004. The performance of the proposed system is evaluated on the database built by the thermal sweep fingerprint sensor.

1 Introduction

With the rapid evolution of mobile technology, mobile phone is not only a communication tool, but also a MMS center, a scheduler, a recorder, a camera, an mp3 player, and even a mobile web explorer. With the advancement of the hardware, mobile phones can store significant amount of sensitive and private information (e.g. address book, SMS, scheduler and even a bank account). Moreover, with the relative low cost, the number of mobile phone user increases rapidly in recent years. Worldwide mobile phone sales in 2003 was 520 million units, by the end of 2004 the estimated sales was in the range of 580 million units [1]. Nowadays, the mobile phone has become a necessary part of our daily life.

Currently, many mobile phones come with a four-digit Personal Identification Number (PIN) and a numerical entry key as a tool for user authentication. Because of the limited length, they may be susceptible to shoulder surfing or systematic trial-and-error attacks [2]. And the PIN may be difficult to remember and prone to input errors when entered via a touch screen.

* This paper is supported by the Project of National Science Fund for Distinguished Young Scholars of China under Grant No. 60225008, the Key Project of National Natural Science Foundation of China under Grant No. 60332010, the Project for Young Scientists' Fund of National Natural Science Foundation of China under Grant No.60303022, and the Project of Natural Science Foundation of Beijing under Grant No.4052026

** Corresponding author: Jie Tian; Telephone: 8610-62532105; Fax: 8610-62527995

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 151–159, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Because the size of the mobile phone becomes smaller and smaller, it can easily misplaced, unattended or stolen. As the information stored in a mobile device is sensitive, the effective protection of the mobile phone against unauthorized access has been increased. Biometrics recognition technology provides replacement or complement passwords to make a higher level of user convenience and security by means of fingerprint, hand geometry, iris, face, and signature etc. Among numerous biometrics technology, fingerprint authentication carries more advantages than others.

Fingerprint authentication has been thoroughly verified through various applications including law enforcement and commercial application for a long time. The fingerprint image may be taken and digitalized by relatively compact and cheap devices. Electronic fingerprint capture has been introduced with much success. Combining such methods with powerful microprocessors and pattern matching software has opened a new application in the mobile phone development.

This paper proposes an authentication system based on fingerprint recognition to improve the security protection of the mobile phone. The authentication system is composed of two parts. One is the front-end fingerprint capture sub-system and the other is back-end fingerprint recognition sub-system based on BIRD mobile phone E868. The fingerprint capture sub-system is an external module. It mainly consists of an ARM-Core processor LPC2106 and an Atmel thermal fingerprint sensor AT77C101B. It is responsible for capturing the fingerprint image frames, reconstructing the image and sending it to the recognition sub-system. As a part of the mobile phone operating system, the fingerprint recognition sub-system includes the enroll unit, match unit and system Application Program Interface (API). An optimized fingerprint recognition algorithm based on the one participated in the FVC2004 is used in the fingerprint recognition sub-system. The programs of both the fingerprint capture and recognition sub-system are optimized for the embedded environment.

This paper is organized as follows. Section 2 describes the structure of the fingerprint authentication system. Section 3 illustrates the software of authentication system. The fingerprint reconstruction algorithm, fingerprint recognition algorithm and the optimization techniques are described in this section. Section 4 shows the experimental results and section 5 concludes our work with future perspectives.

2 The Fingerprint Authentication System

The fingerprint authentication system is composed of two parts. One is the front-end fingerprint capture sub-system and the other is back-end fingerprint recognition sub-system. The structure of the whole system is shown in Fig. 1.

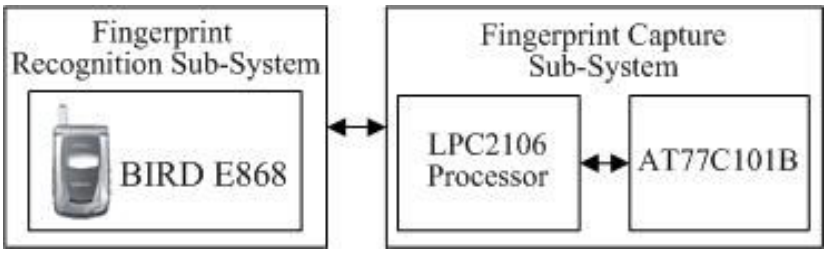


Fig. 1. The fingerprint authentication system block diagram

The fingerprint capture sub-system is an external module, and it is controlled by an ARM-Core processor LPC2106. The external module works in a slave mode. The LPC2106 processor receives the commands from the mobile phone via UART interface and controls the thermal fingerprint sensor AT77C101B to capture the image and reconstructs the original image frames to a full fingerprint image, and sends it to the mobile phone.

As a part of the mobile phone operating system, the recognition sub-system operates in the E868 mobile phone. The functions of the fingerprint recognition subsystem include fingerprint enroll and match. Moreover, the recognition sub-system provides a set of APIs. Other applications on the mobile phone operating system can possess the fingerprint authentication functions by calling those APIs.

2.1 E868 Mobile Phone

The hardware platform of the fingerprint authentication system includes the BIRD mobile phone E868 and the external module. Ningbo Bird Mobile Communications Co. Ltd. unveiled the E868 mobile phone in August 2003 [3]. The E868 mainly targets the high-end business market. It is capable of supporting up to 65,000-colors, touch-screen and handwriting recognition. The mobile phone provides PDA functions, e-mail, internet, camera, mp3, JAVA™ technology and more. The central processing unit of the E868 is a 16-bit embedded processor S1C33. The processor is produced by Epson Company and its working frequency is 13 MHz.

2.2 ARM Core Processor

The fingerprint capture sub-system is based on a LPC2106 ARM-Core embedded processor which is manufactured by PHILIPS [4]. The processor is very powerful. It has a 32-bit ARMTDMI-S core with real-time emulation and embedded trace support, and it can work at 60 MHz clock. Moreover, the LPC2106 processor incorporates 128 KB on-chip Flash and 64 KB on-chip Static RAM. Between the ARMTDMI-S core and the memory block, a 128-bit wide internal memory interface and unique accelerator architecture enable 32-bit code execution at the maximum clock rate.

Furthermore, the LPC2106 processor has a high efficient power management unit. The unit can put the whole processor into three statuses: normal, idle and power down. In addition, LPC2106 may turn off individual peripheral when it is not needed in application, resulting for power saving purpose.

Because of the limitation of the E868 hardware platform, the processor of the mobile phone can not be connected to the fingerprint sensor. So, as the co-processor, the LPC2106 is in charge of capturing the fingerprint image from the fingerprint sensor. After the process of the fingerprint image reconstruction, the LPC2106 sends the image to the mobile phone by using the UART interface. The LPC2106 processor is suitable for the mobile embedded application with the feature of small package (only 7×7 mm^2) and the low power consumption.

2.3 Fingerprint Sensor

The fingerprint authentication system uses the Atmel's AT77C101B FingerChip IC for taking fingerprint image [5]. It captures the image of a fingerprint as the finger

sweeping vertically over the sensor window. The AT77C101B sensor is composed of two main sections: thermal sensor and data conversion. The sensor section comprises a temperature-sensitive pixels array with 8 rows by 280 columns. The data conversion section mainly consists of analog signal amplifier and Analog-to-Digital Converter. The AT77C101B sensor can provide a resolution of 500 dpi fingerprint image. The pixel clock is programmable at up to 2 MHz, giving an output of 1780 frames per second.

The sensor has a very small size. The image zone is only 0.4×14 mm^2, it can be embedded into an external mobile module or even into the mobile phone. The AT77C101B's rectangular sensor window is much smaller than a square window fingerprint sensor with the same image resolution. So it leads to a decreased unit cost and further reduces the cost of the whole fingerprint authentication system. The sweep method of image capture means that the sensor window is self-cleaning with no latent prints left after an image capture. With the technical advance of the smaller size, lower cost and others, AT77C101B sensor is suitable for mobile hand-held devices, such as E868 mobile phone.

3 System Software Descriptions

The function of fingerprint authentication system includes fingerprint capture, enroll and match. The core algorithms are fingerprint reconstruction algorithm and verification algorithm. As the core part of the fingerprint authentication system, the capability of the algorithm influences the system directly. The reconstruction algorithm is based on the linear correlation theory. The verification algorithm is based upon the recognition algorithm participated in the FVC2004. Both of them are optimized prior being ported to E868 mobile phone.

3.1 Fingerprint Capture

The process of fingerprint acquisition is completed in the fingerprint capture subsystem. Because the continuous fingerprint frames are read by the thermal sweep sensor AT77C101B, the reconstruction algorithm is necessary to obtain a full fingerprint image. The fingerprint capture process has three main sections: 1) capturing more than 100 continuous fingerprint image frames; 2) reconstructing the fingerprint image; 3) sending the full fingerprint image to the mobile phone. The flow chart of the three sections is shown in Fig. 2.

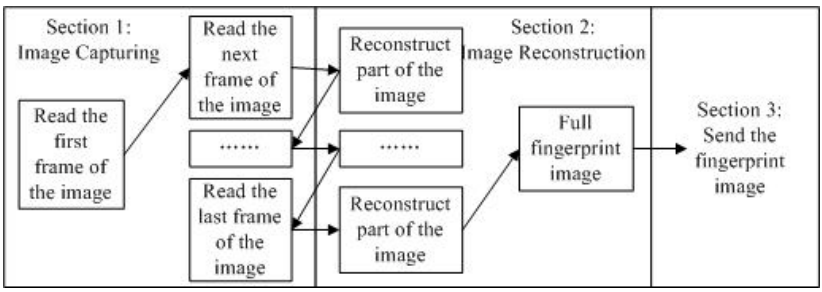


Fig. 2. The fingerprint capturing block diagram

When the fingerprint process starts, the thermal fingerprint sensor will read the image frames no matter a finger is on the sensor or not. One of the image frames' format is 8×280, 16 gray levels. The fingerprint capture sub-system uses a threshold to fix the start point of the swiping movement. First, the system applies local mean and variance to find the real fingerprint image frames. When the number of continuous image frames is larger than the threshold, the fingerprint capture process goes into the second part, the fingerprint reconstruction. Usually, as long as the threshold is not less than 3 frames, it can match the requirement of fixing the accurate starting point of the fingerprint movement.

The speed of the finger swept over the sensor window is not the same at each time. If the speed is slower than a reasonable rate, there will be an overlapping image in the two successive frames. In image reconstruction section, the program deletes the overlapping image and output the registration ones. The fingerprint reconstruction algorithm is based on the linear correlation theory, meanwhile adopts the virtue of the registration algorithm proposed by Hassan Foroosh etc. [6] Because of the practical movement of finger has been counted in fingerprint reconstruction algorithm, the search range of the translation between a pair of continuous swept fingerprint frames is limited from -45 degrees to +45 degrees. As a result, the computation time of the optimized method is only a quart of the normal method in which the search range is in 360 degrees. Moreover, the quality of fingerprint image after registration is the same as the normal method.

In fact, because of the limitation of the LPC2106's memory space, the part one and part two do not execute sequentially. The full fingerprint image is 256×280, 256 gray levels after resolution enhancement, and at least 32 frames are needed to reconstruct a full fingerprint image. Therefore the smallest memory space is 71680 bytes and it is larger than the LPC2106's integrated SRAM memory volume. The real image reconstruction process is shown in Fig. 2. The image capture and reconstruction are executed alternately. The LPC2106's SRAM is used as the temporary memory location for image frame data and other temporary data, and the FLASH is performed to save the program and the full fingerprint image.

In the third part, the LPC2106 sends the full fingerprint image to the mobile phone by UART interface at the rate of 230400 bps. Usually the total time of capturing a full fingerprint image is about 1.5 second.

3.2 Fingerprint Recognition Algorithm

The accuracy and the efficiency of fingerprint recognition algorithm directly influence the performance of the authentication system. Maio and Maltoni have presented the direct gray-scale minutiae detection algorithm [7] for fingerprints. However, in this paper, we present the method based on the fingerprint recognition algorithm participated in the FVC2004 [8]. And our algorithm is modified prior being ported to the E868 mobile phone.

The block diagram of fingerprint recognition sub-system is shown in Fig 3. It basically includes two parts: the enroll part and the match part. Each algorithm is composed of 4 stages. The first three processing stages are the same, reading a fingerprint image (256×280 pixels), applying an image filter with the frequency and extracting the minutiae from the fingerprints. The last stage of enroll is to save the fingerprint

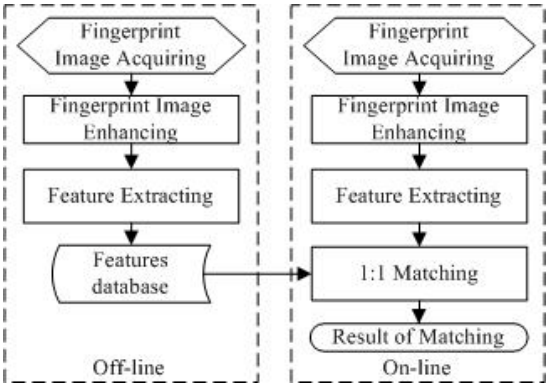
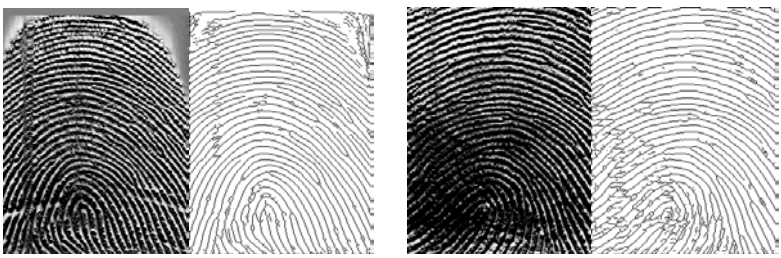


Fig. 3. Fingerprint verification system

template to the template file. While the last stage of the match part is to search the template file in order to find the similar template. Different algorithms of the fingerprint preprocessing, minutiae extraction and template matching are presented in detail as following.

In fingerprint preprocessing stage, a frequency transformation converts an image from its spatial-domain form of bright intensities into a frequency-domain form of frequency components. In this stage, fingerprint enhancement algorithm based on filtering in frequency domain [8] is used. First, Fourier transforming converts the fingerprints from spatial domain to frequency domain, then the fingerprints are enhanced by the proposed filter in frequency domain. The frequency domain shows the frequency of brightness variations, the direction of the variation patterns, and the amplitude of the waveforms representing the patterns.

After the enhanced fingerprint image is obtained, the next stage is to generate the thinned ridged fingerprint image and extract the minutiae of the thinned image. We compute the average grey value in every one of the 8 directions to decide the ridge direction of each pixel. To reduce the effect of noise, we use the algorithm proposed by Yuliang He [9] to get the thinned ridge fingerprint map. The thinned ridge map is shown in fig. 4. After the ridge map of filtered image is obtained, the algorithm proposed in [9] is performed to extract the minutiae.



(a) (b)

Fig. 4. Example of thinned fingerprint ridge images processed by our algorithm: (a) original image of forefinger and the thinned image; (b) original image of thumb and the thinned image

The fingerprint authentication system employs the thermal sweep fingerprint sensor to capture a fingerprint image. As it is well known, the deformations and nonlinear distortions of the sweep style sensor are more serious than those of an ordinary press style sensor. Even the fingerprint reconstruction algorithm removes the translation in the fingerprint capture stage, the other deformations are still big problems for fingerprint matching. We use the Maximum-likelihood estimation to calculate the optimal deformation parameters [10].

The main idea of the match algorithm is using an affine transformation model T to relate the fingerprint template and the minutiae set of the live-scanned fingerprint. The variables in the affine transformation model T represent some styles of deformation, including rotation, scale and so on. The deformations are formulated in terms of maximum-likelihood estimation, namely a probability density function. The experimental results presented in 4.2 demonstrate the good performance of the fingerprint authentication system by using the proposed algorithm.

3.3 Energy Management

Mobile phone is a kind of hand devices. They are characterized by small size, limited storage and processing power, and battery-powered. The battery-powered embedded computing system needs a set of efficient energy management to prolong the work time. Three approaches toward solving the task scheduling and voltage assignment problem are described in [11]. According to the characters of the LPC2106 processor, we propose a method that is suitable for the external module to prolong the system working time and make it accomplish the balance between system performance and battery duration.

The fingerprint capture sub-system works in the slave mode. It waits for the commands come from the application software running into mobile phone, and carries out the relevant operations. The major operations of sub-system are capturing a fingerprint image and sending it to the mobile phone. The sub-system does nothing for most situations. So we set the interrupt flag on for the LPC2106 processor's UART during the sub-system initialization. And afterwards, we make the LPC2106 processor and the sensor, AT77C101B, in sleep mode to save energy. When the mobile phone sends the command to captures a fingerprint image, the communication wakes up the LPC2106 processor to start processing. While the fingerprint capture is finished, the sub-system goes into the sleep mode again. In normal mode, the AT77C101B operates with a power consumption of 20 mW at 3.3V. When the sensor is in the sleep mode, it only consumes less than 10 μA current.

In addition, the LPC2106 processor has another good feature. It possesses a power management unit, which can turn off selected peripheral. When the processor is in the normal mode, the unused peripherals can be turned off automatically to save more energy. For example, when the sub-system is capturing a fingerprint image (described in 3.1), only part 3 needs to use the UART interface to communicate with the mobile phone. The UART interface can be turned off in the period of part 1 and part 2. In the same way, the AT77C101B sensor can be turned off in the period of part 3.

4 Experimental Results

We have developed a prototype of fingerprint recognition mobile phone based on the E868. The prototype has achieved the application of fingerprint enroll and match. The appearance of the fingerprint recognition mobile phone and the fingerprint image captured by the fingerprint recognition system are shown in Fig. 5.



(a) (b)

Fig. 5. (a) The appearance of the fingerprint recognition mobile phone; (b) Fingerprint image captured by the fingerprint recognition system

4.1 The Data Set

To test the performance of the prototype of fingerprint recognition mobile phone, we have built up a small fingerprint database in the mobile phone. 20 person works as volunteers for providing fingerprints. Thumb, forefinger and middle finger of both hands (six fingers total) of each volunteer were captured by the sweep fingerprint sensor. Four fingerprint images were taken for each of the six fingers from each person. The database totally includes 480 fingerprints.

4.2 Results for Fingerprint Match

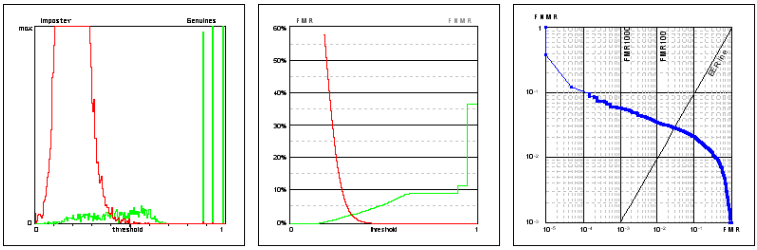
Four performance tests were measured: genuine match, imposter match, average match time and maximum template size. In genuine match, the number of genuine tests is 720. In imposter match, the total number of false acceptance tests is 7140. The definitions of equal error rate (EER), false non-match rate (FNMR), false match rate (FMR), and receiving operating curve (ROC) are defined in [7].

Fig. 6 presents the experimental results of the proposed algorithm on the database. Figure 6(a) shows the match score distributions. In figure 6(b), the value of EER of the proposed system is 4.13%. The value of FNMR equals to 5.64% for FMR = 1%. Figure 6(c) indicates the ROC curve of the proposed algorithm on the database.

We also measure the average match time of a fingerprint image and the maximum template size. Based on the results of both match methods, the average match time is about 6 seconds and the maximum template size is smaller than 128 bytes.

5 Conclusion

Security protection for mobile phone is a desperate issue nowadays. In this paper, we have presented the design and test of a mobile authentication system based on fingerprint recognition for security protection of the mobile phone.



(a) (b) (c)

Fig. 6. Experimental results of the proposed algorithm. (a) Score distributions (b) FMR(t) and FNMR(t) (c) ROC curve

The system consists of front-end fingerprint capture sub-system and back-end fingerprint recognition sub-system. The hardware platform is composed of E868 mobile phone and the external fingerprint capture module. The system software includes both fingerprint capture unit and recognition unit. In the recognition sub-system, the optimized fingerprint recognition algorithm is used. It is based on the algorithm participated in the FVC2004. The performance of the proposed system was evaluated on the 480 fingerprints database. The EER of the experiment is 4.13%.

The average match time will be decreased in the products of the fingerprint recognition mobile phone. Further works will be focused on the system performance optimization and the security implementations of the mobile phone based on the authentication system.

References

1. Alexander wolfe, Worldwide Mobile Phone Sales Surge, http://www.internetnews.com/wireless/article.php/3324061, March 10, 2004
2. Wayne A. Jansen, Authenticating Users on Handheld Devices, Proceedings of the Canadian Information Technology Security Symposium, May 2003. http://csrc.nist.gov/mobilesecurity/Publications
3. Ningbo Bird Mobile Communications Co. Ltd., BIRD DOEASY E868 Mobile Business Elite Introduce, http://doeasy.net.cn/index_2.htm
4. Philips Semiconductors Co. Ltd., LPC2106/2105/2104 USER MANUAL, http://www.semiconductors.philips.com
5. Atmel Corporation, AT77C101B FingerChip Datasheet, Rev. 2150B–BIOM–09/03, http://www.atmel.com
6. Hassan Foroosh, Josiane B. Zerubia, and Marc Berthod, Extension of Phase Correlation to Sub-pixel Registration, IEEE Trans. Image Processing, vol. 11, No.3,pp.188–200, Mar.2002
7. Dario Maio, Davide Maltoni, Raffaele Cappelli, and etc., FVC2004: Third Fingerprint Verification Competition, Proceedings of ICBA 2004, LNCS 3072, pp.1-7, 2004
8. Xinjian Chen, Jie Tian, Xin Yang, A Matching Algorithm Based on Local Topologic Structure, pp. Proceedings of ICIAR2004, LNCS 3211, pp. 360-367, 2004
9. Yuliang He，Jie Tian, Xiping Luo, and etc., Image Enhancement and Minutia Matching in Fingerprint Verification, Pattern Recognition Letters, Vol.24, pp.1349-1360, 2003
10. Yuliang He, Jie Tian, Qun Ren, and etc.，Maximum-Likehood Deformation Analysis of Different-Sized Fingerprints, Proceedings of AVBPA2003,LNCS 2688, pp.421-428, 2003
11. Daler Rakhmatov, Sarma Vrudhula, Energy Management for Battery-Powered Embedded Systems, ACM TECS, Vol. 2, No. 3, 2003

Fingerprint Quality Indices for Predicting Authentication Performance

Yi Chen[1], Sarat C. Dass[2], and Anil K. Jain[1]

[1] Department of Computer Science and Engineering
Michigan State University, East Lansing, MI, 48823
{chenyi1,jain}@cse.msu.edu
[2] Department of Statistics
Michigan State University, East Lansing, MI, 48823
sdass@stt.msu.edu

Abstract. The performance of an automatic fingerprint authentication system relies heavily on the quality of the captured fingerprint images. In this paper, two new quality indices for fingerprint images are developed. The first index measures the energy concentration in the frequency domain as a global feature. The second index measures the spatial coherence in local regions. We present a novel framework for evaluating and comparing quality indices in terms of their capability of predicting the system performance at three different stages, namely, image enhancement, feature extraction and matching. Experimental results on the IBM-HURSLEY and FVC2002 DB3 databases demonstrate that the global index is better than the local index in the enhancement stage (correlation of 0.70 vs. 0.50) and comparative in the feature extraction stage (correlation of 0.70 vs. 0.71). Both quality indices are effective in predicting the matching performance, and by applying a quality-based weighting scheme in the matching algorithm, the overall matching performance can be improved; a decrease of 1.94% in EER is observed on the FVC2002 DB3 database.

1 Introduction

Fingerprint images are usually obtained under different conditions of the skin of a finger (e.g., dry, wet, creased/wrinkled, or abraded), the ergonomics of the acquisition system (e.g., ease of use, alignment and positioning), and the inherent limitations of the sensing equipment (e.g., shadow from optical sensors and electrical noise from capacitive sensors). These conditions, in turn, affect the quality of the acquired fingerprint images (see Figures 1 (a-c)). Fingerprint quality is usually defined as a measure of the clarity of the ridge and valley structures, as well as the "extractability" of features (such as minutiae and singularity points). Poor quality fingerprint images often result in spurious and missed features, and thus severely degrade the performance of an authentication system by increasing the false reject and false accept rates. Recently, NIST [1] has shown that the performance of a fingerprint authentication system is mostly affected, among other factors, by fingerprint image quality. Therefore, it is desirable to assess the quality of a fingerprint image to improve the overall performance of a fingerprint authentication system.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 160–170, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Many on-going and past efforts have tried to address the problem of assessing fingerprint image quality. Bolle et al. [2] used ratio of directional area to nondirectional area as a quality measure. Hong et al. [3] and Shen et al. [4] applied Gabor filters to identify blocks with clear ridge and valley patterns as good quality blocks. Ratha and Bolle [5] computed the ratio of energy distribution in two subjectively selected frequency bands based on the WSQ (Wavelet Scalar Quantization) compressed fingerprint images. Lim et al. [6] combined local and global spatial features to detect low quality and invalid fingerprint images. The most recent work by Tabassi et al. [1] presented a novel definition of fingerprint quality as a predictor for matching performance. They consider quality assessment as a classification problem and use the quality of extracted features to estimate the quality label of a fingerprint image. This approach is effective only when the feature extraction algorithm is reliable and is computationally efficient.

In this paper, we propose two new fingerprint quality indices. The first index measures the entropy of the energy distribution in the frequency domain. The second estimates the local coherence of gradients in non-overlapping blocks. We propose a framework for evaluating and comparing quality indices by assessing how well they predict the system performance at three processing stages: (i) image enhancement, (ii) feature extraction and (iii) matching. Our goal is to determine how each processing stage will be affected by the image quality, and to compare the two quality indices in terms of their predictive capabilities. We also adopt a quality-based weighting scheme to improve the matching performance. To the best of our knowledge, this systematic framework is novel.

The rest of the paper is organized as follows. Section 2 describes in detail the algorithms of each proposed quality index. Section 3 introduces the new framework for evaluating fingerprint quality indices. In Section 4, experimental results are provided and discussed. Summary and future work are included in Section 5.

2 Fingerprint Quality Indices

2.1 A Quality Index in the Frequency Domain

Given a digital image of size $M \times N$, the two-dimensional Discrete Fourier Transformation (DFT) evaluated at the spatial frequency $(\frac{2\pi k}{M}, \frac{2\pi l}{N})$ is given by

$$F(k,l) = \frac{1}{NM} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} f(i,j) e^{-\iota 2\pi (\frac{ki}{N} + \frac{lj}{M})}, \ \iota = \sqrt{-1}\,, \tag{1}$$

where $f(i,j)$ refers to the gray level intensity at pixel (i,j) of the image. Although DFT produces a complex-valued output, only the power spectrum $P(k,l) \equiv |F(k,l)|^2$ is often used as it contains most of the information regarding the geometric structure of an image.

The ridge frequency in a fingerprint image is generally around 60 cycles per image width/height [8]. Since the image width/height is usually between 120 and

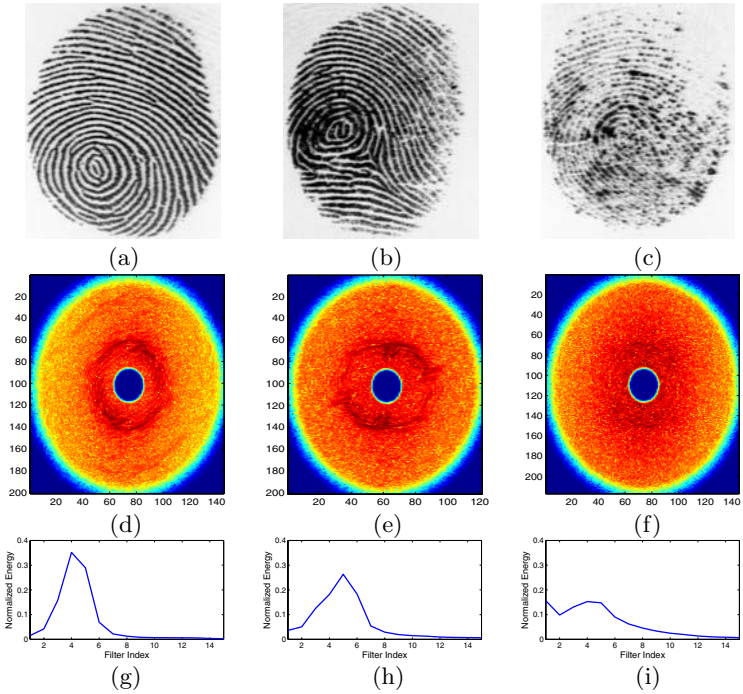


Fig. 1. Computing the quality index Q_f from the power spectrum: Panels (a-c) show three fingerprint images in the decreasing order of quality; Panels (d-f) show their corresponding power spectrums; Panels (g-i) show the energy concentrations in the region of interest. The values of Q_f for the three images are 1.0, 0.6, and 0.3, respectively

1000 pixels, the dominant ridge frequencies should be between $60/1000 = 0.06$ and $60/120 = 0.5$. Therefore, the region of interest (ROI) of the power spectrum is defined to be an annular band with radius ranging from 0.06 to 0.5. Figures 1(a-c) show three fingerprint images of varying quality with their corresponding power spectrums in the ROI shown in Figures 1(d-f). Note that, the fingerprint image with good quality (Figure 1(a)) presents strong ring patterns in the power spectrum (Figure 1(d)), while a poor quality fingerprint (Figure 1(c)) presents a more diffused power spectrum (Figure 1(f)). The global quality index will be defined in terms of the energy concentration in this ROI.

We use a family of Butterworth low-pass filters to extract the ring features from the ROI. A Butterworth function [7], indexed by m and n, is defined as

$$H(k,l\,|\,m,n) = \frac{1}{1+\frac{1}{m^{2n}}((\frac{k-a}{M})^2+(\frac{l-b}{N})^2)^n}\ , \tag{2}$$

where (k,l) is the pixel index in the power spectrum corresponding to the spatial frequency $(\frac{2\pi k}{M}, \frac{2\pi l}{N})$ and (a,b) is the location of the center of the power spectrum corresponding to spatial frequency (0,0). The Butterworth function generates a low-pass filter with the cutoff frequency given by m and the filter order given by

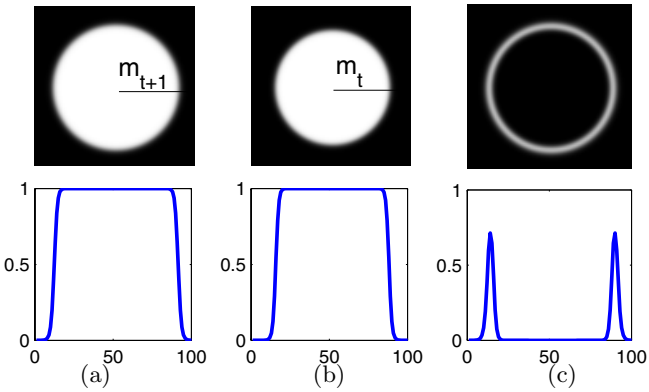


Fig. 2. Taking the differences of two consecutive low-pass filters (a) $H(k,l\,|m_{t+1},\,n)$ and (b) $H(k,l\,|m_t,\,n)$ $(n=20)$ to obtain a bandpass filter (c) $R_t(k,l)$

n. The value of n controls the steepness of the drop at the cutoff frequency; the larger the value of n, the closer H is to an idealized step function.

We construct a total of T equally spaced bandpass filters, R_t, by taking differences of two consecutive Butterworth functions, that is,

$$R_t(k,l) = H(k,l\,|\,m_{t+1},n) - H(k,l\,|\,m_t,n)\ , \tag{3}$$

where $m_t = 0.06 + t\frac{0.5-0.06}{T}$ and $t = 0,1,2,\ldots,(T-1)$. The construction of $R_t(k,l)$ from $H(k,l\,|\,m_{t+1},\,n)$ and $H(k,l\,|\,m_t,\,n)$ is shown graphically in Figure 2. For every t, R_t captures the energy in an annular band with frequencies from m_t to m_{t+1}. The energy concentrated in the t-th band is computed by

$$E_t = \sum_{k=0}^{N-1}\sum_{l=0}^{M-1} R_t(k,l)P(k,l)\ , \tag{4}$$

and the normalized energy for the t-th bandpass filter is defined as $P_t = \frac{E_t}{\sum_{t=0}^{T-1} E_t}$. In Figures 1(g-i), we plot the distribution of P_t for $T = 15$ bandpass filters. A good quality image has a more peaked energy distribution while poor ones have more diffused distribution. The extent of energy concentration is given by the entropy

$$E = -\sum_{t=0}^{T-1} P_t \log P_t\ , \tag{5}$$

which achieves the maximum value $\log T$ when the distribution is uniform and decreases when the distribution is peaked. Our quality score is defined as

$$Q_f = \log T - E\ , \tag{6}$$

so that a fingerprint image with good (bad) quality will have a higher (lower) value of Q_f. We have normalized Q_f on the database so that the values lie between 0 and 1.

2.2 A Quality Index in the Spatial Domain

To assess fingerprint image quality in a local region, we partition a given image into a lattice of blocks of size $b \times b$. An algorithm to distinguish the fingerprint foreground from the background is then applied as described in [2]. For each foreground block B, let $g_s = (g_s^x, g_s^y)$ denote the gradient of the gray level intensity at site $s \in B$. The covariance matrix of the gradient vectors for all b^2 sites in this block is given by

$$J = \frac{1}{b^2} \sum_{s \in B} g_s g_s^T \equiv \begin{bmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{bmatrix} . \tag{7}$$

The above symmetric matrix is positive semidefinite with eigenvalues

$$\begin{aligned} \lambda_1 &= \frac{1}{2}(trace(J) + \sqrt{trace^2(J) - 4\,det(J)}) \\ \lambda_2 &= \frac{1}{2}(trace(J) - \sqrt{trace^2(J) - 4\,det(J)}) \ , \end{aligned} \tag{8}$$

where $trace(J) = j_{11} + j_{12}, det(J) = j_{11}j_{22} - j_{12}^2$ and $\lambda_1 \geq \lambda_2$. The *normalized coherence measure* is defined as

$$\tilde{k} = \frac{(\lambda_1 - \lambda_2)^2}{(\lambda_1 + \lambda_2)^2} = \frac{(j_{11} - j_{22})^2 + 4j_{12}^2}{(j_{11} + j_{22})^2} \ , \tag{9}$$

with $0 \leq \tilde{k} \leq 1$. This measure reflects the clarity of the local ridge-valley orientation in each foreground block B. If the local region has a distinct ridge-valley orientation, then $\lambda_1 \gg \lambda_2$ results in $\tilde{k} \approx 1$. On the contrary, if the local region is of poor quality, we obtain $\lambda_1 \approx \lambda_2$ and consequently $\tilde{k} \approx 0$.

A single quality score can be computed as the weighted average of the block-wise coherence measures given by

$$Q_s = \frac{1}{r} \sum_{i=1}^{r} w_i \tilde{k_i} \ , \tag{10}$$

where r is the total number of foreground blocks, and the relative weight ω_i for the i-th block centered at $l_i = (x_i, y_i)$ is determined by

$$\omega_i = exp\{- \parallel l_i - l_c \parallel^2 /(2q)\} \ , \tag{11}$$

where l_c is the centroid of foreground fingerprint, and q is a normalization constant, which reflects the contribution for blocks with respect to the distance from the centroid [5]. Generally, regions near the centroid of a fingerprint receive higher weights, since they are likely to provide more information than the peripheral.

Figure 3 shows the local quality maps of the three fingerprint images and their overall quality indices. We have also normalized Q_s on the database so that the values lie between 0 and 1.

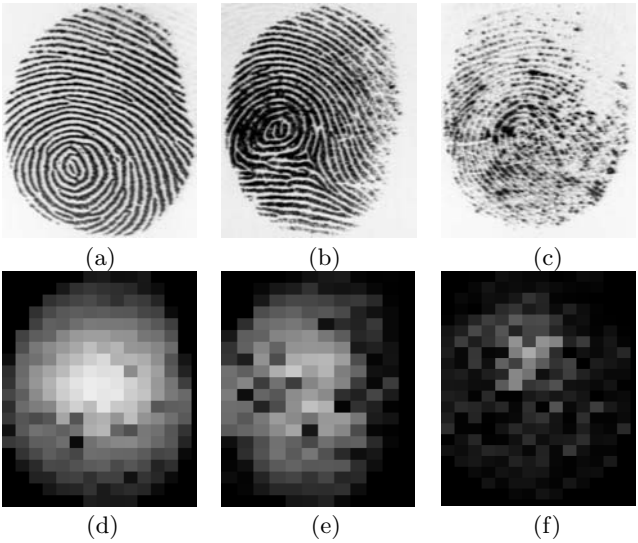


Fig. 3. Computing the quality index Q_s using the spatial coherence measure. Panels (a-c) are the fingerprint images. Panels (d-f) are the block-wise values of $\tilde{k}$; blocks with brighter color indicate higher quality in the region. The values of Q_s for the three fingerprint images are 0.95, 0.56, and 0.20, respectively

3 Evaluation Criteria

In this section, an evaluation criteria is developed for assessing the performance of image enhancement, feature extraction and matching with respect to the proposed quality indices.

3.1 Predicting the Image Enhancement Performance

Our goal is to first quantify the robustness of enhancement for varying values of Q_f and Q_s. A fingerprint image with high values of Q_f and Q_s should be less sensitive (or more robust) to the tuning parameters of an enhancement algorithm than those with low Q_f and Q_s values. The following method is developed to quantify this sensitivity with regard to the tuning of an enhancement algorithm.

Given an enhancement algorithm E, we tune the parameters to obtain a modified version called E'. Run E and E' separately on a fingerprint image to generate two enhanced images I and I'. Let $A = (g^1, g^2, \ldots, g^u)$ and $B = (h^1, h^2, \ldots, h^v)$ be the sets of minutiae detected, respectively, from I and I'. Compute p as the number of paired minutiae in A and B: minutiae $g^i\,(i = 1, \ldots, u)$ and $h^j\,(j = 1, \ldots, v)$ are said to be paired if their distances in position and orientation are within a tolerance bound of 8 pixels and 30 degrees, respectively. The robustness index (RI) of a fingerprint image is given by

$$RI = \frac{p}{u + v - p}\ , \tag{12}$$

where $(u + v - p)$ represents the total number of minutiae detected in both enhanced images. A low RI value indicates large variance in the number of minutiae detected and hence poor image quality due to its sensitivity to the turning of parameters. On the contrary, high RI value indicates consistency in minutiae extraction and consequently, good image quality due to its robustness to parameter tuning (Figure 4).

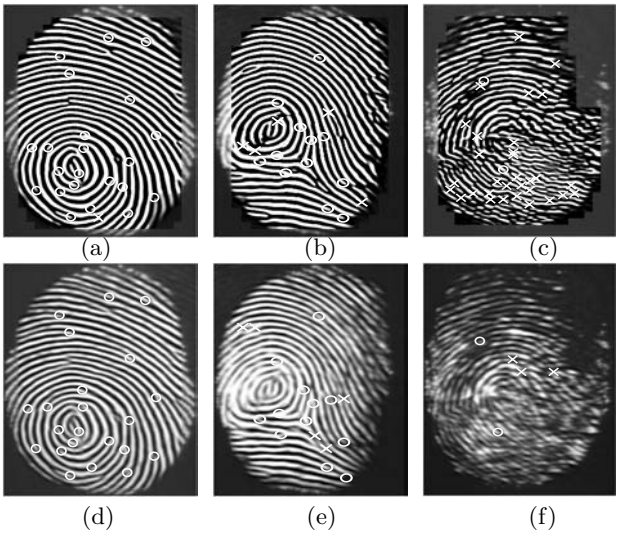


Fig. 4. Sensitivity to the tuning parameters of an enhancement algorithm. Input images are those shown in Figures 3(a-c). Panels (a-c) are obtained using E while panels (d-f) are obtained using E'. Minutiae consistently extracted from both algorithms are considered robust ($\circ$), whereas minutiae detected only by E or E' are non-robust ($\times$)

3.2 Predicting the Feature Extraction Performance

The effects of image quality with regard to feature extraction performance can be measured using the goodness index (GI) defined as

$$GI = \frac{p}{t} - \frac{a+b}{u} \ , \tag{13}$$

where p, a, b, respectively, represent the total number of paired, missed and spurious minutiae among the u detected minutiae when compared to the number of ground truth minutiae t in the given fingerprint image. Here, *missed minutiae* refers to a ground truth minutiae that is missed by the feature extraction whereas *spurious minutiae* represents an extracted minutiae that is not matched with a ground truth minutiae. A low GI value is obtained when the number of missed or spurious minutiae is much larger than the paired minutiae, indicating poor image quality. A high GI value, on the contrary, indicates good quality as most minutiae are correctly matched.

3.3 Predicting the Matching Performance

When matching scores are available, a *Receiver Operating Characteristic* (ROC) curve is plotted to reflect the performance of the matching algorithm. One effective evaluation criterion for a quality index is to rank the ROC as a function of image quality. More specifically, we can divide the quality scores into r equally numbered bins (from low to high) and plot r ROC curves, with the i-th curve reflecting the matching performance after images in the first $i, 0 \leq i \leq 4$ bins are pruned. The 0-th bin is by convention, the original database with no images removed. If a quality index is a good predictor of the matching performance, the ROC curves should consistently rise as more poor quality images are excluded.

3.4 Incorporating Local Quality into the Matching Algorithm

We propose to incorporate the local coherence measure, $\tilde{k_i}$, into the fingerprint matching algorithm [12] that accounts for the reliability of the extracted minutiae points. Prior to finding the matching score between a pair of fingerprint images, we need to align them to remove effects of any translation and rotation of the finger. This is done by maxmizing

$$W = \sum_{i=1}^{p} \sqrt{\tilde{k}^A_{f(i)} \times \tilde{k}^B_{g(i)}} \ , \tag{14}$$

where p is the total number of paired minutiae between A and B, $\tilde{k}^A_{f(i)}$ and $\tilde{k}^B_{g(i)}$ are the local coherence measures associated with the i-th paired minutiae in A and B, respectively; functions f and g return index of the block that contains the paired minutiae belonging to A and B, respectively. Once W is maximized, its corresponding transformation parameters are applied to align the orientation field of the pair, with both results determining the final matching score. Therefore, if the quality is high for both minutiae in a pair, this pairing will contribute more to the estimation of transformation parameters as well as the matching score than a pairing of low quality minutiae.

4 Experimental Results

The quality indices are tested using two databases, namely the IBM-HURSLEY database and FVC2002 DB3 [11]. The IBM-HURSLEY database contains multiple impressions of 269 fingers (a total of 900 images) taken at significantly different times, resulting in large variability in fingerprint quality. The images have different sizes but the same resolution of 500 *dpi*, with "true" minutiae marked by a human expert. The FVC2002 DB3 contains 800 images from 100 fingers (8 impressions per finger), all with the same size (300×300) and the same resolution (500 *dpi*). This database is the most difficult among the four databases in FVC2002 in terms of image quality [11]. No ground-truth is provided for this database, and hence, the quality indices for this database are tested

at the matching stage, while the quality indices for IBM-HURSLEY are tested at the enhancement and the feature extraction stages.

To evaluate the enhancement performance with regard to the proposed quality indices, we employed the enhancement algorithm proposed in [9], and the minutiae feature extraction algorithm given in [10]. We apply a new combination of three tuning parameters of the enhancement algorithm, namely, the minimum inter-ridge distance, the window width and height. The default combination was $E = [12, 11, 11]$ and the new one is $E' = [7, 11, 7]$. The RI value for each fingerprint is obtained as in equation (12) and the quality indices Q_f and Q_s are obtained as in Section 2. Figures 5(a-b) show the scatter plots of RI versus Q_f and RI versus Q_s on the IBM-HURSLEY database. We also sort the images in the increasing order of quality and divide them into $r = 5$ bins (180 images per bin). The median and the lower and upper quartiles of the quality indices in each bin are calculated and shown in the box plots in Figures 5(c-d). It is demonstrated that Q_f has a stronger predictive capability for RI, as it acquires a larger Pearson's correlation (0.70) than Q_s (0.50). In a similar manner, the performance of the feature extraction algorithm (measured by GI in equation (13)) is evaluated with respect to Q_f and Q_s. Here, default settings of the enhancement algorithm is used. Figure 6 gives the corresponding results. Both Q_f and Q_s are effective in predicting the feature extraction performance with Q_s achieving a slightly higher correlation than Q_f (0.71 vs. 0.70).

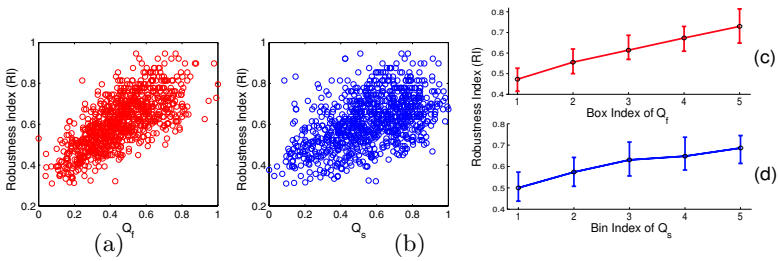


Fig. 5. The effect of the proposed quality indices on image enhancement. (a) gives the scatter plot and (c) the box plot of RI versus Q_f, and (b) gives the scatter plot and (d) the box plot of RI versus Q_s

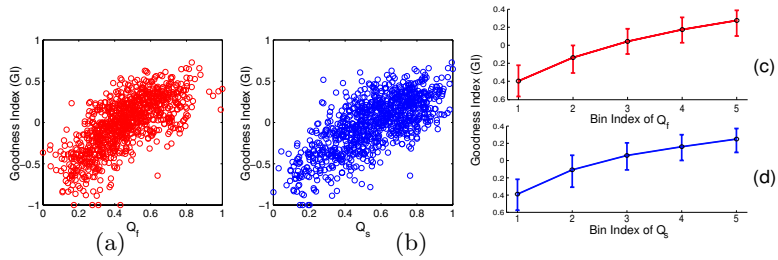


Fig. 6. The effect of the proposed quality indices on feature extraction. (a) gives the scatter plot and (c) the box plot of GI versus Q_f, and (b) gives the scatter plot and (d) the box plot of GI versus Q_s

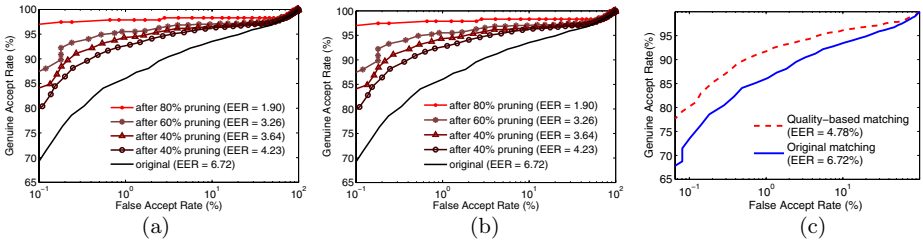


Fig. 7. Improving the matching performance by (a) pruning poor quality images with regard to Q_f and (b) Q_s, and (c) adopting a quality-based weighting scheme in the matcher

Finally, a matcher proposed in [12] is adopted for evaluating and improving the matching performance with respect to the quality indices on FVC2002 DB3. Five ROC curves are plotted in Figures 7(a-b) as suggested in Section 3.3. Figure 7(c) shows the overall improvement in the matching performance when local coherence measures are incorporated by a quality-based weighting scheme in the matcher (see Section 3.4).

5 Conclusion and Future Work

This paper proposes two quality indices, global (Q_f) and local (Q_s), for fingerprint images. We compare the two in a generic evaluation framework and observe the following: (1) Q_f has better predictive capabilities at the image enhancement stage than Q_s. This is because the image enhancement algorithm we use is based on Gabor filtering in the frequency domain, and is therefore directly related to Q_f. (2) Q_s is slightly more effective than Q_f at the feature extraction stage. This is because feature extraction concentrates on local details which is measured directly by Q_s. (3) Both Q_f and Q_s are effective in predicting and improving the matching performance. Future work includes expanding the current framework to other possible representation of fingerprints and biometric identifiers.

References

1. Tabassi, E., Wilson, C., Watson, C.: Fingerprint Image Quality. NIST research report NISTIR7151 (August, 2004).
2. Bolle, R. et al.: System and methods for determing the quality of fingerprint images. United Sates patent number US596356 (1999).
3. Hong, L., Wan, Y., Jain, A.: Fingerprint image enhancement: algorithms and performance evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **20** (1998) 777–789.
4. Shen, L., Kot, A., Koo, W.: Quality measures of fingerprint images. Audio- and Video-based Biometric Person Authentication, (2001).
5. Ratha, N., Bolle, R.: Fingerprint image quality estimation. IBM computer science research report RC21622 (1999).

6. Lim, E., Jiang, X., Yau, W.: Fingerprint quality and validity analysis. IEEE International Conference on Image Processing, **1** (2002) 469-472.
7. Rosenfeld, A., Kak, A.: Digital Picture Processing. Academic Press, (1982).
8. Hong, L., Jain, A., Pankanti, S., Bolle, R.: Fingerprint Enhancement. IEEE Workshop on Applications of Computer Vision, (1996) 202-207.
9. Hong, L., Wan. Y, Jain, A.: Fingerprint Image Enhancement: Algorithms and Performance Evaluation. IEEE Transactions on PAMI, **20-8** (1998) 777-789.
10. Jain, A., Hong, L., Bolle, R.: On-line fingerprint verification. IEEE Trasactions on Pattern Analysis and Machine Intelligence, **19-4** (1997) 302-314.
11. Maltoni, D., Cappelli, R., Wayman, J., Jain, A.: FVC2002: Second Fingerprint Verification Competition. International Conference on Pattern Recognition, **3** (2002) 811-814.
12. Jain, A., Prabhakar, S., Chen, S.: Combining Multiple Matchers for a High Security Fingerprint Verification System. Pattern Recognition Letters, **20** (1999) 1371-1379.

Registration of Fingerprints by Complex Filtering and by 1D Projections of Orientation Images

Kenneth Nilsson and Josef Bigun

School of Information Science, Computer and Electrical Engineering (IDE)
Halmstad University, P.O. Box 823, SE-301 18 Halmstad, Sweden
{Kenneth.Nilsson,Josef.Bigun}@ide.hh.se

Abstract. When selecting a registration method for fingerprints, the choice is often between a minutiae based or an orientation field based registration method. In selecting a combination of both methods, instead of selecting one of the methods, we obtain a *one modality multi-expert* registration system. If the combined methods are based on different features in the fingerprint, e.g. the minutiae points respective the orientation field, they are uncorrelated and a higher registration performance can be expected compared to when only one of the methods are used. In this paper two registration methods are discussed that do not use minutiae points, and are therefore candidates to be combined with a minutiae based registration method to build a multi-expert registration system for fingerprints with expected high registration performance. Both methods use complex orientations fields but produce uncorrelated results by construction. One method uses the position and geometric orientation of symmetry points, i.e. the *singular points* (SPs) in the fingerprint to estimate the translation respectively the rotation parameter in the Euclidean transformation. The second method uses 1D projections of *orientation images* to find the transformation parameters. Experimental results are reported.

1 Introduction

There are numerous techniques that use minutiae points in Automatic Fingerprint Identification Systems (AFIS) as well as low cost silicon sensor systems that are geared toward minutiae based techniques. This is due to long history of minutiae used in crime scene investigations. Consumer uses of biometrics increasingly questions the limitation of identification features to minutiae. Even more interestingly, by selecting a combination of features, instead of selecting minutiae, we can obtain a *one modality multi-expert* registration system. The two registration methods can be expected to be uncorrelated if they are based on different features in the fingerprint, e.g. the minutiae pattern respective the orientation field. By combining the output of uncorrelated methods a gain in the registration performance can be achieved, compared to the use of only one of

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 171–183, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the methods. This is because the methods complement each other in a positive way. When one method fails the other may still have success in the registration.

In the minutiae based registration methods the fingerprints are represented by its minutiae points, i.e. the position and the orientation of their minutiae are elements in their respective feature vector representation. Aligning the two fingerprints is to find the transformation parameters that maximize the number of matching minutiae pairs in the feature vectors [1, 2]. If the transformation is the Euclidean transformation, the parameters are the rotation angle and the translation vector [3] relating the template and the test fingerprint.

However in low quality fingerprints it is difficult to automatically extract the minutia points in a robust way. This often means that genuine minutiae are missed and that false minutiae are added [2]. Also, in cost sensitive applications, because the price of the sensor depends on the sensor area, sensors with small areas are used and therefore fewer numbers of minutiae are present in the captured fingerprint. For these two situations a high performance registration is difficult to obtain if only the minutiae based registration method is used. A higher performance can be expected if the minutiae based method can be combined with an other technique which we suggest to be orientation field features.

In this paper two registration methods are suggested that use the global structure of the fingerprint, and therefore are more robust to low quality fingerprint registration and more suitable to register fingerprints captured from small area sensors. They are therefore candidates to be combined with a minutiae based registration method to build a multi-expert registration system for fingerprints as discussed above. One method uses the position and geometric orientation of symmetry points, i.e. the singular points (SPs) in the fingerprint (see Figure 1) to estimate the translation respectively the rotation parameter in the Euclidean transformation [4]. The second method uses 1D projections of orientation images [5] to find the transformation parameters intended for a situation when SPs are poorly imaged. Both methods complement each other as well as minutiae and used complex orientation fields (see Figure 1).

2 Registration by Symmetry Points

This method (called method 1) extracts automatically the position and the geometric orientation of SPs, from the global structure using complex filters designed to detect rotational symmetries. The translation is estimated from the difference in position, and the rotation parameter from the difference in the geometric orientation of SPs in the test and the template fingerprint. In [4] we have shown that an unbiased alignment error with a standard deviation of approximately the size of the average wavelength (13 pixels) of a fingerprint is possible to achieve using this method.

A common technique to extract SPs (core and delta points) in fingerprints is to use the *Poincaré* index introduced by Kawagoe and Tojo [6]. It takes the values 180°, −180°, and 0° for a core point, a delta point, and an ordinary point respectively. It is obtained by summing the change in orientation following a

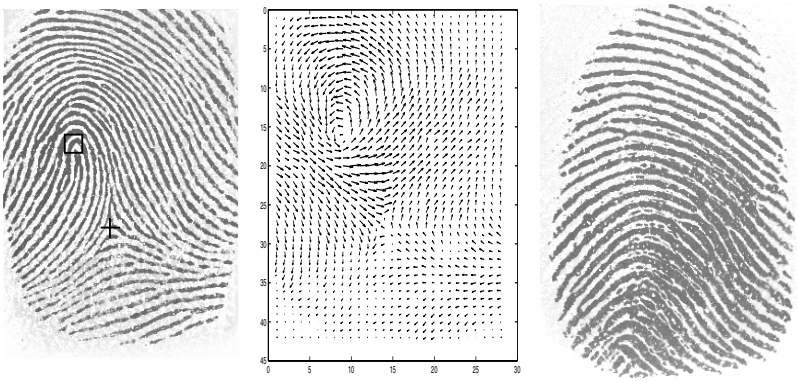


Fig. 1. Left: marked singular points, a core point is marked with a square and a delta point with a cross. Middle: the estimated complex orientation field at level 3 for the fingerprint to the left. Right: a fingerprint of type arch

closed curve counterclockwise around a point [7]. This technique has been used in the studies of Karu and Jain [7], and Bazen and Gerez [8] to define and extract SPs.

Our method using complex filters compared to *Poincaré* index to identify SPs has the advantage to extract not only the position of an SP but also its spatial orientation. When two fingerprints are rotated and translated relative to each other our method can estimate both translation and rotation parameters simultaneously. In the work of Bazen and Gerez [8] the position extraction and the orientation estimation of an SP is done in two sequential steps. The position extraction is performed by using *Poincaré* index. The orientation estimation is done by matching a reference model of the orientation field around an SP with the orientation map of the extracted SP. The orientation maps were obtained by using a technique introduced in [9].

2.1 Filters for Rotational Symmetry Detection

Complex filters, of order m, for the detection of patterns with rotational symmetries are modeled by $e^{im\varphi}$ [10, 11]. A polynomial approximation of these filters in gaussian windows yields $(x+iy)^m g(x,y)$ where g is a gaussian defined as $g(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}$ [12, 13].

It is worth to note that these filters are not applied to the original fingerprint image but instead they are applied to the complex valued orientation field image $z(x,y) = (f_x + if_y)^2$. Here f_x is the derivative of the original image in the x-direction and f_y is the derivative in the y-direction.

In our experiments we use filters of first order symmetry or parabolic symmetry i.e.
$h_1(x,y) = (x+iy)g(x,y) = re^{i\varphi}g(x,y)$ and
$h_2(x,y) = (x-iy)g(x,y) = r\ e^{-i\varphi}g(x,y) = h_1^*$.
Patterns that have a local orientation description of $z = e^{i\varphi}$ (m=1) and $z = e^{-i\varphi}$

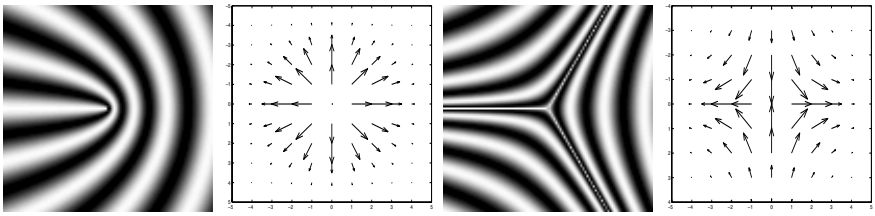


Fig. 2. Patterns with a local orientation description of $z = e^{i\varphi}$ (left) and $z = e^{-i\varphi}$ (right). Both in gray scale (patterns) and in z representation (complex filters)

(m=-1) are shown in Figure 2. As can be seen these patterns are similar to patterns of a core respectively a delta point in a fingerprint and therefore suitable to use as SP-extractors. The SP-extractors are the z representation of the patterns, i.e. the complex filter h_1 and h_2 respectively.

The complex filter response is $c = \mu\ e^{i\alpha}$, where $\mu = \frac{|I_{20}|}{I_{11}}$ is a certainty measure of symmetry, and $\alpha = Arg(I_{20})$ is the "member" of that symmetry family, here representing the geometric orientation of the symmetric pattern. The scalars $I_{20} = \langle h_1, z \rangle$ for the core point extraction, $I_{20} = \langle h_2, z \rangle$ for the delta point extraction, and $I_{11} = \langle |h_1|, |z| \rangle$ are obtained by use of the 2D complex scalar product symbolized by $\langle\ \rangle$ [12]. Representing the certainty measures by μ_1 and μ_2 for core point respectively delta point symmetry, we can identify an SP of type core if $\mu_1 > T_1$ and of type delta if $\mu_2 > T_2$, where T_1 and T_2 are empirically determined thresholds.

2.2 Multi-scale Filtering

Using a multi-resolution representation of the complex orientation field offers a possibility to extract SPs more robustly and precisely compared to a representation at only one resolution level. The extraction of an SP starts at the lowest resolution level (a smooth orientation field) and continues with refinement at higher resolutions. The result at a low resolution guides the extraction at higher resolution levels. This strategy can be taken because SPs have a global support from the orientation field [14].

The complex orientation field $z(x, y)$ is represented by a five level Gaussian pyramid. Level 4 has the lowest, and level 0 has the highest resolution. The core and the delta filtering is applied on each resolution. The complex filter response is called c_{nk}, where k=4, 3, 2, 1 and 0 are the resolution levels, and n=1, 2 are the filter types (core and delta).

Figure 3 (upper row) shows the magnitude of the filter responses of filter h_1 (called μ_{1k}). The filter is applied to the orientation field of the fingerprint to the left in Figure 1.

2.3 Multi-scale Searching for SPs

The extraction of an SP starts at the lower resolution level, i.e. we search for maximum in the certainty image μ_1 and μ_2 for a core and a delta point respec-

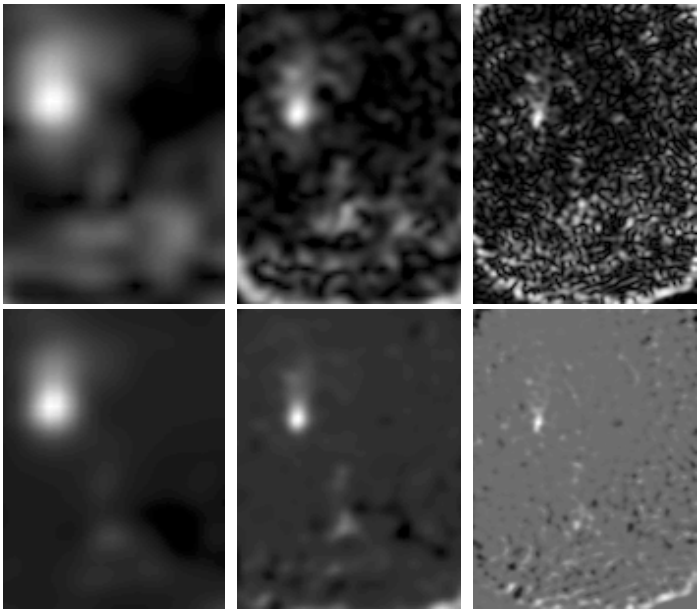


Fig. 3. Filter responses for the fingerprint to the left in Figure 1, core point. Row1: μ_{1k}, k=3, 2, and 1. Row2: enhanced μ_{1k}^{enh}, k=3, 2, and 1

tively. In the found position of maximum $(x, y)_{n4}^{max}$ we extract the complex filter response c_{n4}^{max}, for each type of SP, which is a vector pointing in the geometric orientation of respective SP. The magnitude of this vector is put to one, we call this vector SP_{or} of which there is one per resolution level and SP type. The SP_{or} is then used to define the *search window* (for a core point only) and to increase the signal-to-noise ratio in the certainty images μ_1 and μ_2 when searching for maximum at the next higher resolution level. More precisely, the enhanced certainty image μ_{k-1}^{enh} at level $(k-1)$ is obtained according to equation 1, where φ is the difference in angle between a filter response vector c_{k-1} and the SP_{or} vector at previous lower resolution level k.

$$\mu_{k-1}^{enh} = \mu_{k-1} \cdot cos(\varphi) \tag{1}$$

The quantity μ_{k-1} represents the certainty as described in section 2.1 and the above equation is a vectorial projection of c_{k-1} on SP_{or}. In this way we lower the responses of those complex filter responses that are not coherent with the orientation of the SP_{or} at the previous lower resolution level. Figure 3 (lower row) shows the enhanced certainty image for a core point for the fingerprint to the left in Figure 1. This is repeated for each search of maximum between levels in the Gaussian pyramid.

At each level k we extract in the complex filter response image c_{nk} at the position $(x, y)_{nk}^{max}$ found in the enhanced certainty image μ_k^{enh}. We call these complex filter responses c_{nk}^{max}.

3 Registration by 1D Projections of Orientation Images

One class of fingerprints, i.e. class arch (see Figure 1 to the right), lacks SPs [7]. In noisy fingerprints the complex filtering can give a too weak response to classify the point as a core or a delta point. Also when the sensor area of the capturing device is small the SPs are not always found within the captured fingerprint. In these situations symmetry point extraction will fail and must be complemented by an alternative method. We call this method "Registration by 1D projections of orientation images" which makes use of the global orientation field of the fingerprint but does not need SPs for registration. The method is based on a decomposition of the fingerprint into several images, where each image, O_k, corresponds to a direction. Called *Orientation images* in what follows, they were 6 in number, representing 6 equally spaced directions in our experiments.

By *a pair of Orientation images* we mean two orientation images, one from the template fingerprint and one from the test fingerprint, belonging to the same orientation value. The difference in position of a pair of orientation images, is used to estimate the translation between the template and the test fingerprint (it is assumed that the rotation is negligible, or have been compensated for, between the two fingerprints). From each of the orientation images several 1D projections at different angles (*radiograms*) are computed [5]. We call the two radiograms computed from a pair of orientation images at the same projection angle *a pair of radiograms*. A correlation is computed between each pair of radiograms. From the peak in the correlation measure we estimate a displacement for each such pair of radiograms.

In the estimation of the translation parameter we make use of two displacements computed from pair of radiograms which are perpendicular in projection angle. The final estimate of the translation between the template and the test fingerprint is computed from the total of $n_{or} * \frac{n_{pr}}{2}$ number of estimates, where n_{or} and n_{pr} are the number of orientation images and the number of projection angles respectively.

3.1 Orientation Radiograms

An orientation image is computed according to equation 2.

$$O_k = |z|\, e^{\alpha(cos^2(\theta_k - \varphi) - 1)} \tag{2}$$

In this equation Θ_k is the pass orientation angle for an orientation image O_k and φ is the orientation of the orientation field z in each point in the interval $\left[-\frac{\pi}{2}\ \frac{\pi}{2}\right]$. The constant α controls the sensitivity in the selection of orientation angles close to the pass angle. Figure 4 shows orientation images for pass angles $-\frac{\pi}{3}, -\frac{\pi}{6}, 0, \frac{\pi}{6}, \frac{\pi}{3}$ and $\frac{\pi}{2}$ at level 2 in the Gaussian pyramid when the input is the fingerprint given in Figure 1 to the right.

The Radon transform is used to compute 1D projections of orientation images in the direction of ϕ according to equations 3 and 4. Figure 4 shows radiograms

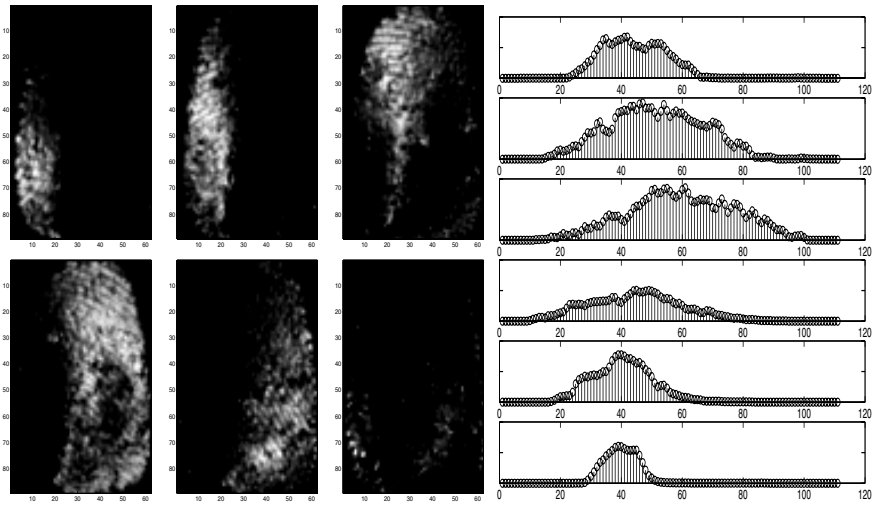


Fig. 4. Left: Orientation images (level 2) for the fingerprint to the right in Figure 1. With pass angles $\theta_k = -\frac{\pi}{3}, -\frac{\pi}{6}, 0$ for the upper row from left to right, and $\theta_k = \frac{\pi}{6}, \frac{\pi}{3}$ and $\frac{\pi}{2}$ for the bottom row from left to right. Right: Radiograms computed from the orientation image to the left in this figure with pass angle $-\frac{\pi}{6}$. Projection angles from top to bottom are $-\frac{\pi}{3}, -\frac{\pi}{6}, 0, \frac{\pi}{6}, \frac{\pi}{3}$ and $\frac{\pi}{2}$

for the orientation image to the left with a pass angle of $-\frac{\pi}{6}$. Radon transform amounts to summing the pixel values along the direction ϕ.

$$R_\phi(x') = \int f(x'cos\phi - y'sin\phi, x'sin\phi + y'cos\phi)\, dy' \tag{3}$$

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\phi & sin\phi \\ -sin\phi & cos\phi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{4}$$

3.2 Translation Estimation

The relation between the displacement dx' between a pair of orientation radiograms and the translation dx and dy of its pair of orientation images is computed according to equation 5.

$$dx'_{\phi_m} = cos\phi_m dx + sin\phi_m dy \tag{5}$$

Where $dx' = x'_{template} - x'_{test}$, $dx = x_{template} - x_{test}$, and $dy = y_{template} - y_{test}$.

The displacement dx' is estimated from data of a certain projection angle ϕ_m by finding the peak in the correlation signal of each pair of orientation radiograms. By using two pairs of radiograms, which are perpendicular in projection angle, we can estimate the translation dx and dy between the template fingerprint and the test fingerprint by equation 6.

$$\begin{bmatrix} dx'_{\phi_m} \\ dx'_{\phi_n} \end{bmatrix} = \begin{bmatrix} cos\phi_m & sin\phi_m \\ -sin\phi_m & cos\phi_m \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} \tag{6}$$

Where $\phi_m = -\frac{\pi}{3}, -\frac{\pi}{6}, 0$ and $\phi_n = \phi_m + \frac{\pi}{2}$, i.e. $\frac{\pi}{6}, \frac{\pi}{3}$ and $\frac{\pi}{2}$. From each pair of orientation images we get $\frac{n_{pr}}{2}$ number of estimates. Out of a total of $n_{or} * \frac{n_{pr}}{2}$ estimates we want to select, in a robust way, the final translation estimate dx, dy. First we apply an outlier detection within an orientation image by disregarding estimates $(dx, dy)^T$ that are most dissimilar to other estimates. Second we take away orientation images that have a high variance in their estimates. Finally we estimate the translation by taking the mean value of the estimates.

4 Experiments

The FVC2000 fingerprint database, db2 set A [15] is used in the experiments. A total of 800 fingerprints (100 persons, 8 fingerprint/person) are captured using a low cost capacitive sensor. The size of an image is 364 x 256 pixels, and the resolution is 500 dpi. It is worth to note that FVC2000 is constructed for the purpose of grading the performance of fingerprint recognition systems, and contains many poor quality fingerprints.

4.1 Symmetry Point Extraction

The filters used in the multi-scale filtering are of size 11 x 11 (a standard deviation of the Gaussian of 1.6). From the multi-scale searching for maximum in the enhanced certainty images μ_{nk}^{enh}, as described in section 2.3, the position of maximum $(x, y)_{nk}^{max}$ is extracted for each level k and for each type n of SP (in the lowest resolution level the search for maximum is done in the ordinary certainty image μ_{nk}). In the position $(x, y)_{nk}^{max}$ the complex filter responses c_{nk}^{max} are extracted and saved for each level k and for each type n of SP. We compute new filter responses R from the extracted complex filter responses c_{nk}^{max} according to equations 7 and 8, i.e. we sum the complex filter responses c_{nk}^{max} in the levels k (vector-sum) for respective type of SP. The final response is the mean of the magnitude of the vector-sum.

$$R_{core} = \frac{1}{4} \left| \sum_{k=1}^{4} c_{1k}^{max} \right| \tag{7}$$

$$R_{delta} = \frac{1}{3} \left| \sum_{k=1}^{3} c_{2k}^{max} \right| \tag{8}$$

To test the performance of the symmetry point extraction (method 1) the true position $(x, y)_n^{true}$ of the SPs have been manually extracted for the fingerprints in the database. Those fingerprints that are lacking SPs are marked manually for being so. An SP is "correct in position" if the Euclidean distance d between the true position and the extracted position at resolution level 1 $(x, y)_{n1}^{max}$, by method 1, is ≤ 15 pixels[1], and the filter responses R_{core} respectively R_{delta} are

[1] Approximately 1.5 wavelength of the fingerprint pattern

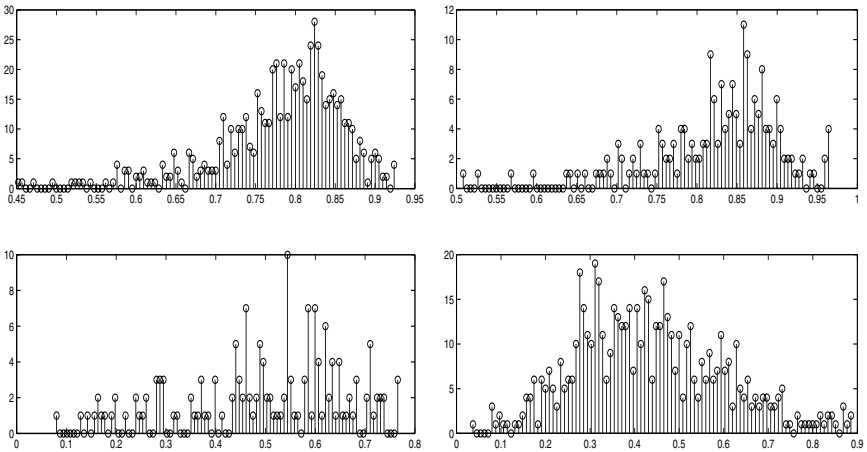


Fig. 5. Distributions of R_{core} (left) and R_{delta} (right). Left: core points that are "correct in position" (top), and core points that are "not correct in position" (bottom). Right: delta points that are "correct in position" (top), and delta points that are "not correct in position" (bottom)

high, i.e. higher than a threshold. Figure 5 shows the distribution of the filter response R_{core} for "correct in position" extracted core points (left/top) and for core points "not correct in position" (left/bottom) and the distribution of the filter response R_{delta} for "correct in position" extracted delta points (right/top) and for delta points "not correct in position" (right/bottom). From these distributions we can estimate the performance for method 1 for different values of thresholds.

If we put the threshold for core point acceptance $th_{core} = 0.63$ we get an EER of 4 % and for the delta points we get an EER of 3 % when the threshold for acceptance for a delta point $th_{delta} = 0.73$. In 665 fingerprints out of 800 we find a core point, or a delta point, or both "correct". By correct we mean both close in distance to the true position (closer than 15 pixels) and a high response from the symmetry filter, i.e. $R_{core} > th_{core}$ respective $R_{delta} > th_{delta}$.

In figure 6 the histograms of the error in distance for the "correctly" estimated SPs are shown. The mean value of the error in distance is approximately 5 $\pm$ 3 pixels.

4.2 Orientation Radiograms

We apply method 1 to obtain SPs. For those fingerprints which does not contain sufficiently strong SPs we apply the alternative method discussed in section 3. Method 1 finds a symmetry point, in 665 fingerprints out of 800. The method 2 is tested on the remaining 135 fingerprints. We call these 135 fingerprints the *SP-free set*. We use a jack-knife strategy to measure the performance of method 2, using the rotation principle because we rotate the test data with the template data to obtain the maximum available trials. This is motivated by that method

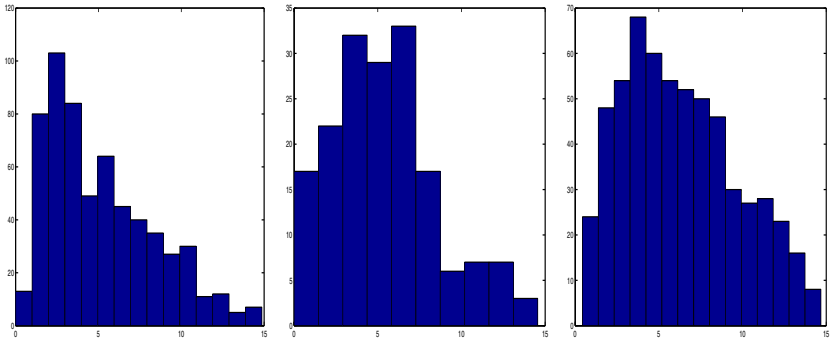


Fig. 6. Histograms of the error in distance for "correct" SPs. To the left for core points and in the middle for delta points. The mean error in distance is approximately 5 ± 3 pixels for both type of SPs. To the right the histogram of the error for the "correctly" estimated translation parameters. The mean error is approximately 6 ± 3 pixels

1 leaves too few samples for method 2 to work on despite the fact that the size of the FVC database is appreciably large. For each fingerprint in the SP-free set (the test fingerprint) we estimate the translation parameters by using the rest of the fingerprints for that person as template fingerprints. The templates may or may not have been found by method 1. In this way we obtain 7 estimates for each test fingerprint, that is a total of $7 * 135 = 945$ translation estimates.

In the experiments we have used 6 orientation images n_{or} with pass angles $-\frac{\pi}{3}, -\frac{\pi}{6}, 0, \frac{\pi}{6}, \frac{\pi}{3}$ and $\frac{\pi}{2}$ and 6 projection angels n_{pr} equal in value to the pass angles. This gives 3 estimates for each orientation pair, and a total of 18 translation estimates. The orientation images O_k are computed using the orientation field z at level 1 in the Gaussian pyramid, the parameter α found empirically is put to 8.6 in equation 2.

In the processing of the 18 estimates of $(dx, dy)^T$ we obtain new estimates stemming from within and between orientation images. First, within an orientation image, we take away one estimate out of 3. The one which is most dissimilar to the other two is disregarded. Second we keep 3, i.e. the 3 orientation images which have minimum variance in their estimates. Third we keep the two orientation images that are closest in the mean of their estimates. Now we have two orientation images, and two estimates of translation for each object. Fourth we take as the final estimate the mean of the two estimates belonging to the orientation image that shows minimum variance.

Figure 7 shows the result. The distance metric is the Euclidean distance between the true translation and the estimated translation. If we assume that the error in the translation estimate is acceptable if the euclidean distance $d \leq 15$ pixels (we name this "correct" estimation) the above method finds the true translation in 588 trials out of 935 possible trials. In figure 6 the histogram of the error for the "correctly" estimated translation parameters is shown. The mean value of the error is approximately 6 ± 3 pixels.

For each test fingerprint (a total of 135 tests, 7 estimates in each test) the possible outcomes that are "correct", i.e. $d \leq 15$, is in the range [0 7]. To the

FVC2000 database, 935 trials in the test.	Number of fingerprints	Ratio %
distance ≤ 15	588	62
distance ≤ 20	647	68
distance ≤ 30	716	76

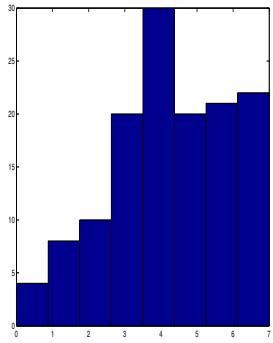


Fig. 7. Results from method 2. Left: Distances from true translation. Right: Histogram number of outcomes (maximum =7 and minimum=0)

right in Figure 7 is the histogram for this test. The mean value is 4.3 which means that in the mean approximately 4 estimates out of 7 are "correct" for each test fingerprint by using method 2 in isolation on fingerprints that are rejected by method 1.

4.3 Registration Using the Combined Methods

Using method 1 we detect the position of an SP (core or delta, or both) in 665 fingerprints out of 800 with an acceptable error in distance of 5 ± 3 pixels. In [4] we have shown that using SP-registration in isolation an unbiased alignment error with a standard deviation of 13 pixels (which approximately is the average wavelength in the fingerprint) can be achieved. We also present a performance measure of the estimation of the geometric orientation of an SP to be unbiased with a standard deviation of less than 4°. Using SP-registration with the 665 correctly extracted SPs, and assuming the same alignment error as in [4], we achieve a registration performance of 83% for SP-registration in isolation.

The alternative method (method 2) running on the 135 fingerprints missed by method 1 estimates correctly 588 trials out of 945 possible trials (62%) with a mean error of 6 ± 3 pixels. With this performance for method 2, we estimate the translation parameter in an acceptable way for 84 fingerprints of 135, missed by method 1. The 135 fingerprints were not compensated for orientation differences. However, for the 84 fingerprints in which a "correct" translation estimate was found, the orientation difference is small (because of how the translation estimation was implemented) and therefore also the rotation difference is small. Accordingly, it can be concluded that a registration performance of 62% is achieved with this method in isolation with an alignment error of similar order as for method 1.

To conclude, by using method 1 and method 2 jointly we estimate the translation parameter "correctly" for 749 $(665 + 84)$ fingerprints out of total 800, yielding an identification performance of 94%. This is done without use of minutiae, and without rotation compensation for method 2.

5 Conclusion and Future Work

In this paper a multi-expert registration system is built using non-minutiae features which makes the suggested method fully complementary to minutiae based methods.

The registration performance for the method *registration by symmetry points* was 83% when running in isolation. Combined with the method *registration by 1D projections of orientation images* the registration performance was increased to 94%. This shows that a combination of registration methods, i.e to use a *one modality multi-expert* registration system, instead of using one registration method in isolation increase the system registration performance. The achieved uncertainty (one standard deviation) of 13 pixels in the alignment error is approximately of the same size as other studies used, e.g. [16].

The 94% performance in the estimation of the translation parameter was achieved when the fingerprints for method 2 were not compensated for rotation differences. Before estimating the translation we can compensate for the rotation differences between the test and the template orientation images by a rough orientation estimation technique, such as orientation histogram correlation. This should increase the performance of registration for method 2.

Acknowledgment

This work has been possible by support from the Swedish national SSF-program VISIT.

References

1. S. Huvanandana, C. Kim, and J. N. Hwang. Reliable and fast fingerprint identification for security applications. *Proc. Int. Conf. on Image Processing*, 2:503–506, 2000.
2. D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of fingerprint recognition.* Springer, New York, 2003.
3. V. Govindu and C. Shekhar. Alignment using distributions of local geometric properties. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):1031–1043, October 1999.
4. K.Nilsson and J. Bigun. Localization of corresponding points in fingerprints by complex filtering. *Pattern Recognition Letters*, 24(13):2135–2144, September 2003.
5. S. Michel, B. Karoubi, J. Bigun, and S. Corsini. Orientation radiograms for indexing and identification in image databases. *In Eusipco-96, European conference on signal processing*, pages 1693–1696, 1996.
6. M. Kawagoe and A. Tojo. Fingerprint pattern classification. *Pattern Recognition*, 17(3):295–303, 1984.
7. K. Karu and A. K. Jain. Fingerprint classification. *Pattern Recognition*, 29(3):389–404, 1996.
8. A. M. Bazen and S. H. Gerez. Systematic methods for the computation of the directional fields and singular points of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):905–919, July 2002.

9. J. Bigun and G. H. Granlund. Optimal orientation detection of linear symmetry. *IEEE Computer Society Press, Washington, DC*, pages 433–438, June 1987. In First International Conference on Computer Vision, ICCV (London).
10. J. Bigun. Recognition of local symmetries in gray value images by harmonic functions. *Ninth International Conference on Pattern Recognition, Rome*, pages 345–347, 1988.
11. H. Knutsson, M. Hedlund, and G. H. Granlund. Apparatus for determining the degree of consistency of a feature in a region of an image that is divided into discrete picture elements. *US. Patent, 4.747.152*, 1988.
12. J. Bigun, T. Bigun, and K.Nilsson. Recognition by symmetry derivatives and the generalized structure tensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1590–1605, 2004.
13. B. Johansson. *Multiscale curvature detection in computer vision.* Tech. lic., Linkoping University, Linkoping University, Dep. Electrical Eng., SE-581 83, 2001.
14. G. Drets and H. Liljenstrom. Fingerprint sub-classification and singular point detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 12(4):407–422, June 1998.
15. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2000: Fingerprint verification competition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):402–412, March 2002.
16. A. M. Bazen and S. H. Gerez. Elastic minutiae matching by means of thin-plate spline models. *Proc of the 16th Int. Conf. on Pattern Recognition*, 2:985–988, August 11-15 2002.

A Feature Map Consisting of Orientation and Inter-ridge Spacing for Fingerprint Retrieval

Sung-Oh Lee[1,3], Yong-Guk Kim[2], and Gwi-Tae Park[1]

[1] Dept. of Electrical Engineering, Korea University, Seoul, Korea
{nbon,gtpark}@korea.ac.kr
[2] School of Computer Engineering, Sejong University, Seoul, Korea
ykim@sejong.ac.kr
[3] NITGEN Co., Ltd., Seoul, Korea

Abstract. We propose a new fingerprint classification method based on a feature map consisting of orientation and inter-ridge spacing for the latent fingerprint retrieval within the large-scale databases. It is designed for the continuous classification methodology. This method captures unique characteristics for each fingerprint from the distribution of combined features of orientation and inter-ridge spacing of local area. The merit of the proposed approach is that it has translation invariant property and is rubust againt registration error since it is not necessary to locate the core position. Our experiments show that the performance of the proposed approach is comparable to the MASK method, and when it is combined with other classifier i.e. PCASYS, the result classifier outperformes any single classifier previously proposed. Moreover, it can be implemented in the low cost hardware such as embedded fingerprint system since the new algorithm saves the processing time.

1 Introduction

In this paper we focus on continuous classification of fingerprints [2]. In that case, each fingerprint can be characterized as a feature vector within the multi-dimensional space. By assuming that similar fingerprints are mapped into close points, the retrieval problem can be solved as a nearest-neighbor search. In such case, we are able to sidestep the problem of exclusive membership of 'ambiguous' fingerprints and to control the reliability of the system by adjusting the size of the neighborhoods.

In general, the goal of feature extraction in a pattern recognition task is to extract information from the input data that is useful for determining its category. Classification of fingerprint is mainly based on its global pattern of ridges and valleys. Since the ridges and valleys normally display well-defined frequency and orientation, we propose that a map consisting of orientation and inter-ridge spacing could be a useful way of classifying fingerprints.

The advantage of the present method is that we do not need to locate the core position for the registration since the map is built by simply scanning the local fingerprint area. We have found that the performance of this system is comparable to that of the MASK method.

In section 2, we review PCASYS developed from NIST. The proposed map and feature extraction are described in section 3. Section 4 discusses about combining classifiers. Then, experimental results are given in section 5. In section 6, conclusions and discussion will be described.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 184–190, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 PCASYS Classifier

PCASYS (Pattern-level Classification Automation SYStem) [4] is designed to classify fingerprints into 6 different classes by analyzing their orientation field. The continuous classification approach for PCASYS is derived in [2] as illustrated in Fig. 1. Segmentation and enhancement of input fingerprint is first carried out, before computing the orientation field based on the method developed by Stock and Swonger [6]. Then, registration of the orientation field around the core position is performed. It is able to reduce the size of the orientation vector by 64 using KL (Karhunen-Loeve) transformation. That vector is used in indexing a fingerprint database by measuring Euclidian distance. Although the auxiliary module of PCASYS is designed for the ridge tracing that works by analyzing the ridge-line concavity under the core position, we would not include such step for our system.

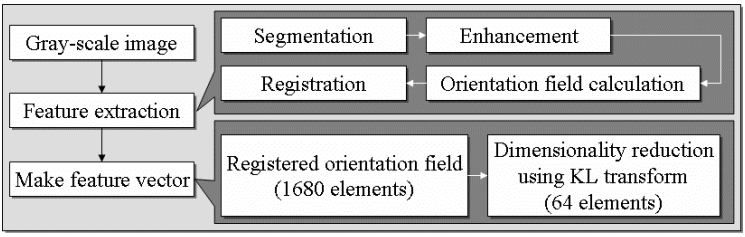


Fig. 1. Block diagram of PCASYS

3 Feature Map Consisting of Orientation and Inter-ridge Spacing

The main idea of the feature map we propose here is to define a ridge feature consisting of unique orientation and inter-ridge spacing within a relatively small finger print area and then to classify the feature map which contains the statistical distribution of the ridge feature of the given fingerprint. Fig. 2 shows the functional block diagram of the proposed method, which consists of 3 steps: first computation of the ridge features (feature extraction); secondly, construction of the feature map (map building) and thirdly, construction of small feature vector for the classification. All the procedures consider only gray-level fingerprint images partitioned into square blocks of NxN pixels. A 512x512 fingerprint image having 256-level from the NIST-SDB4 database [7] is partitioned into 28x30 square blocks with $N = 16$.

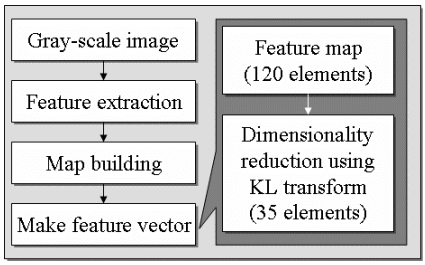


Fig. 2. Block diagram of the proposed method

3.1 Feature Extraction

Although several methods by which the ridge feature is extracted from fingerprint ridge have been proposed [8], we have used that the extraction is carried out by the minutia-basis fingerprint recognition method. It again consists of 4 steps: first, segmentation of the fingerprint image; secondly, enhancement of the segmented image; thirdly, computation of the orientation field and fourthly, inter-ridge spacing field as illustrated in Fig. 3. The fingerprint segmentation, pre-processing, and orientation image computation are performed as described in [4]: the finger area is separated from the background and its quality is improved by a filtering in the frequency domain and then the orientation image is calculated at every pixel. The local ridge orientation is usually specified for a block rather than at every pixel. An image is divided into a discrete grid 28x30 using the method proposed by Hong and Jain [9]. The orientation field O is defined as a 28x30 blocks, where each element represents the local ridge orientation $(0^\circ, 180^\circ)$ at block (i, j) denoted by a value $O(i, j) \in \{0,1,...,7\}$.

The average inter-ridge spacing (ridge distance) of fingerprint images has been used in many cases. In our method, each average local ridge distance is considered as a unique feature that represents local fingerprint area along with local ridge orientation. There are two different approaches to calculate ridge distance: the geometric approach and the spectral approach [10]. For computational efficiency, we use the geometric approach to compute the average ridge distance like the Hong and Jain method [9]. The inter-ridge spacing field R is defined as a 28x30 blocks, where each element represents the average local ridge distance at block (i, j) denoted by a value $R(i, j) \in \{0,1,...,20\}$. We limit the value over 20 in illustrating the steps as shown in Fig. 4.

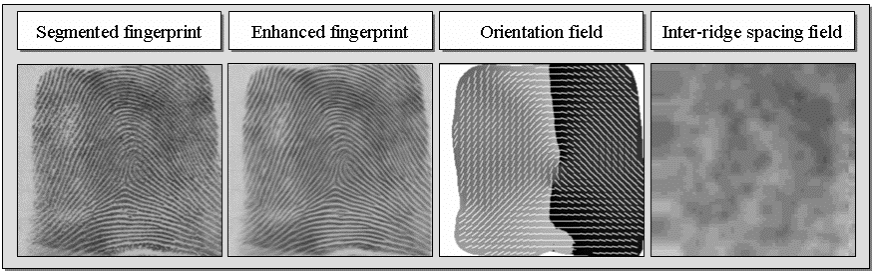


Fig. 3. Four steps of feature extraction

3.2 Map Building

In previous section, both the orientation field and the inter-ridge spacing field are calculated in the same fingerprint area, and then each local fingerprint block (i, j) is represented by a vector $\mathbf{v} = [O(i, j), R(i, j)]$. Notice, however, that we do not consider this information as a vector of 1680 elements by registering with respect to the core position as described in PCASYS, we rather construct a distribution matrix where each vector is mapped repeatedly. The feature map Map is defined as an 8x15 matrix, where each element represents a distribution of the same vector $\mathbf{v}$ at matrix (k, l) .

Each row and column of the matrix represents orientation and inter-ridge spacing value, respectively. The feature map value $Map(k,l)$ is defined as

$$Map(k,l) = \sum_{j=1}^{30}\sum_{i=1}^{28}\{u(\delta(k-1-O(i,j))\cdot\delta(l+5-R(i,j)))\} \quad (1)$$

where $k \in \{1,2,...,8\}$ and $l \in \{1,2,...,15\}$. $u(x)$ is the unit step function defined by $u(x)=0$ for $x \le 0$ and $u(x)=1$ for $x>0$. $\delta(x)$ is Dirac's delta function defined by $\delta(x)=0$ for $x \ne 0$ and $\delta(x)=\infty$ for $x=0$. The number of columns is 21 since the maximum value of ridge distance is 20. And yet the distribution shown in Fig. 4 enables us cropping values below 6.

The feature map is normalized by a constant mean and variance. Such normalization enables us operating within the fixed range and this makes fingerprint retrieval through spatial data structures easier. Let $N(k,l)$ denote the normalized value at matrix (k,l) and M and V the estimated mean and variance, respectively. The normalized feature map is defined as

$$N(k,l) = \begin{cases} M_0 + S(k,l), & if\ Map(k,l) > M \\ M_0 - S(k,l), & otherwise \end{cases} \quad S(k,l) = \sqrt{\frac{V_0}{V}(Map(k,l)-M)^2} \quad (2)$$

where M_0 and V_0 are the desired mean and variance values, respectively. For our experiments, we set the value of both M_0 and V_0 to 100. Fig. 4 shows functional steps of the map building procedure described above.

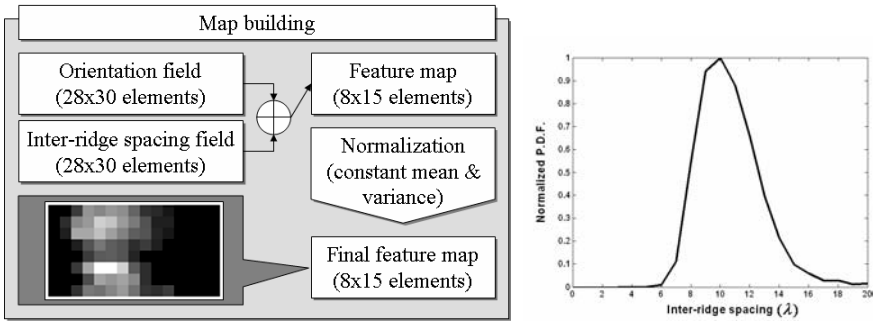


Fig. 4. Left: Functional steps of map building, Right: Distribution of the inter-ridge spacing

Various examples of the final feature map belongs to 5 different classes (Plain Arch, Left Loop, Right Loop, Tented Arch and Whorl) are shown in Fig. 5. One can distinguish the distributions of different fingerprints. The proposed feature map has the following properties:

1. This representation is translation invariant; therefore we do not need to find fingerprint image singularities in alignment step.
2. This representation does not cover rotation but offers some rotation invariance.
3. This representation only fits continuous classification for fingerprint retrieval problem, not exclusive classification since it does not represent global features of each class but each fingerprint.

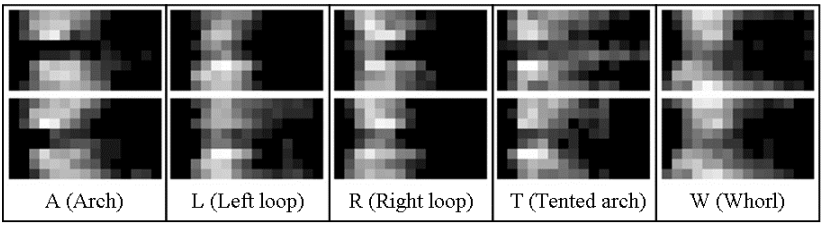


Fig. 5. Each feature map in the figure belongs to a different class: Plain Arch, Left Loop, Right Loop, Tented Arch and Whorl. Top and bottom are first and second instance per finger

3.3 Classification

As shown in Fig. 2, the dimension of the feature map, considered as a vector of 120 elements, is reduced using well-known dimension reduction method, principle component analysis (KL transform) [5]. In continuous classification, each fingerprint is characterized using a vector summarizing its main features, by which similar fingerprints are mapped to the close points in the multidimensional space.

Let $\mathbf{u}$ be a feature map and let $\boldsymbol{\psi}$ be the matrix defining the KL transform which is constituted by the eigenvectors of the covariance matrix. Then the feature vector $\mathbf{w}$ is computed as $\mathbf{w} = \boldsymbol{\psi}^{\mathbf{t}} \cdot \mathbf{u}$. In our experiments, we have found that the optimal number of principle components is 35. If $\mathbf{w}$ is the feature vector extracted from a latent fingerprint and $\mathbf{w}_i$ is the feature vectors in the database, the distances between them is calculated as

$$d_{proposed}(i) = \left\| \mathbf{w} - \mathbf{w}_i \right\|_2 \tag{3}$$

When $\mathbf{w}_i$ is the corresponding one, we can believe that it is sufficiently close to $\mathbf{w}$, so that the distance, $d_{proposed}(i)$, must be smaller than other distances.

4 Combining Classifiers

It is known that combining two classifiers for the continuous fingerprint classification task can outperform using a single classifier only [3]. For instance, performance of combining the MASK and MKL-based classifiers is better than that of each of them. In the similar context, we aim to demonstrate whether when our method is combined with the PCASYS at the measurement level, the overall performance could be improved. According to the method developed by [3], the distance of the combined classifiers is defined as d_{combined}, in which the *bisigm* function is given as below.

$$\begin{aligned} d_{combined} &= (w) \cdot bisigm(d_{proposed}, m_{proposed}, s1_{propoed}, s2_{proposed}) \\ &\quad + (1-w) \cdot bisigm(d_{PCASYS}, m_{PCASYS}, s1_{PCASYS}, s2_{PCASYS}) \\ bisigm(d,m,s1,s2) &= \begin{cases} \dfrac{1}{1+\exp(\dfrac{-2\cdot(d-m)}{s1})} & if\ d < m \\ \dfrac{1}{1+\exp(\dfrac{-2\cdot(d-m)}{s2})} & otherwise \end{cases} \end{aligned} \tag{4}$$

5 Experimental Results

In the present study, NIST Special Database 4 [7] was used as a workbench. DB4 contains 2000 fingerprint pairs, uniformly distributed in the 5 classes. In order to resemble a real distribution (A=3.7%, L=33.8%, R=31.7%, T=2.9%, W=27.9%), we reduce the cardinality of less frequent classes and then obtain eventually 1204 pairs for the experiment. The remaining fingerprints have been used as 'training set'.

In order to compare the performance of continuous classification methods, some fingerprint retrieval operations have been simulated, according to strategies AC and BC [2] by indexing the fingerprint instances F through their feature vectors and by searching the corresponding instances S.

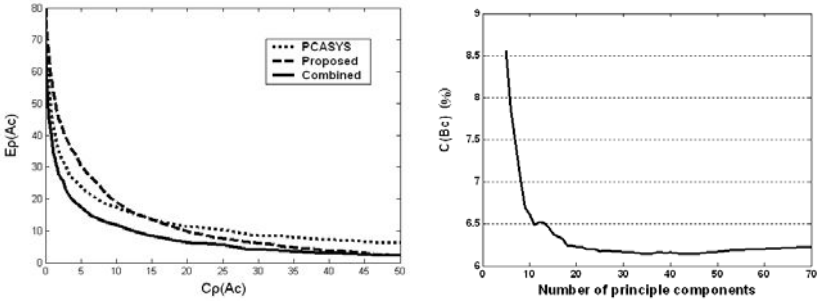


Fig. 6. Left: Trade-off between the portion of database searched and the retrieval error, varying the tolerance ρ , for the individual and the combined classifiers. Some specific values are highlighted in Table 1. Right: Trade-off between the number of principle components and the portion of database searched $C(Bc)$. The best result reported at 35

Table 1 and 2 summarizes performances of four methods, i.e. PCASYS [4], LUMINI [2], MASK [1] and the proposed, for the strategy AC and the strategy BC, respectively, by plotting each pair $E_\rho(Ac)$, $C_\rho(Ac)$ relative to the same ρ as a single point. The present method outperforms the PCASYS and LUMINI with a large margin and yet shows slightly lower performance than MASK. Moreover, notice that when the present method is combined with the PCASYS, the performance is better than any single classifier within most range.

Table 1. Strategy AC: comparison among PCASYS, LUMINI, MASK, Proposed, Combined

Eρ(Ac)	Cρ(Ac)				
	PCASYS	LUMINI	MASK	Proposed	Combined
4%	74%	50%	29%	38%	31%
6%	50%	34%	22%	30%	22%
8%	35%	26%	20%	23%	16%
10%	26%	20%	14%	19%	12%
14%	14%	12%	11%	14%	7%

Table 2. Strategy BC: comparison among PCASYS, LUMINI, MASK, Proposed, Combined

C(Bc)	PCASYS	LUMINI	MASK	Proposed	Combined
	7.52%	6.90%	5.22%	6.14%	3.85%

6 Conclusions and Discussion

In this work, a new fingerprint classification method based on a feature map, consisting of orientation and inter-ridge spacing, is proposed for the latent fingerprint retrieval within the large-scale databases. The proposed method captures unique characteristics for each fingerprint from the distribution of combined features of orientation and inter-ridge spacing of local area. And, it appears to be fit well for the continuous classification methodology. Our experiments show that the performance of the proposed approach is comparable to the MASK method and that of the combined classifier outperformes any single classifier. Although we have only combined the present method with the standard PCASYS, the overall performance would be enhanced greatly if our method would be combined with, for instance, the MASK or MKL method.

The other important merit of the proposed approach is that it has translation invariant property. The registration-based method like the NIST classifier needs to find core point for alignment. However, it is often hard to find core point correctly in every situation. Such cases result in registration fails and evantually increase the portion of database searched. In real situation, registration problem is crucial because the bad area in core region result in false alignment. Our method is robust againt such situation comparing to the registration based approach such as PCASYS.

By using only the common features, i.e. orientation and inter-ridge spacing, which are essential in enhancing the performance of fingerprint extraction, our method add just a small portion of time to calculate the proposed feature vector. In most fingerpint recognition systems, the average time required for the whole extraction processes is an important factor for the low cost hardware specification such as embedded fingerprint system.

References

1. Cappelli R, Lumini A, Maio D, Maltoni D. Fingerprint classification by directional image partitioning. IEEE Trans Pattern Analysis and Machine Intelligence 1999; 21(5): 402-421
2. Lumini A, Maio D, Maltoni D. Continuous vs. exclusive classification for fingerprint retrieval. Pattern Recognition Letters 1997; 18(10): 1027-1034
3. Cappelli R, Maio D, Maltoni D. A multi-classifier approach to fingerprint classification. Pattern Analysis and Applications (2002)5:136-144
4. Candela GT et al. PCASYS – A Pattern-Level Classification Automation System for Fingerprints. NIST Tech. Report NISTIR 5647, 1995
5. Fukunaga K. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, 1990
6. Stock R.M. and Swonger C.W., Development and evaluation of a reader of fingerprint minutiae. Cornell Aeronautical Laboratory tech. Report CAL no. XM-2479-X-1:13-17
7. Watson CI, Wilson CL. NIST Special Database 4, Fingerprint Database, U.S. National Institute of Standards and Technology, 1992
8. Yager N, Amin A. Fingerprint classification: a review, Pattern Anal Applic 2004; 7:77-93
9. Hong L, Wan Y, Jain A. Fingerprint image enhancement: algorithm and performance evaluation. IEEE Trans Pattern Analysis and machine Intelligence 1998; 20:777-789
10. Kovacs-Vajna M, Rovatti R, Frazzoni M: Fingerprint ridge distance computation methodologies. Pattern Recognition 33(1): 69-80 (2000)

A Graph Matching Based Approach to Fingerprint Classification Using Directional Variance

Michel Neuhaus and Horst Bunke

Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
{mneuhaus,bunke}@iam.unibe.ch

Abstract. In the present paper we address the fingerprint classification problem with a structural pattern recognition approach. Our main contribution is the definition of modified directional variance in orientation vector fields. The new directional variance allows us to extract regions from fingerprints that are relevant for the classification in the Henry scheme. After processing the regions of interest, the resulting structures are converted into attributed graphs. The classification is finally performed with an efficient graph edit distance algorithm. The performance of the proposed classification method is evaluated on the NIST-4 database of fingerprints.

1 Introduction

Fingerprint classification refers to the process of assigning fingerprints in a consistent and reliable way to classes. The main objective is to reduce the complexity of the general fingerprint identification problem, where a fingerprint is to be matched against large databases of fingerprints. The fingerprint classification problem is considered to be difficult because of the large within-class variability and the small between-class separation. For many years, classification methods from various pattern recognition areas have been proposed, commonly divided into rule-based, syntactic, statistical, and neural network based approaches [1, 2]. Although the classification problem is intrinsically of structural nature, it was not until recently that classification systems based on structural pattern recognition methods have been developed [3–5]. In comparison to state-of-the-art classification methods, structural approaches often fall behind in terms of performance. Yet, in the context of multiple classifier combination, structural algorithms have proven effective in improving existing classification methods [5, 6]. We furthermore believe that the strength of structural algorithms has not yet been fully exploited in fingerprint recognition.

In fingerprint identification or verification, where identical fingerprints are to be matched, one usually focuses on local characteristics, such as minutiae points. Conversely, in fingerprint classification, the problem is often addressed by extracting and representing global characteristics, such as the ridge flow or

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 191–200, 2005.
© Springer-Verlag Berlin Heidelberg 2005

singular points [1, 2]. In the present paper, we propose an image filter based on a new definition of directional variance. Following the Galton-Henry classification scheme of five classes, we use the filter to extract regions that are relevant for the classification. Our second contribution consists in applying edit distance based graph matching to the classification problem after extracting the characteristic regions.

In Section 2, the directional variance filter on orientation vector fields is described. A brief review of error-tolerant graph matching follows in Section 3. Section 4 gives a number of experimental results, and some concluding remarks are provided in Section 5.

2 A Directional Variance Algorithm

The key procedure of a large number of fingerprint classification algorithms is based on the robust detection of singular points of the ridge orientation vector field [2]. To assign fingerprints to one of the five classic Henry classes, it is in most cases sufficient to know the number and position of singular points [7, 8]. In this paper, we propose an algorithm for the reliable computation of a directional variance value measured at every position of the ridge orientation field. The variance is defined such that high variance regions correspond to relevant regions for the fingerprint classification task, including singular points.

Weakly related to the statistical variance, we define the directional variance of the ridge orientation field at position (x, y) by

$$\sigma^2_{x,y} = \frac{1}{1-n} \sum_{i,j} \sin^2(\alpha_{i,j} - \bar{\alpha}_{x,y}) \ , \tag{1}$$

where $\alpha_{i,j}$ denotes the vector at position (i, j) of the vector field and the summation is performed over a window of size n around (x, y). The average orientation $\bar{\alpha}_{x,y}$ of the local window around position (x, y) is computed by taking into account that two vectors pointing in opposite directions represent the same orientation [9, 10]. The circular nature of the orientation vectors is also accounted for in the sine term; vectors $\alpha_{i,j}$ that are orthogonal to the local average $\bar{\alpha}_{x,y}$ contribute maximally and vectors close to the local average contribute minimally to the variance.

From Eq. 1 it follows that the directional variance is expected to be low everywhere in smooth orientation fields. But in the local neighborhood of singular points, orientations do not follow a single predominant direction, which is equivalent to a high directional variance. In experiments we could confirm this behavior.

Our objective in this paper is not to detect singular points, but rather to extract regions that allow us to discriminate between fingerprint classes. For this purpose, we propose to use a modified directional variance measure, which differs from the directional variance in Eq. 1 in the computation of the local average orientation $\bar{\alpha}_{x,y}$. In a first step, all orientations are normalized to an angle range

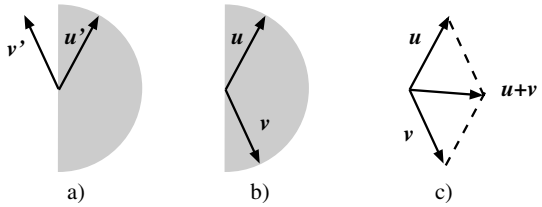


Fig. 1. a) Two vectors representing two ridge orientations, b) the corresponding normalized vectors, and c) their vector sum

of $I = [-\pi/2, \pi/2]$, which corresponds to a vector range of $I' = R_0^+ \times R$. The normalization of two orientation vectors and the normalization range I (see the shaded area) is illustrated in Fig. 1a,b. Normalization consists in reversing any vector that is located outside the shaded area. We proceed by defining the average direction of a number of normalized orientation vectors by their vector sum. In Fig. 1c, the sum of two normalized vectors is illustrated. For a set of vectors in horizontal direction, the vector sum will clearly point in a horizontal direction as well. For a set of vectors in vertical direction, however, the vector sum will not point in vertical direction, but be close to the horizontal direction, as some vectors will point upwards and some will point downwards due to the normalization procedure. In this case, the mean direction $\bar{\alpha}_{x,y}$ does not correspond to the local orientations, which results in a high directional variance. Hence in addition to singular points, the modified directional variance is also responsive to vertical orientation regions. In other words, the proposed new directional variance can be used as a filter that will emphasize not only singular points, but also areas with vertical ridge orientation.

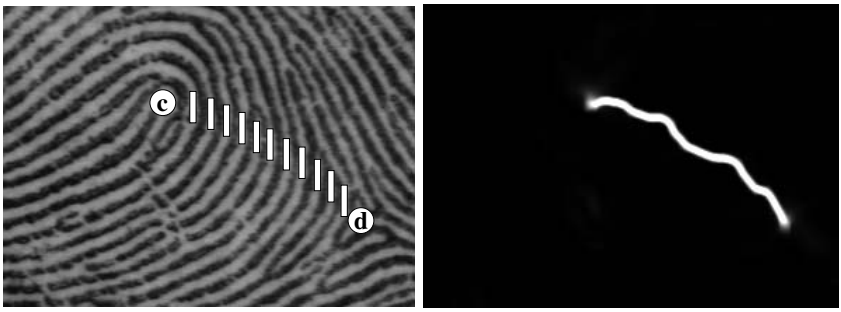


Fig. 2. Left loop fingerprint image a) with core point (c), delta point (d), and marked vertical orientations and b) visualization of the modified directional variance

A closer examination reveals that different fingerprint classes exhibit different characteristics of singular points and vertical orientation regions. Arch fingerprints, for instance, contain no singular points and no vertical ridges, except



a) b) c)

Fig. 3. a) Original fingerprint image, b) visualization of the modified directional variance (bright colors indicate high variance), c) binarized image

for strongly rotated fingerprints. Loop fingerprints, on the other hand, are characterized by a global ridge loop, a core point, and a delta point [7]. The key observation is that one can reach the core point from the delta point via locally almost vertical ridge segments, which is due to the nature of the ridge flow around the delta point and the core point. An illustration of this observation is provided in Fig. 2, where the vertical orientation segments are clearly visible in the loop fingerprint image and in the image resulting from applying the directional variance filter. The same properties are also present in right loop, whorl, and tented arch fingerprints. In a number of experiments, it turns out that the directional variance filter detects the connection between core and delta point much more reliably than a filter simply enhancing vertical orientations. In contrast to other classification methods, the directional variance approach does not solely rely on the detection of singular points, but can also be employed if singular points are not present in the image or distorted by noise.

After filtering the fingerprint, the resulting image is binarized and undergoes a noise removal procedure. The extracted regions can then be used for the purpose of classification. Possible classification criteria include the number of extracted regions and the position and main direction of the regions. An illustration of the extraction of the characteristic regions in a whorl image is shown in Fig. 3. It is easy to verify that the ending points of the two extracted regions correspond to the four singular points, and the regions to the vertical orientation areas of the ridge orientation field. Further examples from the *left loop*, *right loop*, and *whorl* class are shown in Fig. 4. To perform the actual classification based on the extracted regions, various classifiers could potentially be employed. One such method based on graph matching is described in the following sections.

3 Error-Tolerant Graph Matching

Graph matching refers to the process of evaluating the structural similarity of attributed graphs, that is, the similarity with respect to nodes, edges and attributes attached to nodes and edges. A large number of graph matching methods from various research fields have been proposed in recent years [11], ranging from isomorphism-based systems to algorithms based on the spectral decomposition of graph matrices and the definition of positive definite kernel functions on graphs.

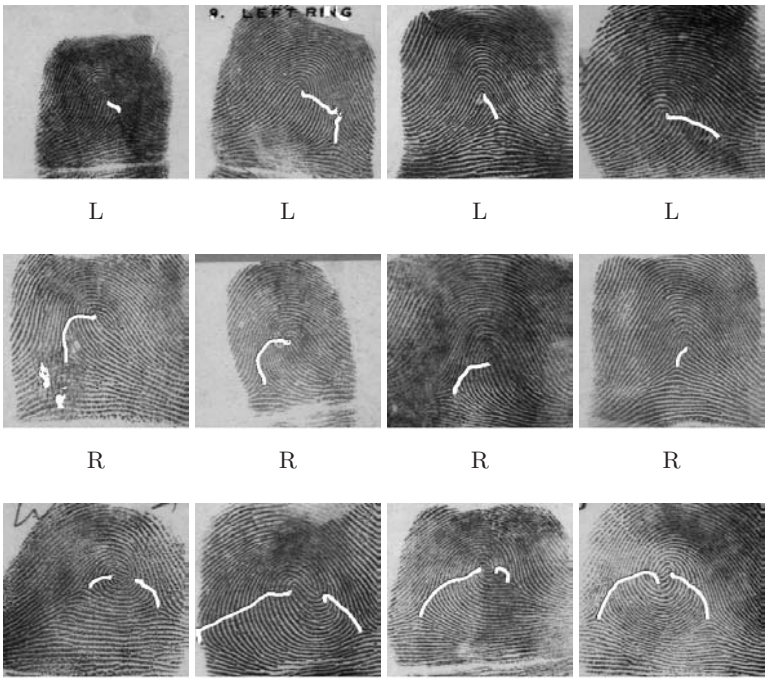


Fig. 4. Visualization of the modified directional variance for *left loop* (L), *right loop* (R), and *whorl* (W) fingerprints

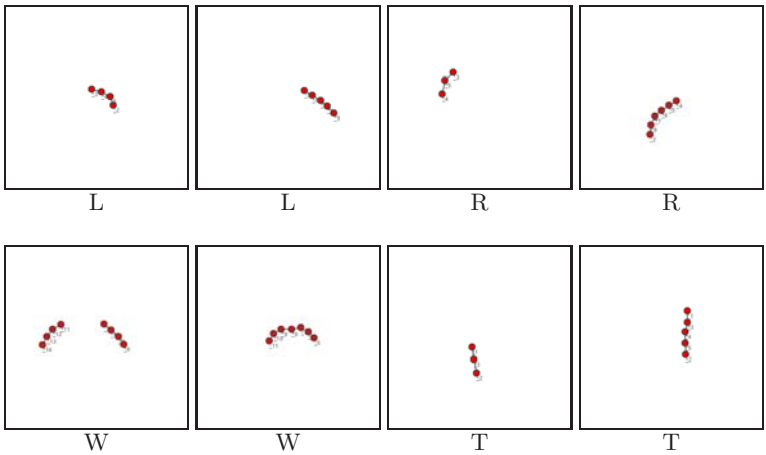


Fig. 5. Sample prototype graphs for the *left loop* (L), *right loop* (R), *whorl* (W), and *tented arch* (T) class

3.1 Graph Edit Distance

One of the most intuitive error-tolerant graph matching approaches consists in the computation of the graph edit distance [12, 13]. The graph edit distance is defined in the context of a basic distortion model, where structural distortions are performed by edit operations on graphs. The standard set of edit operations comprises a node insertion, a node deletion, a node substitution, an edge insertion, an edge deletion, and an edge substitution operation. A sequence of edit operations $(e_1, \ldots, e_l)$ transforming graph g into graph g' is termed an edit path from g to g', and $E(g, g')$ denotes the set of edit paths from g to g'. Given a cost function $c : E(g, g') \rightarrow R^+ \cup \{0\}$ assigning non-negative costs to edit paths, we can then define the edit distance of g and g' by

$$d(g, g') = \min_{p \in E(g,g')} c(p) \quad . \tag{2}$$

The edit distance of two graphs is thus given by the least expensive transformation of the first graph into the second graph in the underlying distortion model. The cost function is usually defined on single edit operations with respect to the attributes attached to nodes and edges.

An edit distance based system can be tailored to a specific application by adjusting the cost functions accordingly. The basic idea is that weak distortions should result in low costs, whereas strong distortions should correspond to higher costs. The cost functions implicitly define, for instance, when the removal of a node n followed by the insertion of another node n' is less expensive than the substitution of n with n', and therefore preferred in an optimal edit path. In other words, the edit distance is derived from the most reasonable explanation of the structural differences of two graphs in the edit operation framework.

The actual computation of the edit distance is performed by constructing and traversing a search tree. In spite of pruning criteria and look-ahead techniques, however, the computational complexity both in terms of running time and memory requirements is high – in fact, it is exponential in the number of nodes of both graphs. For unconstrained graphs of arbitrary size, the edit distance approach is largely unfeasible. Therefore, a fast approximate version of the edit distance algorithm for large graphs is employed in the experiments of this paper. A brief description of the algorithm follows in the next section.

3.2 Approximate Graph Edit Distance

The development of efficient graph matching algorithms for special classes of graphs, for instance bounded-valence graphs, trees, or planar graphs, has been an issue in the graph matching literature for years [11]. In the graph edit distance context an efficient approximate algorithm has recently been proposed that turns out to be very fast and sufficiently accurate for certain graph problems [14]. This approximate algorithm requires the graph nodes to be embedded in the plane. That is, for every node a meaningful position attribute providing a spatial context needs to be present. For graphs extracted from images it is usually easy to derive such a node embedding. Examples include graphs representing interest points and their relations, or region adjacency graphs.

Instead of exploring the full search space, only a subset of all edit paths is considered in the approximate algorithm. Starting from an initial node substitution $n \rightarrow n'$, the least costly transformation from the neighborhood of n to the neighborhood of n' is computed by optimizing local minimum-cost criteria. The computation is performed by means of an efficient cyclic string matching algorithm based on dynamic programming. The result is a valid edit path between two graphs, but not necessarily the optimal one. To account for the dependence on the initialization, the computation is carried out for a number of initial substitutions, and the minimum cost edit path among them is kept. In contrast to the exponential computational complexity of the exact edit distance, the approximate algorithm runs in polynomial time. In practical experiments, the approximation has shown to be feasible and fast, even for large graphs with more than 200 nodes and edges, whereas the exact edit distance algorithm can only be computed for graphs with a size of about 10 nodes [14]. In the following, the approximate edit distance will be used to obtain distance values between fingerprint graphs and subsequently perform the classification.

3.3 Fingerprint Graph Representation and Edit Cost Function

From the results of the region extraction process based on the modified directional variance filter described in Section 2, an attributed graph can be extracted in various ways. In this paper, we follow a simple method to generate structural skeletons. We proceed by applying a one-pass thinning operator [15] to the extracted regions and represent ending points and bifurcation points of the resulting skeleton by graph nodes. Additional nodes are inserted along the skeleton at regular intervals. An attribute giving the position of the corresponding pixel is attached to each node. Edges containing an angle attribute are used to connect nodes that are directly connected through a ridge in the skeleton. An illustration of several graphs of this kind is given in Fig. 5. The simple edit cost function we employ assigns constant costs to insertions and deletions independent of involved attributes; substitution costs are defined proportional to the Euclidean distance of attributes.

4 Experimental Results

The NIST-4 database [16] consists of 4,000 grayscale images of fingerprints with class labels according to the five most common classes of the Galton-Henry classification scheme: *arch*, *tented arch*, *left loop*, *right loop*, and *whorl*. We proceed by extracting an attributed graph from every image as described previously to obtain 4,000 graphs. On the average, these graphs contain 6.1 nodes and 10.3 edges. To classify fingerprint graphs by means of the edit distance, a set of reference graphs for each class needs to be defined. Although an automatic method would be desirable, it proved efficient in recent studies [17, 18] to use a manual construction procedure for this purpose. Adopting a similar approach, we define prototype graphs by manually selecting promising candidates from a training

set of graphs. Where appropriate, a few nodes are deleted from prototype candidates to provide for class representatives as general as possible. By means of this procedure we obtain about 60 prototypes overall. The classification can then be performed based on the nearest-neighbor paradigm: An input graph is assigned the class of the most similar prototype graph. The structural similarity is derived from the corresponding approximate graph edit distance between prototype graph and input graph. An illustration of some prototype graphs is provided in Fig. 5.

The first 1,000 fingerprints from the database are used for the development of the class prototypes and are therefore considered a *Training set*. The remaining 3,000 fingerprints constitute the independent *Test set 1*, and the subset of *Test set 1* consisting of the last 2,000 fingerprints of the database is termed *Test set 2*. The classification rates obtained on the various data sets are summarized in Table 1, where GED refers to the graph edit distance approach proposed in this paper, MASKS, RNN, and GM refer to graph matching approaches reported in [18] using dynamic masks, recursive neural networks, and graph edit distance, respectively, whereas MLP refers to a non-structural neural network approach [19].

From the experimental results we find that the proposed method performs clearly better than the best graph matching approach reported in [18]. A comparison of the training error and test error reveals that a slight overfitting occurs. However, the ability of the graph matching approach to generalize well on unseen data seems to be sufficiently strong. Using the approximate matching algorithm, the classification runs very fast in comparison to other graph edit distance methods. On a regular workstation it takes 27 minutes to conduct a (non-optimized) graph classification of all 4,000 fingerprints of the NIST-4 database. Although the exact edit distance computation would be feasible for these graphs, experiments indicate that the classification takes by far longer (100h instead of 3 minutes for 500 graphs) and results in a lower classification rate.

It is well known that the definition of adequate cost functions is crucial for the performance of a graph edit distance based classification system. In our experiments, we used simple edit costs based on constant costs and Euclidean distances. One major drawback of this edit cost function, and thus a shortcoming of our classification approach, is that all costs are defined in a location-independent way; that is, the information where in the attribute space an edit operation occurs is not taken into account. For a number of graph matching problems, it turns out that location-dependent edit cost functions automatically learned beforehand from a sample set of graphs can significantly improve the recognition performance [20], which may also be of interest in future investigations in the context of fingerprint graph classification.

5 Conclusions

In the present paper we propose a fingerprint classification system by means of error-tolerant graph matching. Our main contribution is an algorithm for the extraction of regions in the ridge orientation field that are relevant for the

Table 1. Fingerprint classification rate on the NIST-4 database

Data set	Classifier	5 classes
Training set	GED	82.6
Test set 1	GED	80.27
Full database	GED	80.85
Test set 2	GED	80.25
	RNN [5, 18]	76.75
	MASKS [17, 18]	71.45
	GM [18]	65.15
	MLP [18, 19]	86.01

classification. Extracted regions correspond to singular points and characteristic connections between core and delta points. To assign one of the five most common Henry classes to fingerprints, we use a graph edit distance approach. In experiments on the NIST-4 fingerprint database, the proposed method is found to outperform graph matching systems reported in recent years. In the future we intend to address the classification problem based on the proposed directional variance with non-structural classifiers and study whether combinations of classifiers may lead to more robust performance results. In addition we plan to investigate if more complex edit cost functions than the one used in this paper could further improve the classification accuracy.

Acknowledgment

This research was supported by the Swiss National Science Foundation NCCR program "Interactive Multimodal Information Management (IM)2" in the Individual Project "Multimedia Information Access and Content Protection". The authors would also like to thank Alessandra Serrau and Prof. Fabio Roli for helpful discussions.

References

1. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer (2003)
2. Yager, N., Amin, A.: Fingerprint classification: A review. Pattern Analysis and Applications **7** (2004) 77–93
3. Maio, D., Maltoni, D.: A structural approach to fingerprint classification. In: Proc. 13th Int. Conf. on Pattern Recognition. (1996) 578–585
4. Lumini, A., Maio, D., Maltoni, D.: Inexact graph matching for fingerprint classification. Machine Graphics and Vision, Special Issue on Graph Transformations in Pattern Generation and CAD **8** (1999) 231–248
5. Marcialis, G., Roli, F., Serrau, A.: Fusion of statistical and structural fingerprint classifiers. In Kittler, J., Nixon, M., eds.: 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication. LNCS 2688 (2003) 310–317
6. Yao, Y., Marcialis, G., Pontil, M., Frasconi, P., Roli, F.: Combining flat and structured representations for fingerprint classification with recursive neural networks and support vector machines. Pattern Recognition **36** (2003) 397–406

7. Karu, K., Jain, A.: Fingerprint classification. Pattern Recognition **29** (1996) 389–404
8. Kawagoe, M., Tojo, A.: Fingerprint pattern classification. Pattern Recognition **17** (1984) 295–303
9. Kass, M., Witkin, A.: Analyzing oriented patterns. Computer Vision, Graphics, and Image Processing **37** (1987) 362–385
10. Bigun, J., Granlund, G.: Optimal orientation detection of linear symmetry. In: Proc. 1st Int. Conf. on Computer Vision, IEEE Computer Society Press (1987) 433–438
11. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. Int. Journal of Pattern Recognition and Artificial Intelligence **18** (2004) 265–298
12. Sanfeliu, A., Fu, K.: A distance measure between attributed relational graphs for pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics **13** (1983) 353–363
13. Bunke, H., Allermann, G.: Inexact graph matching for structural pattern recognition. Pattern Recognition Letters **1** (1983) 245–253
14. Neuhaus, M., Bunke, H.: An error-tolerant approximate matching algorithm for attributed planar graphs and its application to fingerprint classification. In Fred, A., Caelli, T., Duin, R., Campilho, A., de Ridder, D., eds.: Proc. 10th Int. Workshop on Structural and Syntactic Pattern Recognition. LNCS 3138 (2004) 180–189
15. Zhou, R., Quek, C., Ng, G.: A novel single-pass thinning algorithm and an effective set of performance criteria. Pattern Recognition Letters **16** (1995) 1267–1275
16. Watson, C., Wilson, C.: NIST special database 4, fingerprint database. (1992)
17. Cappelli, R., Lumini, A., Maio, D., Maltoni, D.: Fingerprint classification by directional image partitioning. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 402–421
18. Serrau, A., Marcialis, G., Bunke, H., Roli, F.: An experimental comparison of fingerprint classification methods using graphs. In: Proc. 5th Int. Workshop on Graph-based Representations in Pattern Recognition. (2005) Submitted.
19. Jain, A., Prabhakar, S., Hong, L.: A multichannel approach to fingerprint classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 348–359
20. Neuhaus, M., Bunke, H.: A probabilistic approach to learning costs for graph edit distance. In Kittler, J., Petrou, M., Nixon, M., eds.: Proc. 17th Int. Conference on Pattern Recognition. Volume 3. (2004) 389–393

Fingerprint Singular Points Detection and Direction Estimation with a "T" Shape Model

Tong Liu[1,2], Pengwei Hao[1,3] and Chao Zhang[1]

[1] National Laboratory on Machine Perception, Peking University, Beijing, 100871, China
{liutong,phao,chzhang}@cis.pku.edu.cn
[2] School of Mathematical Sciences, Peking University, Beijing, 100871, China
[3] Deptartment of Computer Science, Queen Mary, University of London, E1 4NS, UK

Abstract. As a sort of evident landmark features of fingerprints, singular points (SPs) play important roles in fingerprint alignment, classification and recognition. We present an adaptive "T" shape model of SPs and develop a robust and generic approach to detect SPs and their directions simultaneously. The proposed approach utilizes homocentric sectors around candidate SPs to pick out lateral-axes and further main-axes based on the proposed model. The results of the experiment conducted on a public database, FVC 2002, demonstrate the effectiveness of the method in this paper.

1 Introduction

Among various biometric techniques, automatic fingerprint identification is one of the most popular and reliable techniques for personal authentication. Conventional fingerprint features can be classified into two categories: the minutiae features and the global features [1]. Minutiae which provide the details of the ridge-valley structures are often used in fingerprint matching. On the other hand SPs (including core and delta which can be seen in Fig 1) are points of discontinuity in directional field (DF) of fingerprint images. They are considered as important features of fingerprint macrostructure. Total number of SPs with their individual information (including type, position and direction) and mutual relationships (including relative distance and inclination) can be utilized in fingerprint alignment [2], classification [3] and matching [4]. As an important property of an SP, direction, defined as in which ridges flow away from the SP, provides significant information in fingerprint image registration especially for rotary normalization.

Many methods for detecting SPs have been presented in literature. Nilsson K. and Bigun J. [2] designed a multi-scale filter for complex valued directional tensor field image and measured certainty of SPs with intensity of the response. Hsieh C. and Lu Z. [5] extracted SPs in thinned image by analyzing block directions. Tico M. and Kuosmanen P.'s method [6] was based on multi-resolution representation of DF and certainty level.

Poincaré index (PI), which was first introduced in [7], is a widely exploited feature in detection of SPs [3, 8]. By following a counterclockwise closed contour, which is often chosen as a 2*2 square around a possible SP in the DF, and adding up the difference between the subsequent angles, the resulting cumulative change of orientation is PI. The type of point can be determined by its corresponding PI to core (PI=π), delta (PI=π) or plain one (PI=0). But due to the existence of noise, which is usually caused by scar, bad skin condition or uneven intensity of pressure, many spurious SPs may occur in immediate results.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 201– 207, 2005.
' Springer-Verlag Berlin Heidelberg 2005



Fig. 1. SPs in fingerprint images (core marked by a circle and delta by a square)

Karu K. and Jain A. K. [3] used a 3*3 mean mask to smooth original BDF (Block-wise DF) recursively until reaching a reasonable SP number. But smoothing operation like this often results in transition of SPs' positions or even losing of true SPs. Bazen A.M. and Gerez S.H. [9] performed Gaussian filtering on PDF (Pixel-wise DF) to diminish false SPs. But the choice of the key parameter σ (standard variance) is a dilemma. If it is too big, the speed and accuracy of the algorithm to locating SPs are not ensured, while if it is too small, spurious SPs can't be removed entirely. So simply using a Gaussian filtering operation to solve SPs detection of various fingerprint images doesn't seem to have a good prospect.

To estimate orientation of a core, Wang F. and Zou X. [8] defined Block Orientation Template and searched for least difference neighbor of the core. Then average direction of resulting block and some of its neighbors was accepted as the core's orientation. This method is over-dependent on least difference of orientation which is not credibly an evidence for that the corresponding block's orientation is also the core's. Bazen A.M. and Gerez S.H. [9] introduced two reference models (of core and delta respectively) to simulate neighboring PDF of typical SPs. They took the element-by-element product of the complex conjugated of models and the observed squared gradient data to get rotary angles between observed SPs and ideal patterns. Since constant template can't model various core and delta region, the generalization of this model is not well-founded.

In our approach, PI calculation following comparatively smaller σ's Gaussian filtering of BDF only gives coarse result. Candidate SPs are analyzed with an adaptive model to estimate direction and confidence simultaneously. Furthermore, three directions of delta are estimated apart, which gives more information for further processing than most traditional methods.

This paper is organized as follows. First, in Section 2, the methods of background segmentation, DF estimation and PI calculation are given. Then in Section 3, a "T" shape model is introduced and main-axes are detected. In Section 4, some experimental results are presented. Finally, the conclusion is drawn in Section 5.

2 Detection of Candidate SPs

To get rid of disturbance of the background, which often brings in a number of spurious SPs, we use approach based on transition-number minimization in [10] to segment the

original fingerprint image. Then PDF expressed by $\begin{bmatrix} PDF_x(x,y) \\ PDF_y(x,y) \end{bmatrix}$, is calculated with (1), where $PDF_x(x,y)$ and $PDF_y(x,y)$ represent sine and cosine values of doubled local orientation angle at *(x, y)* [9]. *W* is a small square region centered at *(x, y)*. PDF can be exploited to obtain BDF by (2), where *W′* is the block *(i, j)* and *m* is the size of it. In order to get more precise locations of SPs, *m* is set to be 2. Note that the average orientation of homocentric sectors can also be calculated with PDF which will be described in Section 3.

$$\begin{bmatrix} PDF_x(x,y) \\ PDF_y(x,y) \end{bmatrix} = \begin{bmatrix} G_{xx} - G_{yy} \\ 2G_{xy} \end{bmatrix}$$
$$G_{xx} = \sum_W G_x^2$$
$$G_{yy} = \sum_W G_y^2 \qquad (1)$$
$$G_{xy} = \sum_W G_x G_y$$
where $\begin{bmatrix} G_x \\ G_y \end{bmatrix}$ is gradient vector of point (x, y).

$$\begin{bmatrix} BDF_x(i,j) \\ BDF_y(i,j) \end{bmatrix} = \begin{bmatrix} \frac{1}{m*m} \sum_{W'} PDF_x(x,y) \\ \frac{1}{m*m} \sum_{W'} PDF_y(x,y) \end{bmatrix} \qquad (2)$$

Because of the presence of noise in original BDF, we use Gaussian filter to smooth the BDF to alleviate noise before calculating PI. By setting σ to 3.5, the filter can eliminate the spurious SPs significantly while preserving true ones. An example of filtering effect is demonstrated in Fig 2. In this way, a relatively smaller number of candidate SPs are obtained for further analysis.



Fig. 2. Candidate SPs detected by PI method on (a) original BDF (b) Gaussian filtered BDF (locations of candidate cores and deltas are marked by white and black blocks respectively)

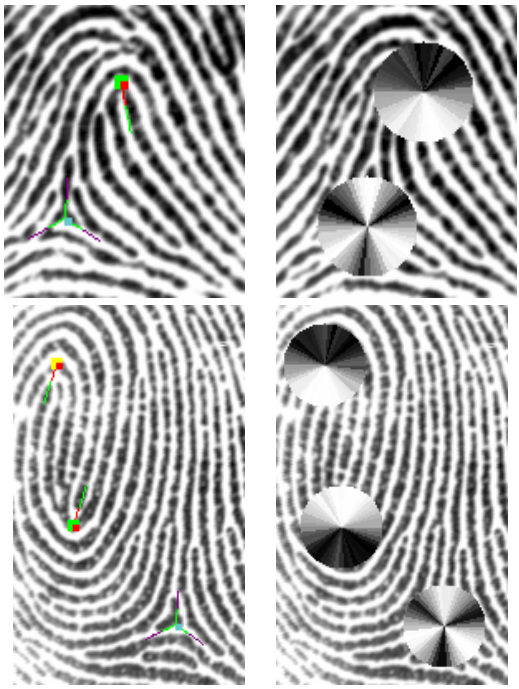


Fig. 3. Illumination of the "T" shape model: grayscale of each sector in circle expresses parallelism of its AOS and the radial orientation

3 Detection of SPs and Estimation of Their Directions

To analyze an SP finely, a direction that can be robustly detected and precisely expressed must be identified. Here we define a main-axis (direction) of an SP as the direction in which ridges in SP's neighborhood tend to leave the SP. Obviously a delta has three main-axes according to this definition. Some examples of SPs' main-axes are shown in Fig 3.

Considering an SP as a starting point, ridges that are passed by while following the radius along the SP's main-axis, probably seem to flow away from the SP, namely parallel to the radial direction. While following the radius in the opposite direction, crossed ridges are flowing around the SP, i.e. perpendicular to radial direction. We believe this rule is an intrinsic nature of all SPs and is obviously invariant to the translation and rotation of fingerprint images. We visually name this pattern "T" shape model. Furthermore, how well a region matches this model can be utilized to measure the confidence of whether an SP locates in this region.

To apply this idea on a candidate SP acquired in Section 2, we first cut out a circle in PDF centered at it with certain radius, which is decided empirically in relation to the resolution of fingerprint images, as the region of interest. Then we divide the circle into homocentric sectors with a series of equally spaced radiuses. The average orientation of every sector, namely AOS, is calculated in a similar way to (2) except for the set of average.

Now we are about to detect the candidate SP's main-axis with AOS. First, we pick out lateral-axes whose corresponding sector fulfill (3), where k' is the opposite sector of k, $\begin{bmatrix} S_x(i) \\ S_y(i) \end{bmatrix}$ represent the *i*-th sector's orientation, θ_i is the *i*-th radial angle, λ is the weight of parallelism and *TH* is a predefined threshold. Each lateral-axis is probably a main-axis or almost parallel to it on the whole. Then we make use of the value of $f(k)$ to measure the confidence of each lateral-axis to be true main-axis or parallelism to it.

$$
\begin{aligned}
& f(k) = \frac{\lambda T_{//} + T_{\perp}}{\lambda + 1} - TH > 0 \\
& \text{where } T_{//} = |\cos(\frac{1}{2} tg^{-1}(S_y(k)/S_x(k)) - \theta_k)| \\
& T_{\perp} = |\sin(\frac{1}{2} tg^{-1}(S_y(k')/S_x(k')) - \theta_{k'})|
\end{aligned}
\tag{3}
$$

To avoid accidental peak values of $f(k)$, which are usually caused by noise, badly affecting the result of main-axis detection, we consider connecting sets that consist of consecutive lateral-axes. We select the set(s) with the largest one (for core) or three (for delta) of total summations of $f(k)$ as winner(s) for further process and use $C(\phi) = \sum_{k \in \phi} f(k)$ to represent the confidence of corresponding winning set ϕ. Next, in each ϕ we calculate a weighted average direction with all candidate lateral-axes as main-axis, where the weights are corresponding $f(k)$. Finally, one main-axis with corresponding confidence is obtained to a core and three main-axes to a delta. As a part of "T" shape model, we use the summation of all main-axes to measure the candidate SP's confidence. With this confidence, the determination of a candidate SP's authenticity can be realized by a simple thresholding. In practice, the thresholds for a core and for a delta are different.

4 Experimental Results

We choose dataset made up of the first prints (up to 100) of every finger in FVC2002 [11] DB1 set A to examine our method. When calculating lateral-axis, we consider certain number of consecutive sectors together as a BIG sector to enhance robustness. If the BIG sector satisfied (3), the sector in it with the max $f(k)$ is given as the result of main-axis. If more than two cores or two deltas are detected, the most confident two of them are output as the result.

Some of resulting images are presented in Fig 4. Note that several false acceptations are caused by similar patterns occurring close to the true SPs. It can be easily identified and cleaned by a post-process. Statistical data are given in Table 1 and Table 2. In fact, if ignoring the SPs located near the foreground edges, FRR and FAR of detection are 1.16% and 4.62% respectively. The accuracy of direction estimation is also satisfactory.

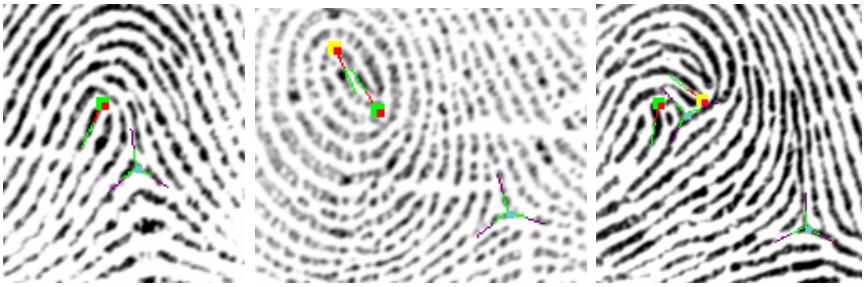


Fig. 4. Some resulting images: although the quality of the left two images is not very good and the right one is an unusual type of fingerprint, our approach works well on them

Table 1. Statistics of Detection

Number of Total SPs	177	100%
Correctly Detected	171	96.61%
Ignored for being close to background	4	2.26%
False Ignored	2	1.13%
False Accepted	8	4.52%

Table 2. Statistics of Direction Estimation

SPs Detected correctly	171	100%
Direction Error within 10 degrees	158	92.40%
10-20 degrees	9	5.26%
Over 20 degrees	4	2.34%

5 Conclusion

In this paper, a robust algorithm to locate the singularities and their directions has been proposed. An adaptive "T" shape model has been suggested which is abstracted from SP's intrinsic nature. Based on the "T" model, the proposed approach detects SPs and estimates their directions simultaneously. Experimental results show that the proposed model is quite effective in correctly assessing singularities and their directions.

Acknowledgements

This work described in this paper is supported by the National Key Basic Research Project of China under Grant No. 2004CB318005 and FANEDD China under Grant 200038.

References

1. Moayer, B., Fu, K.S.: A Syntactic Approach to Fingerprint Pattern Recognition, Pattern Recognition, Vol. 7, (1975), pp. 1-23
2. Nilsson, K., Bigun, J.: Localization of corresponding points in fingerprints by complex filtering, Pattern Recognition Letters 24 (13), (2003), pp. 2135-2144
3. Karu, K., Jain, A.K.: Fingerprint Classification, Pattern Recognition, Vol.18, No.3, (1996), pp. 389-404

4. Zhang, W., Wang, Y.: Core-based structure matching algorithm of fingerprint verification, Proceedings. 16th International Conference on Pattern Recognition, Vol: 1, (2002) pp. 70 -74
5. Hsieh, C., Lu, Z.Y., Li, T.C., Mei, K.C.: An effective method to extract fingerprint singular point, Proceedings of the Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region, 2000, Vol: 2, pp. 696 - 699
6. Tico, M., Kuosmanen, P.: A multi-resolution method for singular points detection in fingerprint images, Proceedings of the 1999 IEEE International Symposium on Circuits and Systems, vol.4, pp. 183 -186
7. Kawagoe, M., Tojo, A.: Fingerprint Pattern Classification, Pattern Recognition, Vol. 17, no. 3, (1984), pp. 295-303
8. Wang, F., Zou, X., Luo, Y., Hu, J.: A hierarchy approach for singular point detection in fingerprint images, Proceedings of First International Conference on Biometric Authentication, ICBA 2004, pp. 359-365, Hong Kong, China
9. Bazen, A. M., Gerez, S. H.: Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.7, (2002), pp. 905-919
10. Candela G.T., Grother P.J., Watson C.I., Wilkinson R.A., Wilson C.L.: PCASYS - A Pattern-level Classification Automation System for Fingerprints, Technical Report NISTIR 5647 & CD-ROM, April 1995
11. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: FVC2002: Second Fingerprint Verification Competition, Proceedings of 16th International Conference on Pattern Recognition, (2002)., Vol.3, 11-15, pp. 811 - 814

Face Detection Based on the Manifold

Ruiping Wang[1], Jie Chen[1], Shengye Yan[1], and Wen Gao[1,2]

[1] ICT-ISVISION Joint R&D Lab for Face Recognition, Institute of Computing Technology, Chinese of Academy of Sciences, Beijing, 100080, China
{rpwang,jchen,syyan,wgao}@jdl.ac.cn
[2] School of Computer Science and Technology, Harbin Institute of Technology, 150001, China

Abstract. Data collection for both training and testing a classifier is a tedious but essential step towards face detection and recognition. It is a piece of cake to collect more than hundreds of thousands of examples from web and digital camera nowadays. How to train a face detector based on the collected immense face database? This paper presents a manifold-based method to select a training set. That is to say we learn the manifold from the collected enormous face database and then subsample and interweave the training set by the estimated geodesic distance in the low-dimensional manifold embedding. By the resulting training set, we train an AdaBoost-based face detector. The trained detector is tested on the MIT+CMU frontal face test set. The experimental results show that the proposed method based on the manifold is efficient to train a classifier confronted with the huge database.

1 Introduction

Over the past ten years, face detection has been thoroughly studied in computer vision research for its wide potential applications, such as video surveillance, human computer interface, face recognition, and face image database management etc. Face detection is to determine whether there are any faces within a given image, and return the location and extent of each face in the image if one or more faces present [31]. Recently, the emphasis has been laid on data-driven learning-based techniques, such as [7, 13, 14, 15, 19, 20, 21, 22, 30]. All of these schemes can be found in the recent survey by Yang [31]. After the survey, one of the important progresses is the boosting-based method proposed by Viola [23] who uses the Haar features for the rapid object detection, and a lot of related works then followed [11, 12, 28].

The performance of these learning-based methods highly depends on the training set, and they suffer from a common problem of data collection for training. It is a piece of cake to collect more than hundreds of thousands of examples from web and digital camera nowadays. How to train a classifier based on the collected immense face database? This paper will give a solution.

In nature, how to train a classifier based on the collected immense face database is a problem of data mining. In this paper we will use the knowledge of the manifold to subsample a small subset from the collected huge face database and then interweave some big hole among the manifold embedding. Manifold can help us to transform the data to a low-dimensional space with little loss of information, which can enable us to visualize data, perform classification and cluster more efficiently. Recently, some representative techniques, including isometric feature mapping (ISOMAP) [25], local

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 208–218, 2005.
© Springer-Verlag Berlin Heidelberg 2005

linear embedding (LLE) [17], and Laplacian Eigenmap [1], have been proposed. The ISOMAP algorithm is intuitive, well understood and produces reasonable mapping results [9, 10, 29]. Also, it is supported theoretically [2, 5, 32], which has been developed by [3, 8, 16, 18, 24, 26, 27].

The main contributions of this paper are:

1. Subsample a small but efficient and representative subset from the collected huge face database based on the manifold embedding to train a classifier.
2. Interweave the subsampled manifold embedding to fill in the big holes to complete the training set furthermore.
3. The performance is instable to train a detector based on the random subsampling face set from a huge database. However, a detector trained based on the subsampled face set by the data manifold is not only stable but also can improve the detector performance.

The rest of this paper is organized as follows: After a review of ISOMAP, the proposed subsampling and interweaving method based on the manifold embedding is described in section 2. Experimental results are presented in section 3, followed by discussion in section 4.

2 Subsampling Based on ISOMAP

2.1 ISOMAP Algorithm

The goal of learning the data manifold is to show high-dimensional data in its intrinsic low-dimensional structures and use easily measured local metric information to learn the underlying global geometry of a data set [25]. Given a set of data points $X = \{x_1, \cdots, x_n\}$ in a high dimensional space, let $d_X(x_i, x_j)$ be the distance between x_i and x_j; let $y_i \in R^d$ be the co-ordinates corresponding to x_i and $Y = \{y_1, \cdots, y_n\}$. Let $d_Y(y_i, y_j)$ be the distance between y_i and y_j, which is an Euclidean distance in a *d*-dimensional Euclidean space *Y*. ISOMAP attempts to recover an isometric mapping from the co-ordinate space to the manifold. The neighborhood is necessary to be specified by ISOMAP. It can be *knn* -neighborhood, where x_i and x_j are neighbors if x_i (x_j) is one of the *k* nearest neighbors (*knn*) of x_j (x_i).

Let the vertex $v_i \in V$ corresponding to x_i; between v_i and v_j, an edge $e(i, j)$ exists iff x_i is a neighbor of x_j. The weight of $e(i, j)$ is simply $d_X(x_i, x_j)$. And then a weighted undirected neighborhood graph $G = (V, E)$ is constructed. Let $d_G(v_i, v_j)$ denote the length of the shortest path $sp(i, j)$ between v_i and v_j. The shortest paths can be found by the Dijkstra's algorithm, and the shortest paths can be stored efficiently by the predecessor matrix τ_{ij}, where $\tau_{ij} = k$ if v_k is immediately before v_j in $sp(i, j)$. We may call $d_G(v_i, v_j)$ "geodesic distance". That is to say after embedding the high-dimensional data manifold into low-dimensional structures, we can use straight lines in the embedding to approximate the true geodesic path.

2.2 Subsampling Algorithm

As discussed in [25], with the increase of the embedding dimensionality d, the difference between the Euclidean distance in the d-dimensional Euclidean space Y and the true geodesic path decreases. Therefore, after learning its manifold and embedding it in low-dimensionality, we can use the Euclidean distance in the d-dimensional Euclidean space Y to delete some examples from the huge database. And the remained examples can still keep the data's intrinsic geometric structure basically. By this means, we can get a small representative subset of the huge data.

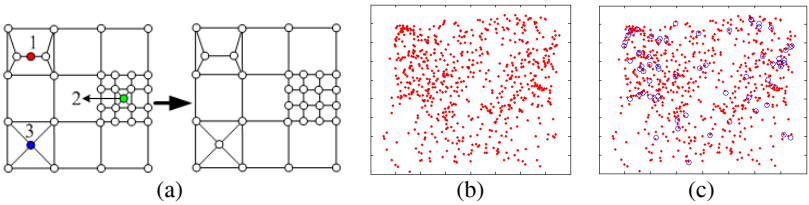


Fig. 1. Subsampling based on the manifold embedding. (a) The schematic of subsampling based on the estimated geodesic distance; (b) the manifold embedding of the 698 raw face images [25]; (c) the results of subsampling based on the estimated geodesic distance

The scheme is demonstrated in Fig. 1 (a). We sort all of the Euclidean distance (i.e., the estimated geodesic distance) between pairs of points y_i and y_j in the d-dimensional Euclidean space Y in increasing order. If the estimated geodesic distance between an example and its neighbor examples is smaller than a given threshold, it will be deleted while its neighbor examples will be reserved. For example, as shown in Fig. 1 (a), the data point 1 and 2 will be deleted when subsampling in the embedding while its neighbors are reserved. As to the data point 3, it is preserved since the estimated geodesic distance between it and its neighbors are bigger than the given threshold. From the figure of top right in Fig. 1 (a), the remained examples can still maintain the data's intrinsic geometric structure basically.

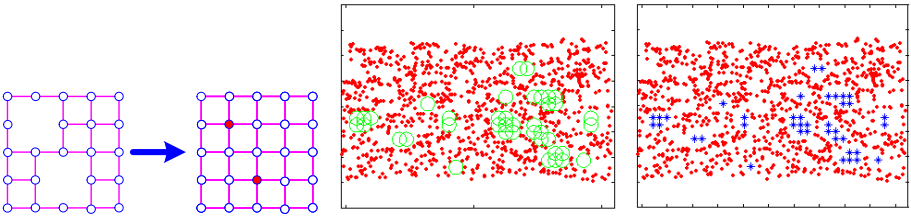
As demonstrated in Fig. 1 (b), it is a two-dimensional projection of 698 raw face images where the three-dimensional embedding of data's intrinsic geometric structure is learned by ISOMAP (K=6) [25]. Fig. 1 (c) is the results of subsampling where the data points (blue circle) are deleted and the remained data points are still in red dots.

If we want to subsample 90% examples from a whole set, what we need to do is to delete its 10% examples since their corresponding estimated geodesic distances to their neighbors are in the front of the sorted distance sequence.

2.3 Interweaving Algorithm

To complete the training set furthermore, we fill in the hole among the manifold embedding after the subsampling. The basic idea is as shown in Fig.2. The solid circle points in Fig 2 (a) are the filled examples. How to search these holes in the manifold embedding? In our case, we calculate the ***median*** of all of the Euclidean distance between pairs of points y_i and its nearest neighbor y_j in the d-dimensional Euclidean space Y. The median is used as the radius of the searching ring as demonstrated in

Fig.2 (b). Moving the searching ring along the embedding, we can get several holes wanted. As shown in Fig 2 (c), the centers of these holes are the places where we generate virtual samples.



(a) (b) (c)

Fig. 2. Interweaving based on the manifold embedding. (a) The schematic of interweaving based on the estimated geodesic distance; (b) the searched holes among the manifold embedding of the Swiss roll [25]; (c) the results of interweaving based on the estimated geodesic distance

Having found the holes in the embedding, the next step is to generate the virtual examples to fill in these holes. In our case, after we have learned the manifold embedding about the face example set, we fill in the holes with some virtual face examples. The basic idea is as following:

1. Perform PCA to the face database, then a coefficient vector $\vec{e}_i$ is computed for each sample f_i in our face database $\{f_1, f_2, ..., f_n\}$, where n is the number of the face examples;
2. Get the K neighbors $\{f_{n1}, f_{n2}, ..., f_{nK}\}$ of a virtual example VE_p $(p = 1,...,m)$ and the Euclidean distance $d_Y(VE_p, f_i)(i = 1,...,K)$ between it and its neighbors in the d-dimensional Euclidean space Y, where m is the number of virtual examples;
3. Calculate the weight $\omega_{pi} = \omega_Y(VE_p, f_i) = 1/d_Y(VE_p, f_i)$, and are normalized: $\tilde{\omega}_{pi} = \dfrac{\omega_{pi}}{\sum_{i=1}^{K} \omega_{pi}}$;
4. The coefficient vector $\vec{e}_p$ of the virtual example VE_p is generated by the linear combination of the corresponding coefficient vectors $\{\vec{e}_{n1}, \vec{e}_{n2}, ..., \vec{e}_{nK}\}$: $\vec{e}_p = \tilde{\omega}_{p1}\vec{e}_{n1} + \tilde{\omega}_{p2}\vec{e}_{n2} + ... + \tilde{\omega}_{pK}\vec{e}_{nK}$;
5. Reconstruct the virtual example VE_p with the coefficient vector $\vec{e}_p$.

Some synthetic virtual samples are shown in table 1, while some synthetic virtual samples and its neighbors are shown in table 2. From the table 1, one can conclude that these synthetic virtual samples look like the real faces very much. From the table 2, one can conclude that the virtual face in the white complexion is constructed by several faces also in the white complexion, while the virtual face in the black is constructed by several faces also in the black and the female by several females.

3 Experiments

3.1 Detector Based on the MIT Face Database

The data set is consisted of a training set of 6,977 images (2,429 faces and 4,548 non-faces) and the test set is composed of 24,045 images (472 faces and 23,573 non-faces). All of these images are 19×19 grayscale and they are available on the CBCL webpage [33].

Table 1. Original samples vs. synthetic virtual samples

Original samples	
Synthetic virtual samples	

Table 2. Synthetic virtual sample and its conresponding original samples

Synthetic virtual sample:  (in the white complexion)						
Original samples						
weight	0.1172	0.1109	0.0883	0.0816	0.0803	0.0800
Original samples						
weight	0.0780	0.0776	0.0758	0.0734	0.0690	0.0675
Synthetic virtual sample:  (in the black complexion)						
Original samples						
weight	0.1179	0.1132	0.0951	0.0850	0.0838	0.0770
Original samples						
weight	0.0760	0.0716	0.0712	0.0703	0.0692	0.0690
Synthetic virtual sample:  (a female)						
Original samples						
weight	0.0887	0.0874	0.0874	0.0847	0.0843	0.0828
Original samples						
weight	0.0828	0.0825	0.0811	0.0803	0.0796	0.0779

We let K=6 when ISOMAP learns the manifold of 2,429 faces in MIT database. The intrinsic dimensionality of the database can be estimated by the inflexion of the curve [25]. As to the MIT face database, its residual variance decreases while the dimensionality d increases as shown in Fig. 3. We can let d=10 for the MIT database. However, the residual variance is still 0.097. It is because the face examples in MIT database are too complex, such as different people and variations in poses, facial expressions, lighting conditions.

Note that all of these examples are performed by histogram equalization before learning the manifold. It is because all examples to train a classifier are needed histogram equalization which maps the intensity values to expand the range of intensities.

In order to study the relationship between the distribution of the training set and the trained detector, we subsample the MIT face database by 90%, 80% and 70% (named as ISO90, ISO80, ISO70 later) as discussed in section 2.2. Subsampling by 90% is to

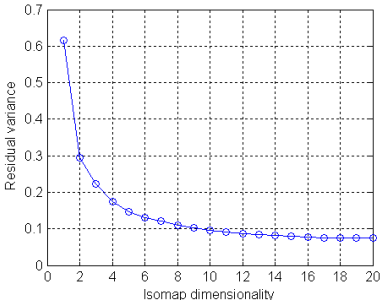


Fig. 3. The residual variance of ISOMAP on the MIT face database

say we reserve 90% examples of the database and the same meaning of 80% and 70%. Note that ISO70 is a subset of ISO80 and ISO80 is a subset of ISO90 in fact.

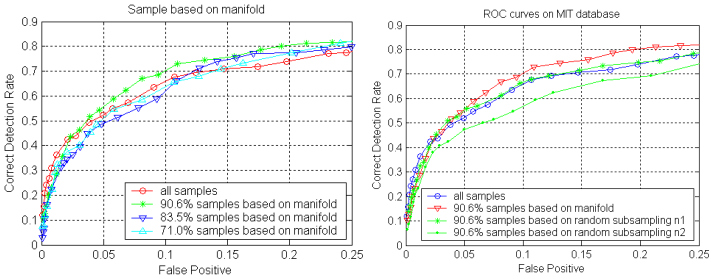
The three subsampled face sets together with the non-face are used to train three classifiers based on the AdaBoost as demonstrated in [23]. And then they are tested on the test set of MIT database. The ROC curves of these three classifiers are shown in Fig. 4 (a). From these ROC curves, one can conclude that the detector trained by ISO90 is the best of all and improves the performance of the detector distinctly compared with the detector even by the entire face examples in MIT database. Even the detector trained on ISO70 is a little better than the detector trained on the entire examples. Some possible reasons: the first one is the examples of ISO90 distribute evenly in the example space and has no example congregate compared with the whole set; the second is that the outliers in the whole set deteriorate its performance which has been discarded during the manifold learning [25] (During the ISOMAP learning, we get 30 outliers.).

However, random subsampling from the MIT database is not so lucky. We choose four subsets randomly-subsampled from the MIT database and each subset has the same number of examples as in ISO90. After trained on these four sets, they are also tested on the same test set. The ROC curves are shown in Fig. 4 (b). In this figure, we plot the resulting ROC curves of detectors on the whole set, ISO90, and two randomly chosen subsets. Herein, the curve "*90.6% examples based on the random subsampling n1*" and the curve "*90.6% examples based on the random subsampling n2*" represent the best and the worst results of these four random sampling cases. From these ROC curves, one can conclude that the detector based on ISO90 is still the best of all and the results based on random subsampling is much instable. We also think that the evenly-distributed examples and no outliers contribute to this kind of results.

After the subsampling, we interweave the manifold embedding as discussed in section 2.3. As shown in Fig. 5 (a), we add the different number of virtual examples in the set ISO90. There are 100 or 500 examples are added into ISO90, respectively. One can conclude that a few numbers of added examples is valuable for training a detector. When the number is up to 500, it will be deteriorate. As shown in Fig. 5 (b), we change the radius of the searching ring. The first 100 examples are searched by the radius equal to the ***median***, while the second 100 examples are searched by the radius equal to the 1.1×***median*** and the third 100 equal to the 1.2×***median***. One can conclude when the radius of the searching ring is equal to the ***median***, the added 100 examples is most valuable for training a detector.

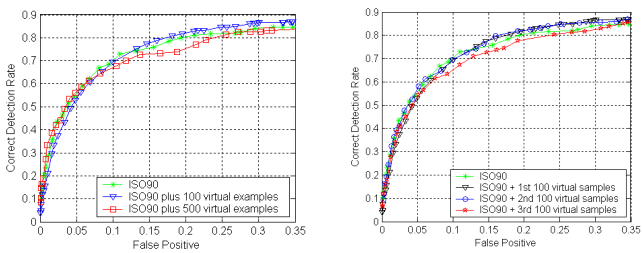
3.2 Detector Based on the Huge Database

To compare the performance difference on different training sets further, we also use another three different face training sets. The face-image database consists of 10,000 faces (collected form web, video and digital camera), which cover wide variations in poses, facial expressions and also in lighting conditions. To make the detection method robust to affine transform, the images are often rotated, translated and scaled [6]. After these preprocessing, we get 90,000 face images which constitute the whole set. The first group is composed of 14,000 face images which are sampled by the ISOMAP (called ISO14000, here). The second and third group are also composed of 14,000 face images which are random subsampling examples from the whole set (named Rand1-14000 and Rand2-14000, respectively).



(a) (b)

Fig. 4. The ROC curves on the MIT test set. (a) Using the subsampling face example sets based on the manifold embedding and the whole set as training set for a fixed classifier. (b) Using the subsampling face example sets based on the manifold embedding, two random sampling sets and the whole set as training set for a fixed classifier



(a) (b)

Fig. 5. The ROC curves on the MIT test set using the interweaving face example sets based on the manifold embedding. (a) Add the different number of virtual examples in the training set ISO90. (b) Change the radius of the searching ring

It is hard to learn the manifold from 90,000 examples by the ISOMAP because it needs to calculate the eigenvalues and eigenvectors of a $90{,}000 \times 90{,}000$ matrix. In order to avoid this problem, as demonstrated in Fig. 6, we divide the whole set into 30 subsets and each subset has 3,000 examples. We get 1,000 examples by the proposed method from each subset and then incorporate every three subsampled sets into one subset. We will have 10 subsets with the total 30,000 examples. With the same proce-

dure, we can get 1,400 examples by the proposed method from each incorporated subset and then incorporate all subsampled examples into one set with the total 14,000 examples.

The non-face class is initially represented by 14,000 non-face images. Each single classifier is then trained using a bootstrap approach similar to that described in [22] to increase the number of negative examples in the non-face set. The bootstrap is carried out several times on a set of 13,272 images containing no faces.

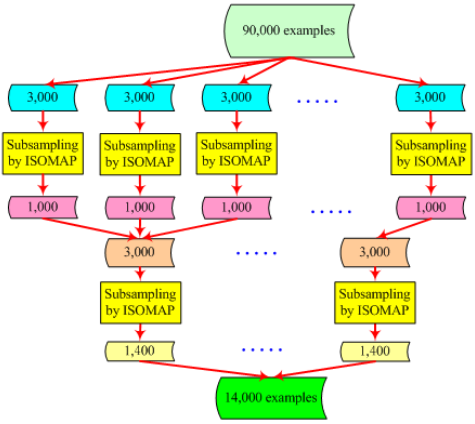
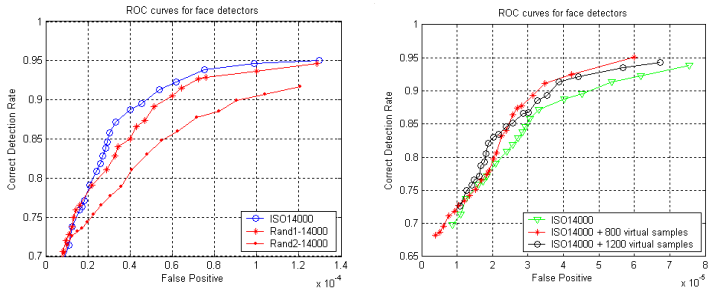


Fig. 6. Subsampling procedure by ISOMAP to get 14,000 examples from 90,000 examples

The resulting detectors, trained on the three different sets, are evaluated on the MIT+CMU frontal face test set which consists of 130 images showing 507 upright faces [19]. The detection performances on this set are compared in Fig. 7 (a). From these ROC curves one can conclude that the detector based on ISO14000 is the best of all and the results based on random subsampling is also much instable. During the ISOMAP learning, we get 838 outliers. We think that the evenly-distributed examples and no outliers contribute this kind of results, again.



(a) (b)

Fig. 7. The ROC curves for the trained detectors. (a) Train detector based on the sampled training set by the ISOMAP and the random subsampling training set. (b) Train detector by adding the virtual samples

Based on the subsampled training set ISO14000, we add some virtual examples (800, 1200 respectively) by the proposed method. As shown in Fig. 7 (b), the detectors by ISO14000 together with adding virtual example outperform the detector only by ISO14000. Some results of our trained detector based on ISO14000 + 800 examples are shown in Fig.8.



Fig. 8. Some results of our trained detector

4 Conclusion

In this paper, we present a novel manifold-based method to subsample a small but efficient and representative training subset from the collected enormous face database. After learning the manifold from the collected face database, we subsample the training set by the estimated geodesic distance in the manifold embedding and then interweave the big holes in the embedding. An AdaBoost-based face detector is trained on the resulting training set in the low-dimensional manifold embedding, and then we test it on the MIT+CMU frontal face test set. Compared with the AdaBoost-based face detector using random subsampling examples, the detector trained by the proposed method is more stable and achieve better face detection performance. We conclude that the evenly-distributed examples, due to the subsampling training set based on the manifold embedding, and no outliers, discarded during the manifold learning, contribute to the improved performance. The added virtual examples can improve the performance of the detector further.

Acknowledgement

This research is partially sponsored by Natural Science Foundation of China under contract No. 60473043, National Hi-Tech Program of China (No. 2003AA142140), and ISVISION Technologies Co., Ltd.

References

1. M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Inform. Proc. Systems 14*, pp.585-591. MIT Press, 2002.

2. M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Technical report, Stanford University*, 2000.
3. M. Brand. Charting a manifold. In *Advances in Neural Information Proc. Systems 15*, pp. 961-968. MIT Press, 2003.
4. J. Chen, X. Chen and W. Gao. Expand Training Set for Face Detection by GA Resampling. *The 6th IEEE Intern. Conf. FG2004*. pp. 73-79. 2004.
5. D. L. Donoho and C. Grimes. When does ISOMAP recover natural parameterization of families of articulated images? *Technical Report 2002-27,, Stanford Universit*y, 2002.
6. B. Heisele, T. Poggio, and M. Pontil. Face Detection in Still Gray Images. *CBCL Paper #187*. MIT, Cambridge, MA, 2000.
7. R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Anal. Machine Intell.*, pp.696–706, 2002.
8. D. R. Hundley and M. J. Kirby. Estimation of topological dimension. In *Proc. SIAM International Conference on Data Mining*, 2003. http://www.siam.org/meetings/sdm03/proceedings/ sdm03_18.pdf
9. O. C. Jenkins and M. J Mataric. Automated derivation of behavior vocabularies for autonomous humanoid motion. In *Proc. of the Second Int'l Joint Conference on Autonomous Agents and Multiagent Systems*, Melbourne, Australia, July 2003.
10. M. H. Law, N. Zhang, A. K. Jain. Nonlinear Manifold Learning for Data Stream. In *Proc. of SIAM Data Mining*, pp. 33-44, Florida, 2004.
11. S. Z. Li, L. Zhu, Z.Q. Zhang, A. Blake, H. J. Zhang, and H. Shum. Statistical Learning of Multi-View Face Detection. *In Proc. of the 7th ECCV*. 2002.
12. C. Liu, H. Y. Shum. Kullback-Leibler Boosting. *Proceedings of the 2003 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'03)*. 2003.
13. C. J. Liu. A Bayesian Discriminating Features Method for Face Detection, *IEEE Trans. Pattern Anal. and Machine Intel.*, pp. 725-740. 2003.
14. E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. IEEE Conf. on CVPR*, pp. 130–136. 1997,
15. C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proc. 6th Int. Conf. Computer Vision*, pp.555–562. 1998,
16. K. Pettis, T. Bailey, A. K. Jain, and R. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Trans. of Patt. Anal. and Machine Intel.*, pp.25-36, 1979.
17. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290: pp.2323-2326, 2000.
18. S. T. Roweis, L. K. Saul, and G. E. Hinton. Global coordination of local linear models. In *Advances in Neural Information Processing Systems 14*, pp. 889-896. MIT Press, 2002.
19. H. A. Rowley, S. Baluja, and T. Kanade. Neural Network-Based Face Detection. *IEEE Tr. Pattern Analysis and Machine Intel.* pp. 23-38. 1998.
20. H. A. Rowley, S. Baluja, and T. Kanade. Rotation Invariant Neural Network-Based Face Detection. *Conf. Computer Vision and Pattern Rec.*, pp. 38-44. 1998.
21. H. Schneiderman and T. Kanade. A Statistical Method for 3D Object Detection Applied to Faces. *Comp. Vision and Pattern Recog.*, pp. 746-751. 2000.
22. K. K. Sung, and T. Poggio. Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. on PAM*. pp. 39-51. 1998.
23. P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *Conf. Comp. Vision and Pattern Recog.*, pp. 511-518. 2001.
24. Y. W. Teh and S. T. Roweis. Automatic alignment of local representations. In *Advances in Neural Information Processing Systems 15*, pp. 841-848. MIT Press, 2003.
25. B. J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, Volume 290, pp.2319-2323, 2000
26. J.J. Verbeek, N. Vlassis, and B. Krose. Coordinating principal component analyzers. In *Proc. of International Conf. on Artificial Neural Networks*, pp. 914-919, Spain, 2002.

27. J.J. Verbeek, N. Vlassis, and B. Krose. Fast nonlinear dimensionality reduction with topology preserving networks. In *Proc. 10th European Symposium on Artificial Neural Networks*, pp.193-198, 2002.
28. R. Xiao, M. J. Li, H. J. Zhang. Robust Multipose Face Detection in Images, *IEEE Trans on Circuits and Systems for Video Technology*, Vol.14, No.1 pp. 31-41. 2004,
29. M.-H. Yang. Face recognition using extended ISOMAP. In *ICIP*, pp.117-120, 2002.
30. M. H. Yang, D. Roth, and N. Ahuja. A SNoW-Based Face Detector. *Advances in Neural Information Processing Systems 12*, MIT Press, pp. 855-861. 2000.
31. M. H. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Tr. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 34-58. 2002.
32. H. Zha and Z. Zhang. Isometric embedding and continuum ISOMAP. In *ICML*, 2003. http://www.hpl.hp.com/conferences/icml2003/ papers / 8.pdf
33. http://www.ai.mit.edu/projects/cbcl/software- dataset / index.html

Local and Global Feature Extraction for Face Recognition

Yongjin Lee[1], Kyunghee Lee[2], and Sungbum Pan[3]

[1] Biometrics Technology Research Team
Electronics and Telecommunications Research Institute
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
solarone@etri.re.kr
[2] Department of Electrical Engineering
The University of Suwon, Korea
khlee@suwon.ac.kr
[3] Division of Information and Control Measurement Engineering
Chosun University, Korea
sbpan@chosun.ac.kr

Abstract. This paper proposes a new feature extraction method for face recognition. The proposed method is based on Local Feature Analysis (LFA). LFA is known as a local method for face recognition since it constructs kernels which detect local structures of a face. It, however, addresses only image representation and has a problem for recognition. In the paper, we point out the problem of LFA and propose a new feature extraction method by modifying LFA. Our method consists of three steps. After extracting local structures using LFA, we construct a subset of kernels, which is efficient for recognition. Then we combine the local structures to represent them in a more compact form. This results in new bases which have compromised aspects between kernels of LFA and eigenfaces for face images. Through face recognition experiments, we verify the efficiency of our method.

1 Introduction

For face recognition, feature extraction is required to represent high dimensional image data into low dimensional feature vectors. In general, there are two approaches to feature extraction, a global and a local method. The most famous one among the global methods is Principal Component Analysis(PCA) [1]. PCA for face recognition is known as a global method since it extracts face features using the bases which describe a whole face. The bases are eigenvectors of the covariance matrix of face images and thought as face models, called *eigenfaces*. By projecting a face image onto the eigenfaces, the linear combination weights for eigenfaces are calculated. These weights are used as representations of a face. Eigenface method is simple and fast, but there are limitations in recognition under illumination and pose variation.

On the contrary, it is known that local methods are robust to such variations. One of the methods, we consider in this paper, is Local Feature Analy-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 219–228, 2005.
© Springer-Verlag Berlin Heidelberg 2005

sis(LFA) [2]. LFA is referred to as a local method because it constructs a set of kernels which detect local structures such as nose, eye, jaw-line, and cheek-bone. It, however, addresses only image representation and has a problem to be used for recognition. In this paper, we point out the problem of LFA and present a new feature extraction method for face recognition by modifying LFA. This results in new bases for feature extraction which have compromised aspects between kernels of LFA and eigenfaces.

The rest of the paper is organized as follows. Fisher Score and LFA will be briefly reviewed in Sec. 2. In Sec. 3, we propose our method. In Sec. 4. experimental results are given to verify the efficiency of our method. And conclusions are drawn in Sec. 5.

2 Backgrounds

In this section, we will give brief overview of LFA and Fisher Score, which are main ingredients of our method.

2.1 Local Feature Analysis

LFA is a topographic representation based on second-order statistics [2] [3]. The kernels of LFA are derived by enforcing topology into eigenvectors of PCA. Then selection, or sparsification, step is used to reduce and decorrelate the outputs.

Let x and y in a parenthesis indicate indexes of elements of vectors and matrices and suppose that we have N eigenvectors, $\boldsymbol{\Psi}_r$, and a set of V-dimensional input images, $\{\boldsymbol{\phi}_t\}_{t=1}^{n}$. Then a kernel is defined as follows

$$\boldsymbol{K}(x,y) = \sum_{r=1}^{N} \boldsymbol{\Psi}_r(x) \frac{1}{\sqrt{\lambda_r}} \boldsymbol{\Psi}_r(y) \tag{1}$$

where λ_r and $\boldsymbol{\Psi}_r$ denotes rth eigenvalue and eigenvector of covariance matrix of face images. And the output for tth input image, $\boldsymbol{O}_t$, and correlation of the outputs, $\boldsymbol{P}$, are written as

$$\boldsymbol{O}_t(x) = \sum_{y=1}^{V} \boldsymbol{K}(x,y)\,\boldsymbol{\phi}_t(y) = \sum_{r=1}^{N} \frac{A_t(r)}{\sqrt{\lambda_r}} \boldsymbol{\Psi}_r(x) \tag{2}$$

$$\boldsymbol{P}(x,y) = \langle \boldsymbol{O}_t(x)\,\boldsymbol{O}_t(y) \rangle = \sum_{r=1}^{N} \boldsymbol{\Psi}_r(x)\,\boldsymbol{\Psi}_r(y) \tag{3}$$

where $A_t(r) = \sum_{y=1}^{V} \boldsymbol{\Psi}_r(y)\,\boldsymbol{\phi}_t(y)$.

In a matrix form, a set of kernels, the output matrix, and the covariance matrix are written as

$$\boldsymbol{K} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^T \tag{4}$$

$$\boldsymbol{O} = \boldsymbol{K}^T\boldsymbol{\Phi} \tag{5}$$

$$\boldsymbol{P} = \boldsymbol{\Psi}\boldsymbol{\Psi}^T \tag{6}$$

where $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1 \ldots \boldsymbol{\Psi}_N]$, $\boldsymbol{\Lambda} = diag\left(\frac{1}{\sqrt{\lambda_r}}\right)$, and $\boldsymbol{\Phi} = [\phi_1 \cdots \phi_n]$ ($diag\,(d_i)$ denotes the matrix with the elements $d_1, d_2, \ldots$ on the leading diagonal and zeros elsewhere). The rows or columns of $\boldsymbol{K}$ contain kernels since $\boldsymbol{K}$ is symmetric. For clarity, we consider that the columns of $\boldsymbol{K}$ contain kernels. The kernels are bases like eigenfaces. The difference is that the kernels of LFA have spatially local properties(see Fig. 3), and are *topographic* in the sense that they are indexed by spatial location.

Note that the number of kernels constructed by LFA is the same as the input dimension, V. The dimension of the outputs is reduced by choosing a subset of kernels, $\mathcal{M}$. $\mathcal{M}$ is constructed by adding iteratively the kernel corresponding to the output with the largest mean reconstruction error across all of the images [3].

At each step, the point added to $\mathcal{M}$ is chosen as the kernel corresponding to location, x,

$$\arg\max_{x} \langle \| \boldsymbol{O}_t(x) - \boldsymbol{O}_t^{rec}(x) \|^2 \rangle \tag{7}$$

where $\boldsymbol{O}_t^{rec}(x)$ is the reconstruction of the output, $\boldsymbol{O}_t(x)$. The reconstruction of tth output is

$$\boldsymbol{O}_t^{rec}(x) = \sum_{m=1}^{|\mathcal{M}|} \boldsymbol{C}(m,x)\, \boldsymbol{O}_t(y_m) \tag{8}$$

where $\boldsymbol{C}(m,x)$ is the reconstruction coefficient and $y_m \in \mathcal{M}$.
For all images, the reconstruction is written in a matrix form as follows

$$\boldsymbol{O}^{rec} = \boldsymbol{C}^T \boldsymbol{O}(\mathcal{M},:). \tag{9}$$

$\boldsymbol{O}(\mathcal{M},:)$ denotes the subset of $\boldsymbol{O}$ corresponding to the points in $\mathcal{M}$ for all n images. And, $\boldsymbol{C}$ is calculated from:

$$\boldsymbol{C} = \boldsymbol{P}(\mathcal{M},\mathcal{M})^{-1} \boldsymbol{P}(\mathcal{M},:). \tag{10}$$

2.2 Fisher Score

Fisher Score is a measure of discriminant power. It estimates how well classes are separated from each other by the ratio of the *between-class scatter* and the *within-class scatter* [4].

For the c class problem, suppose that we have a set of n d-dimensional samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and n_i in the subset $\mathcal{X}_i$ labelled c_i. Then, *between-class scatter*, $\boldsymbol{S}_B$, and *within-class scatter*, $\boldsymbol{S}_W$, are defined as

$$\boldsymbol{S}_W = \sum_{i=1}^{c} \sum_{\boldsymbol{x} \in \mathcal{X}_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^T \tag{11}$$

$$\boldsymbol{S}_B = \sum_{i=1}^{c} n_i (\boldsymbol{m} - \boldsymbol{m}_i)(\boldsymbol{m} - \boldsymbol{m}_i)^T \tag{12}$$

where

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x}, \ \boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{c} n_i \boldsymbol{m}_i \ .$$

A simple scalar measure of scatter is the determinant of the scatter matrix. Using this measure, Fisher Score is defined as

$$\mathcal{J} = \frac{|\boldsymbol{S}_B|}{|\boldsymbol{S}_W|} \tag{13}$$

where $|\cdot|$ denotes determinant.

3 Local and Global Feature Extraction

In this section, we address the problems of kernel selection of LFA. Then we propose a new feature extraction method based on LFA. Our method consists of three steps: construction, selection, and combination of local structures. The last step causes reduction of dimensions in outputs in a more compact form.

3.1 LFA

As mentioned in previous section, LFA chooses a set of kernels whose outputs produced the biggest reconstruction error in the sense of minimum reconstruction error. Although mean reconstruction error is a useful criterion for representing data, there is no reason to assume that it must be useful for discriminating between data in different classes. This problem can be easily verified through an experiment with face images which include some background.

An example which addressed the problem is shown in Fig. 1. We used the first 120 eigenvectors to construct a set of 120 kernels. Dots are placed in their locations on the mean of face images and the order of the first 25 are written. It can be seen that kernels which belong to the outside of the face are also selected. It aims at reducing reconstruction error on a whole picture not on a face. Note that it is difficult to select kernels more than the number of eigenvectors used for kernel construction since the algorithm involves matrix inversion and may cause rank deficiency(see Eq. 10).

3.2 Proposed Method

After constructing kernels using LFA, we calculated their Fisher Scores(Eq. 13) using the outputs. In Fig. 2, the score values are displayed on the location of the corresponding kernels. It shows that kernels belonging to the meaningful areas for recognition, such as eyebrow, nose, cheekbone, and jaw-line, received higher scores than the rest. This verifies the usefulness of Fisher Score for kernel selection. But kernel selection by Fisher Score does not regard the redundancy between outputs of kernels. To cover the meaningful area of a face, a large number

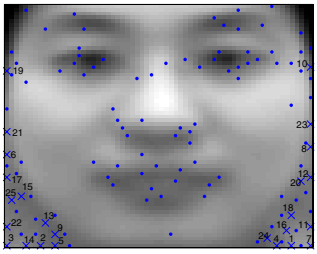


Fig. 1. The locations of 120 kernels selected according to minimum reconstruction error

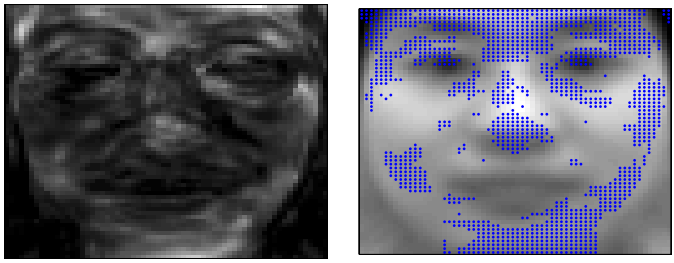


Fig. 2. Left: The Fisher Scores of the kernels. White color is corresponding with larger magnitude. Right: The locations of the first 1500 kernels selected according to Fisher Score

of kernels are required. However, this problem can be solved by overlaying the set of local structures onto a single sheet, *composite template*[1].

Suppose that we choose a subset, $\mathcal{I}$, of column vectors(i.e., kernels) from the matrix $\boldsymbol{K}$ in Eq. 4 according to Fisher Score and their elements are denoted by $\boldsymbol{K}(:, x_i)$. We compose a composite template, $\boldsymbol{g}$, by linear combination of local structures as follows.

$$\boldsymbol{g} = \sum_{i=1}^{|\mathcal{I}|} w_i \boldsymbol{K}(:, x_i) \tag{14}$$

where w_i is a linear combination weight and $x_i \in \mathcal{I}$.

However, we do not want to lose information by combining them. We, thus, select the combination weights, w_i, so as to maximize the entropy of the outputs of $\boldsymbol{g}$ [5]. For simplicity, we assume Gaussian density for the outputs. Other criterions and density assumptions can be used for different combination strategies.

Let s_t be the final output for the tth input, $\boldsymbol{\phi}_t$,

$$s_t = \boldsymbol{g}^T \boldsymbol{\phi}_t = \sum_{i=1}^{|\mathcal{I}|} w_i \boldsymbol{O}_t(x_i). \tag{15}$$

[1] We call our method and derived bases *composite template*, because the bases of our method consist of a set of local structures

Without loss of generality, we assume zero mean for s.

$$p(s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{s^2}{2\sigma^2}\right] \tag{16}$$

where σ^2 is the variance of s. Then, the entropy of density for s is written as

$$\begin{aligned} H(p(s)) &= -\int p(s) \log p(s)\, ds \\ &= \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log 2\pi + \frac{1}{2} \end{aligned} \tag{17}$$

Since the last two terms are constants, we only concern the variance, σ^2. Maximization of Eq. 17 is equivalent to the maximization of σ^2. It can be rewritten as

$$\begin{aligned} \sigma^2 &= \frac{1}{n}\sum_{t=1}^{n} s_t^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\left[\sum_{i=1}^{|\mathcal{I}|} w_i \boldsymbol{O}_t(x_i)\right]^2 \\ &= \sum_{i=1}^{|\mathcal{I}|}\sum_{j=1}^{|\mathcal{I}|} w_i \boldsymbol{P}(x_i, x_j)\, w_j \\ &= \boldsymbol{w}^T \boldsymbol{P}(\mathcal{I}, \mathcal{I})\, \boldsymbol{w} \end{aligned} \tag{18}$$

where $\boldsymbol{w}^T = [w_1 \ldots w_{|\mathcal{I}|}]$.

The linear combination weights, $\boldsymbol{w}$, which maximizes the above equation can be easily estimated, if we constrain $\boldsymbol{w}^T\boldsymbol{w} = 1$, since $\boldsymbol{P}(\mathcal{I}, \mathcal{I})$ is symmetric. In this case, it is equivalent to carry out PCA for the outputs from $\mathcal{I}$. This makes it clear of how many composite templates, $\boldsymbol{g}$, should be constructed and how we construct them in the framework of information theory. The maximum number(also to be believed as optimal) of composite templates is N since a set of kernels, $\boldsymbol{K}$, the outputs, $\boldsymbol{O}$, and the covariance matrix, $\boldsymbol{P}(\mathcal{I}, \mathcal{I})$, are all based on the N eigenvectors, i.e., their rank is N. By constructing N composite templates, we can keep all information. Practically, N is much smaller than $|\mathcal{I}|$.

Remark. The Fig. 3 shows eigenfaces, kernels, fisherfaces [6] and composite templates. Eigenfaces and kernels are at the extreme sides, one is global and the other is local. Our composite templates, which are constructed by combining kernels, show intermediate aspects between kernels(local) and eigenfaces(global). Fisherfaces, which also use Fisher Score, are somehow similar to composite templates. It can be, however, thought that fisherfaces are constructed by combining the global structures(i.e., eigenfaces) since Fisher Liner Discriminant(FLD) is applied after the dimension of the images is reduced using PCA.

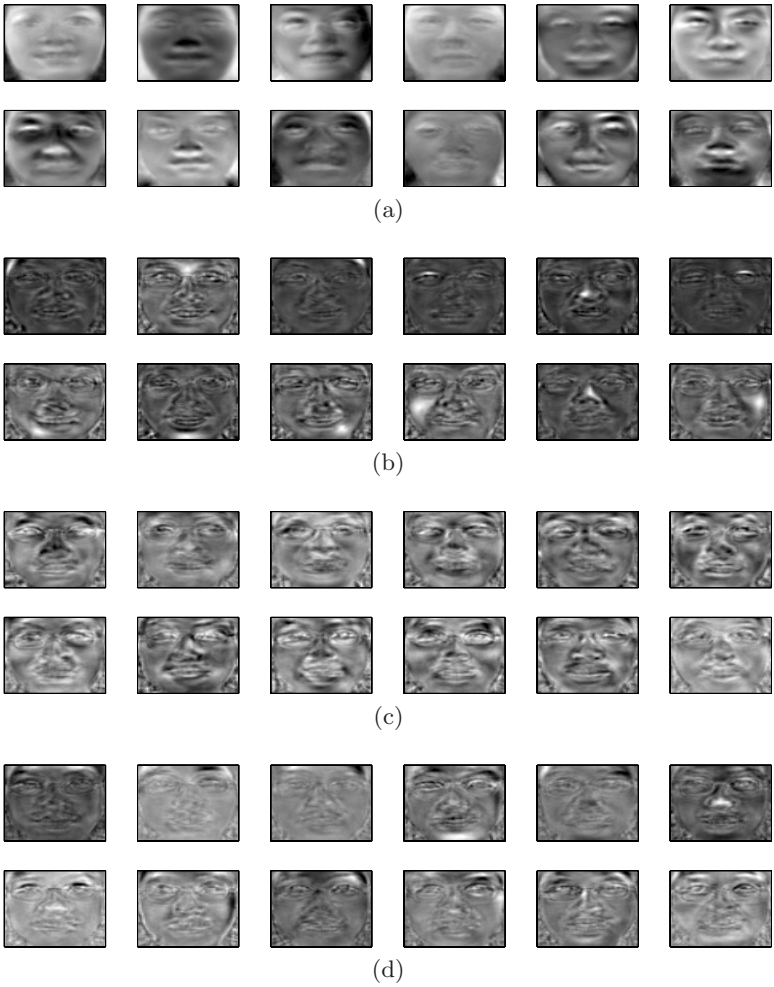


Fig. 3. (a) The first 12 eigenfaces. (b)The kernels of LFA selected manually. (c) The first 12 fisherfaces. (d)The first 12 composite templates

4 Experiments

In this section, we verify efficiency of our suggested method through face recognition experiments. The data used in the experiments is consisted of 55 people from our research institute. Example faces are shown in Fig. 4. The 20 pictures were taken for each person in a normal office. The images from 20 people were used to construct bases for feature extraction and the images from the rest were used for training(gallery) and test(probe). 10 images per person were used for training and test respectively. The size of image is 64×64. Using Euclidean distance, the face recognition experiments were conducted.



Fig. 4. Example faces in our database

We carried out experiments for each feature extraction method: (a) eigenface, (b) LFA, (c) fisherface, and (d) composite template. 120 eigenvectors were used to construct a set of kernels for both LFA and composite template. Then 120 and 1500 kernels were chosen by their own methods. The positions of the selected kernels can be seen in Fig. 1 and Fig. 2. In fisherface, the dimension of the images was reduced into 120 using PCA, and then FLD was applied. The bases constructed by each method are shown in Fig. 3.

The performance for each method is shown in Table 1 and Fig. 5. Fig. 5 shows recognition rate as increasing the number of features. For the rank 1, the best recognition rates of eigenface, LFA, fisherface, and composite template are 61.43%, 62.57%, 79.14%, and 86.57%, respectively. In LFA and fisherface, we increased the number of the eigenvectors, but could not obtain a better result. Additionally, we conducted the experiments that FLD was applied to the outputs of all kernels without feature selection. But almost the same results as fisherface were obtained. The explicit exclusion of the unnecessary features through the selection step may be one of the reasons for success of the proposed method.

To achieve the best recognition rate in each method, 144, 120, 17, and 84 features were needed. Note that the number of fisherfaces is bounded on the number of classes(people), which are used in basis construction. Although LFA gave a poor result, our method showed the best performance among the methods. It can be also seen that the size of kernels chosen by Fisher Score was reduced effectively and performance of composite template was in stable condition after 120 features as discussed in the previous section. The number is the same as the size of the eigenvector set used for kernel construction.

Table 1. Accuracy(%) which test images are matched within rank 1, 2, 3, 4, 5, 10 and 20 using the number of features in the first row. The numbers in the first column are the minimum number of features with the best performance for the rank 1. (a) eigenface, (b) LFA, (c) fisherface, and (d) composite template

Rank	(a): 144	(b): 120	(c): 17	(d): 84
1	61.43	62.57	79.14	86.57
2	73.43	67.71	86.86	89.14
3	77.43	71.43	89.71	90.57
4	80.86	74.00	92.00	93.43
5	83.71	77.71	93.71	95.14
10	92.29	86.57	98.57	98.86
20	98.86	97.43	99.71	99.71

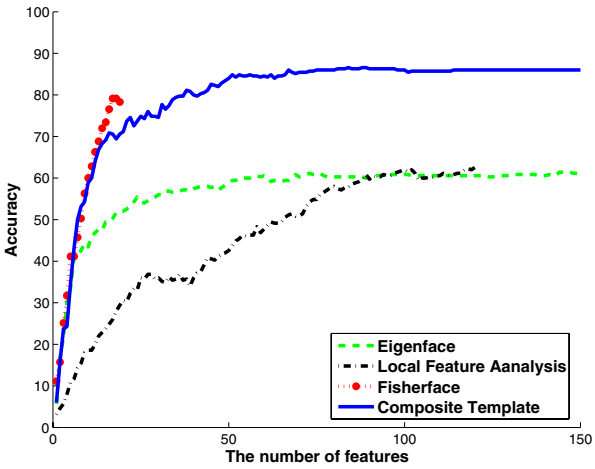


Fig. 5. Accuracy(%) which test images are matched at rank 1 as increasing the number of features

5 Conclusions

By modifying LFA suitable for recognition, we propose a new feature extraction method for face recognition. Our method consists of three steps. First we extract local structures using LFA and select a subset of them, which is efficient for recognition. Then we combine the local structures into composite templates. The composite templates represent data in a more compact form and show compromised aspects between kernels of LFA and eigenfaces. Although LFA is originally problematic for recognition, in the experiments the proposed method has shown better recognition performance than fisherface.

References

1. Turk, M.A., Petland, A.P.: Face recognition using eigenface. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Maui, Hawaii (1991)
2. Penev, P., Atick, J.: Local feature analysis: A general statistical theory for object representation. Network: Computation in Neural Systems **7** (1996) 477–500
3. Bartlett, M.: Face Image Analysis by Unsupervised Learning. Kluwer Academic Publisher (2001)
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiely & Sons (2001)
5. Principe, J.C., Xu, D., Fisher III, J.W.: Information-theoretic learning. In Haykin, S., ed.: Unsupervised Adaptive Filtering: Blind Source Separation. John Wiley & Sons, Inc. (2000) 265–319
6. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence **19** (1997) 711–720

Video-Based Face Recognition Using Earth Mover's Distance

Jiangwei Li[1], Yunhong Wang[1,2], and Tieniu Tan[1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080, P.R. China
[2] School of Computer Science and Engineering, Bei Hang University, Beijing, 100083, P.R. China

Abstract. In this paper, we present a novel approach of using Earth Mover's Distance for video-based face recognition. General methods can be classified into sequential approach and batch approach. Batch approach is to compute a similarity function between two videos. There are two classical batch methods. The one is to compute the angle between subspaces, and the other is to find K-L divergence between probabilistic models. This paper considers a most straightforward method of using distance for matching. We propose a metric based on an average Euclidean distance between two videos as the classifier. This metric makes use of Earth Mover's Distance (EMD) as the underlying similarity measurement between two distributions of face images. To make the algorithm more effective, dimensionality reduction is needed. Fisher's Linear Discriminant analysis (FLDA) is used for linear transformation and making each class more separable. The set of features is then compressed with a signature, which is composed of numbers of points and their corresponding weights. During matching, the distance between two signatures is computed by EMD. Experimental results demonstrate the efficiency of EMD for video-based face recognition.

1 Introduction

Recently, more and more researchers are focusing on face recognition from video sequences [1][2][3][4][5][6], which is very useful in applications of surveillance and access control. Compared to still-based face recognition technologies, multiple frames and temporal information facilitate the process of face recognition. The discriminative information can be integrated across the video sequences. However, poor video quality, large illumination and pose variations, partial occlusion and small size image are the disadvantages of video-based face recognition. To overcome above problems, many approaches, which attempt to utilize multiple frames and temporal information in video, are proposed. Based on whether the temporal information is utilized or not, these schemes can be divided into sequential approach and batch approach.

Sequential approach assumes temporal continuity between two adjacent samples. The continuity property propagates face position and identity frame by

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 229–238, 2005.
© Springer-Verlag Berlin Heidelberg 2005

frame. The previous tracking and recognition result can be utilized for current face tasks. Zhou[2] proposes a tracking-and-recognition approach, which utilizes a very powerful unified probabilistic framework to resolve uncertainties in tracking and recognition simultaneously. Lee[3] represents each person with an appearance manifolds expressed as a collection of pose manifolds. In recognition, the probability of the test image from a particular pose manifold and the transition probability from the previous frame to the current pose manifold are integrated. Liu[4] applies adaptive HMM to perform video-based face recognition task.

The other is batch approach, which assumes independence between any two samples, thus the dynamics of image sequences are ignored. It is particularly useful to recognize a person from sparse observations. The main idea of batch approach is to compute the similarity between two videos. For instance, Mutual Subspace Method (MSM)[5] defines the similarity by the angle between two subspaces spanned by the basis of image sets. Shakhnarovich [6] used multivariate Gaussian models to represent the densities of face sets, and K-L divergence between models is used for matching.

The main problems of the above batch methods are heavy computational cost and not precise models. It is not efficient to estimate the subspace or Gaussian model directly in image space. Moreover, they are not considering the complex data distribution of video data. Both of the subspace and the Gaussian model are only effective to the convex data sets. But in video, head poses, face expressions and illumination change constantly, the shape of data distribution is largely non-convex, more robust model is needed.

Our algorithm is a novel method of batch approaches. In the paper, instead of modeling the data distribution directly in high dimensional image space, we firstly reduce the dimensionality. There we use Fisher's Linear Discriminate (FLDA)[7] to map sets of images to groups of points in low-dimensional feature space. With a linear transformation, FLDA makes sets of images more compact and separable. Furthermore, it reduces the computational consuming. Each video yields a set of points in feature space. We consider a more reasonable model to estimate the distribution of each set. The match of videos can be viewed as the geometric match of sets in feature space. We use the conception of signature to represent each set. By clustering algorithm, the points in a set are grouped into several clusters. The signature is composed of means and weights of these clusters. In fact, it reflects complex data distribution of the set in feature space. So the match problem turns to be the distance measurement of two signatures. Earth Mover's Distance (EMD) is proposed for this purpose. EMD is based on an optimization method for the transportation problem[8]. It computes the minimum work done by moving the weights of one signature to another. EMD is good metric for the comparison of two distributions and in addition, it is adaptive for partial matching, since some faces with large pose variations are thought to be useless and should be discarded in matching. However, when partial matching, EMD is not a metric. In our method, with FLDA for linear transformation, face images are well represented and the computational cost becomes low. In addition,

each distribution of observations in feature space can be efficiently modeled as a signature and the similarity of two videos can be easily and accurately estimated by EMD.

2 Earth Mover's Distance for Recognition

Earth Mover's Distance is a general metric to compare two distributions that have the same weights. To accommodate pairs of distributions that are "not rigidly embedded"[12], the definition of EMD is:

$$EMD(A,B) = min_{f \in F} EMD(A, f(B)) \tag{1}$$

where A and B are two distributions. The purpose of this equation is to seek a transformation f that minimizes $EMD(A,B)$. "FT iteration"[12] is proposed to the solution of object function f. In this paper, considering its application to video-based face recognition, we define it as:

$$EMD(A,B) = \max_{g \in G} EMD(g(A), g(B)) \tag{2}$$

where g is a linear transformation to project two distributions onto feature space so as to maximize $EMD(A,B)$.

2.1 Linear Transformation

As mentioned above, considering the efficiency, the techniques of linear subspace, e.g., PCA[10], FLDA[7] and ICA[11], are taken into account. For simplicity and validity, we use FLDA. In FLDA, between-class matrix S_b and within-class matrix S_w are defined as:

$$S_b = \sum_{i=1}^{H} N_i (m_i - m)(m_i - m)^T \tag{3}$$

$$S_w = \sum_{i=1}^{H} \sum_{x_k \in L_i} (x_k - m_i)(x_k - m_i)^T \tag{4}$$

where m_i is the mean of the image set L_i , and N_i is the number of images in L_i.

The linear transformation matrix W maximizes the following optimal criterion:

$$W = argmax_{\phi} \frac{|\Phi^T S_b \Phi|}{|\Phi^T S_w \Phi|} \tag{5}$$

For video-based face recognition, S_w is generally a full rank matrix. So W can be obtained by seeking the eigenvectors of $S_w^{-1} S_b$ directly.

Using linear transformation, the dimensionality of video data is much reduced. It preliminary solves the problem of heavy computation for video-base face recognition. Furthermore, with FLDA, some variations contained in video are modelled, so the sets of images become compact and separable.

2.2 Earth Mover's Distance

After linear transformation, we obtain two feature distributions $g(A)$ and $g(B)$. In order to define the similarity function $f(A, B)$ between two videos, we introduce the notion of Earth Mover's Distance (EMD). EMD is a general distance measure with application to image retrieval[12][13] and graph matching [16][17]. It is proved much better than other well-known metrics (e.g., Euclidean distance between two vectors). The name is suggested by Stolfi for road design[15].

Given a set of points in feature space, we represent the set with a *signature*. The signature is composed of numbers of clusters of similar features in a Euclidean space. Each cluster is attached to a weight, which reflects the ratio of the number of features in this cluster to the total number of features in the set. During the process of video-based face recognition, each video corresponds to a feature distribution in feature space and it can be modelled as a signature. For simplicity and efficiency, we apply K-Means algorithm[14] for clustering. Each cluster contributes a pair (μ, p_μ), where μ is the mean of the cluster and p_μ is the weight of the cluster. For videos, poses and expressions change constantly. The images in a video form a complex manifold in high dimensional image space. It can not be simply expressed by a single subspace or a single multivariate Gaussian model. Since clustering algorithm is used, signature can well represent the overall feature distribution in a set. In addition, with clustering, some degree of variations, e.g., illumination, poses and expressions, can be tolerated. Moreover, changing the number of clusters, it provides a compact and flexible method to represent data distribution.

Assume two distributions $g(A)$ and $g(B)$. We can imagine $g(A)$ is a mass of earth, and $g(B)$ is a collection of holes. EMD is a measurement of the minimal work needed to fill the holes with earth. This is the reason why it is named "Earth Mover's Distance". Figure 1 shows an example with three piles of earth and two holes. When $g(A)$ and $g(B)$ are represented with signatures, EMD is defined as the minimal "cost" needed to transform one signature to the other. EMD can be formalized as the following linear programming problem: Let $g(A) = \{(\mu_1, p_{\mu_1}), \ldots, (\mu_m, p_{\mu_m}\}$ and $g(B) = \{(\nu_1, p_{\nu_1}), \ldots, (\nu_n, p_{\nu_n}\}$, where μ_i, ν_j are the mean vectors of clusters of $g(A)$ and $g(B)$, respectively, and p_{μ_i}, p_{ν_j} is their corresponding weight. The cost to move an element μ_i, to a new position ν_j is the cost coefficient c_{ij}, multiplied by d_{ij}, where c_{ij} corresponds to the portion of the weight to be moved, and d_{ij} is the Euclidean distance between μ_i and ν_j. EMD is the sum of cost of moving the weights of the elements of $g(A)$ to

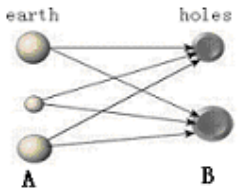


Fig. 1. An example of EMD

those of $g(B)$. Thus the solution to EMD is to find a set of cost coefficients c_{ij} to minimize the following function:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}d_{ij} \tag{6}$$

subject to: (i) $c_{ij} \geq 0$, (ii) $\sum_{i=1}^{m} c_{ij} \leq p_{\nu_j}$, (iii) $\sum_{j=1}^{n} c_{ij} \leq p_{\mu_i}$, and (iv) $\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij} = \min(\sum_{i=1}^{m} p_{\mu_i}, \sum_{j=1}^{n} p_{\nu_j})$. Constraint (i) indicates only positive quantity of "earth" is allowed to move. Constraint (ii) limits the quantity of earth filled to a "hole". Each hole is at most filled up all its capacity. Constraint (iii) limits the quantity of earth provided to holes. Each pile of earth provides at most its capacity. Constraint (iv) prescribes that at least one signature contributes all its weights. If the optimization is successful, then EMD can be normalized as:

$$EMD(A,B) = EMD(g(A), g(B)) = \frac{\min(\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}d_{ij})}{\min(\sum_{i=1}^{m} p_{\mu_i}, \sum_{j=1}^{n} p_{\nu_j})} \tag{7}$$

As illuminated above, EMD reflects the average ground distance between two distributions. The cost of moving indicates the nearness of the signatures in Euclidian space. In our method, after linear transformation with FLDA, corresponding to each distribution, a signature is built with K-Mean algorithm as shown in Figure 2. Each signature contains a set of mean feature vectors and their corresponding weights. In Figure 2, the mean of each cluster is labelled with a red '$\star$' and the weight is denoted under the corresponding image. With more clusters are used, more precise the model is, and more difficult the problem of linear optimization is to solve. Particularly, if some weights of clusters are smaller than a threshold, we discard these clusters since it contributes a little for matching. For videos, these cluster generally consist of faces under bad condition, which deviate far away from normal face clusters. EMD provides a natural solution to this kind of partial matching. However, EMD with partial matching is not a metric for the distance measure of two distributions. Based on the above description, the similarity function between the training video A and the testing video B can be defined as:

$$f(A,B) = \exp(-\frac{EMD(A,B)}{\sigma^2}) \tag{8}$$

where σ is a constant for normalization. The value of the function changes from 0 to 1. Bigger value means more similarity between A and B.

3 Experimental Results

We take a combined database to evaluate the performance of our algorithm. Two experiments are performed. The first experiment fixes the sizes of image sets, and compares the recognition rate with changing the number of eigenvectors or

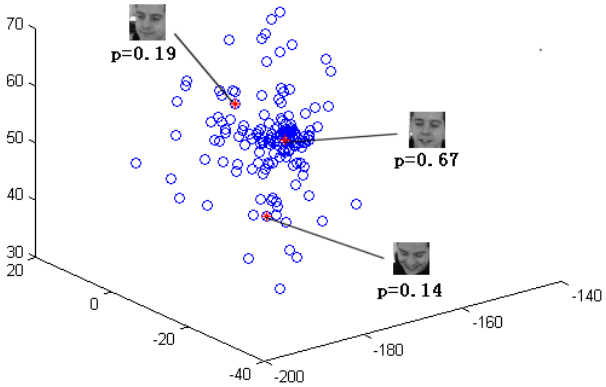


Fig. 2. A signature for video-based face recognition

features. The second experiment changes the sizes of image sets and records the recognition results. The methods are used for comparison is listed as follows:
- Mutual Subspace Method (MSM);
- K-L divergence for classification in original image space (K-L);
- K-L divergence in FLDA feature space (FLDA+K-L);
- EMD in FLDA feature space (FLDA+EMD);
- EMD in original image space (EMD).

We apply the following experimental methods for these algorithms. For MSM, K-L and EMD, we evaluate their performance directly in high dimensional image space. For the other two methods, we firstly reduce dimensionality based on FLDA. For FLDA+K-L, Gaussian function is used to model the set of feature data and K-L divergence between Gaussian models are estimated for classification. For FLDA+EMD, video's matching is based on the measurement of Earth Mover's Distance in the reduced dimensionality space. The label K is assigned to the testing video if the following formula is satisfied:

$$K = argmax_A f(A, B) \tag{9}$$

where A is the reference video in training sets, B is the querying video, and $f(A, B)$ is the similarity function.

3.1 The Combined Database

We use a combined database to evaluate the performance of our algorithm. The database can be divided into two parts: (i) Mobo (Motion of Body) database. Mobo database was collected at the Carnegie Mellon University for human identification. There are 25 individuals in the database. (ii) Our collected database. This part is collected from advertisements, MTV and personal videos. There are 25 subjects in the database. Totally, our combined database contains 50 subjects, and each subject has 300 face images. Figure 3 shows some faces cropped from



Fig. 3. Some cropped faces from sequences

sequences in the database. Using the very coarse positions of eyes, we normalize it to 30×30 pixels and use it for experiments. Some location errors, various poses and expressions can be observed in the database.

3.2 Recognition Rate vs. Number of Eigenvectors or Features

In this experiment, 60 frames of a video are for training and the remaining are for testing. The recognition results are shown in Figure 4. In Figure 4, the remaining frames in a video are divided into 4 testing sets. Each set contains 60 frames. The number of features is changing from 2 to 24. When more features are used, no changes can be observed.

Three methods, i.e., MSM, FLDA+K-L, FLDA+EMD, are compared in the experiment. Those methods have a common that the similarity function can be computed with changeable number of eigenvectors or features. For MSM,

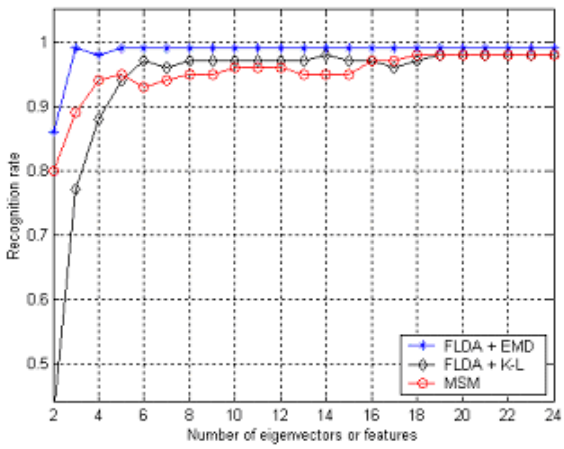


Fig. 4. Recognition rate vs. Number of features

we change the number of eigenvectors to obtain the recognition rate. For other methods, we change that of features. When we use EMD for matching, only 3 clusters in a signature are used. Even more clusters are taken, no significant improvements of recognition rate are made. From Figure 4, we note that the recognition performance of FLDA+EMD is the best. Especially when less than 8 eigenvectors or features are for the experiment, FLDA+EMD outperforms MSM and FLDA+K-L. It is worth noting that the reason why FLDA+EMD is better than FLDA+K-L is that the Gaussian model for K-L divergence is too simple to reflect the data distribution in feature space, while the signatures are competent for this task. We also note that MSM is better than FLDA+K-L for less than 4 or 5 eigenvectors or features. With the increasing of eigenvectors or features, FLDA+K-L will be better.

3.3 Recognition Rate vs. Size of Sets

In the experiment, fixing the number of eigenvectors as their maximal value and changing the size of the sets, we evaluate all the five algorithms in image space. They are: MSM, K-L, FLDA+K-L, FLDA+EMD and EMD. The different partition method of the sets in a video are listed as follows:
(i). A set of 60 images is for training, 8 sets of 30 images are for testing;
(ii). A set of 60 images is for training, 4 sets of 60 images are for testing;
(iii). A set of 100 images is for training, 5 sets of 40 images are for testing;
(iv). A set of 100 images is for training, 4 sets of 50 images are for testing.

The recognition result is shown in Table 1. From this table, we know that the recognition rate of FLDA+EMD is higher than the others. We also note that FLDA+K-L is better than K-L and FLDA+EMD is better than EMD. This phenomenon demonstrates that FLDA is an effective method to reduce dimension and make the features more discriminative. In addition, though EMD directly in image space is not comparable to MSM, but it is superior over K-L. It also demonstrates EMD is an effective metric for classification.

Table 1. Recognition rate vs. Size of sets

Training size	Testing size	MSM	K-L	FLDA+K-L	FLDA+EMD	EMD
60	8×30	95%	70.5%	96%	98%	87%
60	4×60	98%	66%	98%	99%	90%
100	5×40	97%	90%	98%	99%	91%
100	4×50	97%	90%	97%	100%	92%

4 Conclusions

In this paper, we consider a most straightforward method of using distance for matching. The similarity function is established based on the computation of Earth Mover's Distance (EMD) between two distributions. The features are

obtained by mapping the images from high dimensional image space to low dimensional FLDA feature space. Each set is represented with a signature. The solution to EMD is a linear optimization problem to find the minimal work needed to fill up one signature with the other. Experimental results show the performance of EMD and compare it to other batch methods. In future, we will consider the updating method to improve the representative capability of signatures. Moreover, as in [4], time information and transformation probability will be considered to build a more reasonable model to represent a video.

Acknowledgements

This work is funded by research grants from the National Basic Research Program of China (No. 2004CB318110) and the National Natural Science Foundation of China (No. 60332010).

References

1. W.Zhao, R.Chellappa, A. Rosenfeld and P.J Phillips, "Face Recognition: A Literature Survey", Technical Reports of Computer Vision Laboratory of University of Maryland,2000.
2. S. Zhou and R.Chellappa, "Probabilistic Human Recognition from Video", In Proceedings of the European Conference On Computer Vision, 2002.
3. K.C.Lee, J.Ho, M.H.Yang, D.Kriegman, "Video-Based Face Recognition Using Probabilistic Appearance Manifolds", In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2003.
4. X.Liu and T.Chen, "Video-Based Face Recognition Using Adaptive Hidden Markov Models", In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2003.
5. O.Yamaguchi, K.Fukui, K.Maeda, "Face Recognition using Temporal Image Sequence," In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 1998.
6. G.Shakhnarovich, J.W.Fisher, T.Darrell, "Face recognition from long-term observations", In Proceedings of the European Conference On Computer Vision, 2002.
7. P.N.Belhumeur, J.P.Hespanha, D.J.Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, 1997.
8. G. B. Dantzig, "Application of the simplex method to a transportation problem", In Activity Analysis of Production and Allocation, 359-373. John Wiley and Sons, 1951.
9. B. Moghaddam, A. Pentland, "Probabilistic visual learning for object representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, no.7, pp. 696-710, 1997.
10. M.Turk, A.Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, 1991, 3(1): 71-86.
11. M.S.Bartlett, H.M.Lades and T.Sejnowski, "Independent Component Representations for Face Recognition", In Proceedings of SPIE, 2399(3), pp. 528-539, 1998.

12. S.Cohen, L.Guibas, "The Earth Mover's Distance under Transformation Sets", In Proceedings of the 7th IEEE International Conference On Computer Vision, 1999.
13. Y.Rubner, C.Tomasi, L.J.Guibas, "Adaptive Color-Image Embedding for Database Navigation", In Proceedings of the Asian Conference on Computer Vision, 1998.
14. J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.
15. J.Stolfi, "Personal Communication", 1994.
16. Y.Keselman, A.Shokoufandeh, M.F.Demirci, S.Dickinson, "Many-to-Many Graph Matching via Metric Embedding", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2003.
17. M.F.Demirci, A.Shokoufandeh, Y.Keselman, S.Dickinson, L.Bretzner, "Many-to-Many Feature Matching Using Spherical Coding of Directed Graphs", In Proceedings of the 8th European Conference on Computer Vision, 2004.

Face Recognition Based on Recursive Bayesian Fusion of Multiple Signals and Results from Expert Classifier Sets

Michael Hild and Ryo Kuzui

Osaka Electro-Communication University, Graduate School of Engineering
Neyagawa, Osaka, Japan
hild@hilab.osakac.ac.jp

Abstract. We report on a system for person identification based on face images. The system uses sequences of visual wavelength intensity and thermal image pairs as input and carries out classification with a set of expert classifiers (such as ANN or SVM) for each input signal separately. The decisions of the classifiers are integrated both over the two signals and over time as new image pairs arrive, using stochastic recursive inference based on Bayes formula. Our experimental results indicate that both recognition and rejection rates are higher than those for the expert classifiers alone.

1 Introduction

Image-based face recognition systems have reached a high level of performance and technical sophistication, and several commercial systems have appeared on the market. However, benchmark tests indicate that there are still unsolved problems [1]. Some of these problems are:

- Robustness against *illumination-, head pose-, and distance-to-face changes* is still not high.
- Robustness against change in *facial expression* is still difficult to realize.
- Robustness against simple *forgeries*, such as presenting photos instead of real faces to the system, is still difficult to achieve.
- Achieving *very high recognition rates* for registered faces and *very high rejection rates* for unregistered faces is still unaccomplished.

Partly because of these problems, the conviction of many researchers in the field that face recognition should be viewed as just one of several components of a comprehensive biometric person identification system is gaining support.[2, 3] With such systems it may not be necessary to aim at close-to-perfect face identification rates. Nonetheless, we take the stance that it is worth the effort to gain a deeper understanding of face recognition methodologies and to further develop the capabilities of face identification technology.

In this paper we present an approach to face identification which uses a richer input data representation than is usually used. Although we have no intention

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 239–249, 2005.
© Springer-Verlag Berlin Heidelberg 2005

to provide solutions for the facial expression problem in this paper, we intend to show that a richer input data representation in conjunction with an appropriate signal and decision fusion algorithm can be effective for overcoming the high recognition and rejection rates problem. Our method also provides (at least partial) solutions for some of the other problems. In Section 2 we discuss the utility of a specific richer input data representation, in Section 3 we introduce our face identification approach in more detail, and in Section 4 we present experimental results in order to demonstrate the effectiveness of the proposed method. The presented face identification system can be thought of as an extension of an earlier system we described in [4].

2 Face Recognition Based on Multiple Signals

In recent years there have been various efforts to improve the recognition rates of face recognition systems; for example, multiple classifier systems have been applied [5] or methods for information fusion have been explored [6]. These methods were shown to raise the recognition rate, but not enough to achieve reliable face recognition in many application areas. As an extension of these ideas, we propose to combine the usual *light intensity* face images with *thermal* images, and in addition use image sequences instead of single still images for recognition. Furthermore, we suggest to utilize range data at the preprocessing stage. The motivation for this proposal is as follows:

In [7] it was shown that thermal facial images can be used for face recognition, although the recognition rate was not very high. Furthermore, a good deal of the visual information in thermal images seems to be complementary to that of visual spectrum images. Using thermal imagery also helps alleviate the problem of changing facial appearance due to changes of illumination direction. Thermal imagery also can be useful for making face recognition more robust against forgeries, and it enables the system to function in darkness, although at reduced reliability.

Using image sequences instead of single still images provides more information mainly due to head pose variation, which increases the statistical confidence of recognition results. The idea of using faces in motion for face recognition has been discussed in [8], but the number of concrete studies is still limited.

The advantage of using range data is due to the fact that range data allow us to determine the apparent size and position of face image frames within scene images more accurately.

In order to capture this kind of input data we use an image acquisition system which consists of a camera that is sensitive in the far infrared wavelength band for capturing thermal image sequences and a stereo camera which takes color image sequences in the visible spectrum of the scene. The stereo camera provides color and light intensity images together with range data. The two cameras have parallel lines-of-sight in the same direction. Thermal images and visual spectrum images are synchronized and mutually registered. Image registration is carried out by using the range data from the stereo camera system. Face images are cut

out from the scene images by first determining the "center-of-neck" reference point of a person in the image and then determining a *face frame* which is positioned relative to this point. This computation, too, is based on 3D point stereo measurements as well as 2D image data (see [4]). As a result, all images can be aligned with respect to the common center-of-neck reference points of the images.

3 Hybrid System of Trainable Classifiers and Stochastic Decision Fusion

Since in our system the input data are temporal sequences of mutually registered intensity and thermal image pairs, we need an algorithm for (a) integrating information from both intensity and thermal images and (b) sequentially integrating the information from the incoming image pairs. Our proposal for a system with this capability is presented in this section. The system structure is shown in the diagram on the left, and its function is summarized below.

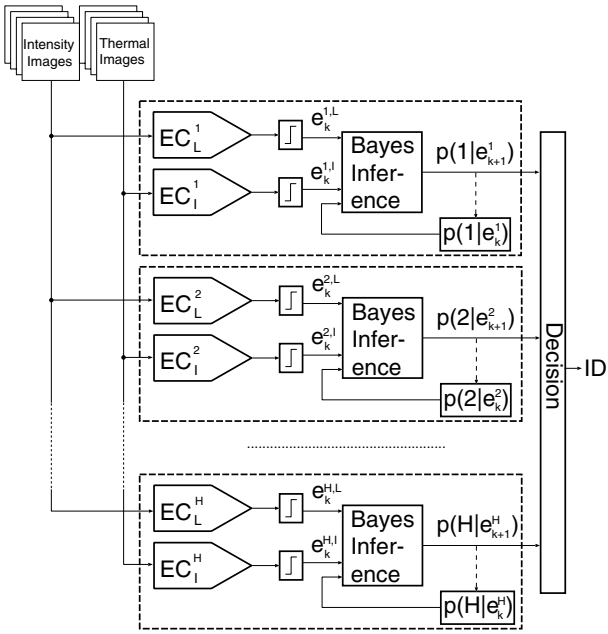


Fig. 1. System for face identification

3.1 System Function

At the start, a face image pair from the input image sequence is fed into all classification channels in parallel, assuming that each channel has been trained as an "expert" for identifying just one *target person*[1]. It makes a decision about whether the presented image pair represents the target person for which it has been trained, or not. Each classification channel includes two trainable classifiers (such as ANNs or SVMs or any other appropriate classifier), one for the intensity image and the other one for the thermal image. The two classifiers make binary decisions about whether the images are instances of the channel's

[1] Details on training are given below

target person or not. These decisions are combined in the following stage by using Bayes formula. For this purpose the prior class probabilities and likelihoods of decision occurrence obtained during the training phase are used. A more detailed discussion of this issue is deferred until after the description of the training process.

This computation is carried out for the same input image pair in every classification channel separately, leading to as many posterior probabilities as there are classification channels. Then the next image pair from the sequence is fetched and processed in the same way, leading to an update of all the channel posterior probabilities. When after a certain number of iterative steps the channel probability of one of the channels supersedes those of all other channels, the target person of this channel is determined as the identity of the present input image sequence[2].

This type of system is a type of modular classification system which combines the decisions from individual experts based on the performance statistics of the expert classifiers.

3.2 System Training

The system is trained with two data sets. Set No.1 includes just nine light intensity-thermal image pairs for each person registered in the face database. These nine image pairs represent faces that are oriented in the characteristic directions "frontal, left, right, up, down, left-down, left-up, right-down, right-up". These images are manually selected from training image sequences with the aim of making the expert classifiers head-pose invariant (see [4]). Set No.2 consists of general intensity-thermal face image pair sequences of all persons registered in the database.

Training is carried out in two phases. In Phase 1 only the expert classifiers are trained with training data set No.1. An individual data configuration is prepared for each expert classifier in the following way: Half of the data consist of the data of the expert classifier's target person, and the other half consists of data from all the other persons in the database. This way, the training process can be kept balanced. Furthermore, for some types of classifiers (such as ANNs or SVMs) it is necessary to alternate target person data and non-target person data. Separate data configurations for light intensity image data and thermal image data are provided.

In Phase 2, all the image pairs of training data set No.2 are fed into all expert classifiers in order to obtain their decisions as to whether the images are instances of their respective expert classifiers's target person or not. When an expert classifier that is in charge of target person h makes a correct decision, the evidence for person h is $\{e^{h,L} = 1, e^{h,I} = 1\}$, otherwise 0. Based on this evidence, the following classifier performance probabilities are computed: $p(e^{h,L}|h)$ is the likelihood that the classifier in charge of target person h indeed makes a decision

[2] A discussion of possible conditions for terminating the recognition process is included in Section 3.3

for person h when the light intensity image L of person h is presented; $p(e^{h,I}|h)$ is the corresponding likelihood for the case of the infrared image; and $p(h|e_k^h)$, $k = 0$ is the a priori probability of the occurrence of the face of a particular person h before any evidence has been collected (i.e. at time $k = 0$).

These probabilities are approximated as frequencies of occurrence. Assuming that the number of persons registered in the face database (or equivalently, the number of classification channels) is H and the number of images contained in each training image sequence is K, the probabilities are estimated as

$$p(e^{h,L}|h) \approx \frac{N(e^{h,L} \equiv 1)}{K} \tag{1}$$

$$p(h|e_0^h) = \frac{1}{H} \tag{2}$$

where $N()$ is a function which counts how often the condition expressed by the function argument is true over the entire length of image sequence.

3.3 Bayesian Decision Fusion

When the system has to classify new face image pairs from the input image sequence, the posterior probabilities that the face images are those of a given channel's target person h are computed by evaluating Bayes' formula for each classification channel. For these computations we can take a narrow, or *local*, view, namely that the computation is limited to the data directly relevant within each classification channel, or a wider, *global* view according to which the computation comprises all data across classification channels. If we take the local view, Bayes formula for the classification channel that is in charge of target person h takes the following form:

$$p(h|e_{k+1}^h, e^{h,L}, e^{h,I}) = \frac{p(e^{h,L}|h) \cdot p(e^{h,I}|h)}{A} \cdot p(h|e_k^h) \tag{3}$$

$$A = p(h|e_k^h) \cdot p(e^{h,L}|h) \cdot p(e^{h,I}|h) + p(\neg h|e_k^h) \cdot p(e^{h,L}|\neg h) \cdot p(e^{h,I}|\neg h) \tag{4}$$

It should be noted that this is a recursive formulation, in which $e^{h,L}$ and $e^{h,I}$ represent the evidence that person h is depicted in the presented image pair $\{L_{k+1}^h, I_{k+1}^h\}$, and e_k^h represents the accumulated evidence for person h up to, but not including the present image pair. The computed posterior probability $p(h|e_{k+1}^h, e^{h,L}, e^{h,I})$, which at the next iteration step replaces the probability of the accumulated evidence $p(e_k^h)$, represents the stochastic fusion of the evidences for person h up to the present image pair. In this case, $p(h|e_k^h, e^{h,L}, e^{h,I}) + p(\neg h|e_k^h, e^{h,L}, e^{h,I}) = 1$ holds.

If we take the global view, Bayes formula has to be modified as follows:

$$p(h_0|e_{k+1}^{h_0}, e^{h_0,L}, e^{h_0,I}) = \frac{p(e^{h_0,L}|h_0) \cdot p(e^{h_0,I}|h_0)}{B} \cdot p(h_0|e_k^{h_0}) \tag{5}$$

$$B = \sum_{h=1}^{H} p(h|e_k^h) \cdot p(e^{h,L}|h) \cdot p(e^{h,I}|h) \tag{6}$$

The difference to the former formulation is mainly in the normalizing factor B. In this case, $\sum_{h=1}^{H} p(h|e_k^h, e^{h,L}, e^{h,I}) = 1$ holds.

The decision stage makes the final decision about the identity of the person based on the computed posterior probabilities. As the decision rule we use the integrals of the posterior probabilities for each classification channel computed over a preset length of the image sequence, and the target person of the channel with the highest integral value is used as the identifier of the input image sequence.

4 Experimental Results

We have carried out a series of face recognition experiments using the proposed method. We also investigated which one of the system components contributes most to the success of the method.

4.1 Experimental Environment

All images were acquired indoors, and the camera was kept stationary. The Digiclops Stereo Vision System made by Point Grey Research was used for acquiring the light intensity image sequences and 3D data points of the scene, and the Thermal Imager made by Mitsubishi Corporation was used for the acquisition of thermal image sequences (see Fig.2). The distance between camera and subject was limited to 1 to 3 Meters and the time delay between acquisitions from the two cameras was kept small. Image processing included background-frame-differencing (see [4] for details).



Fig. 2. Thermal imager (left) and stereo camera (right)

4.2 Configuration of Expert Classifiers Used in Experiments

As expert classifiers we used multilayer feed-forward neural networks (NN). This choice was motivated by the known ability of NNs to identify persons from their face (light intensity) images with reasonably high accuracy (about 90 %). The structure of the NNs and the size of the input face images was determined during a preliminary experimental phase with the objective of obtaining reasonably high recognition rates and generalization ability, but not at the expense of overly long training time requirements. As a result, the images were sub-sampled to a size of 25×25 pixels, and the resulting NNs had a (625-225-1) neurons structure. Examples of input images are shown in Fig. 3 and 4.

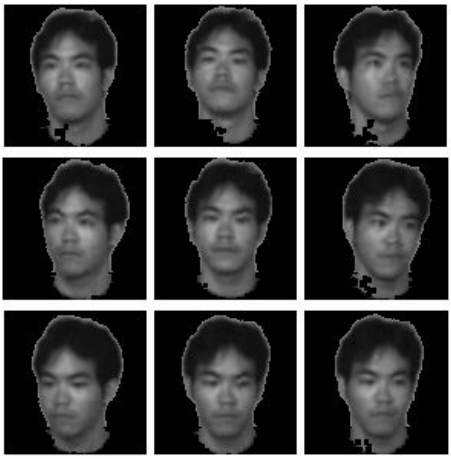


Fig. 3. Input faces for training, oriented in nine directions

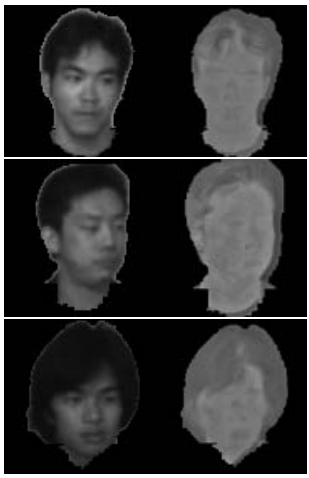


Fig. 4. Examples of intensity-thermal image pairs

4.3 Data Sets Used for Experiments

The light intensity-thermal image pair sequences were taken with the subjects facing the camera system in a generally frontal view, but they were allowed to rotate their heads freely within a $\pm 45^\circ$ angular range in both horizontal and vertical directions. All face images were automatically cut out from the scene images using the automatically determined neck reference point. The nine image pairs of *Training Data Set No.1* (see Section. 3.2) representing nine facial orientations were selected by hand for each of the 30 subjects for training the expert classifiers. In addition, image pair sequences of all 30 subjects were acquired in which the subjects are moving their heads arbitrarily, and each one having a total length of 50 image pairs. The first 20 image pairs of these 30 sequences were used as *Training Data Set No.2* for training the likelihood values in Bayes formula. The latter 30 image pairs of the 30 sequences were assigned to *Recognition Data Set A*, which was used for carrying out recognition experiments. In addition, a *Recognition Data Set B* was created by conjoining *Recognition Data Set A* and *Training Data Set No.2*.

4.4 Results of Experiments with the Proposed Method

We conducted **Recognition Experiment A** by using the two recognition methods described by (3) and (5) and using *Recognition Data Set A* as input data. The obtained recognition rates are shown in Table 1. These rates were computed as follows: First the output posterior probabilities were integrated over the 30 image pairs of a given input image sequence for each classification channel separately, and the recognized person identity was determined as the ID of the

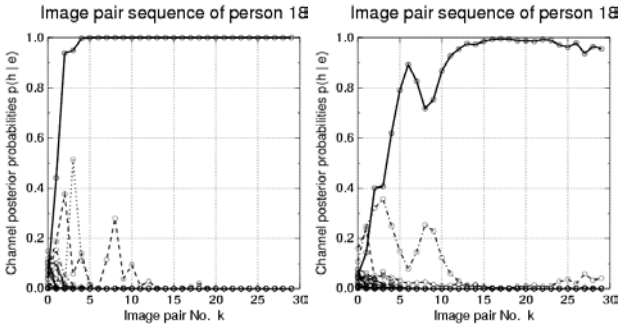
Table 1. Recognition rates for proposed method

	Recog. rate due to (3)	Recog. rate due to (5)
Exp. A	100.0 %	96.67%
Exp. B	96.67%	100.0%

target person from the channel with the highest integral value of posterior probability. Then the recognition rate was computed from the thus obtained correct recognition results divided by the number of input image sequences (here: 30).

Next we carried out **Recognition Experiment B** with *Recognition Data Set B*. This result is also included in Table 1. Inspecting the results of Table 1 reveals that the method using (3) does better than the method using (5) in Experiment A, but with Experiment B, this is reversed. The probable cause for this is the insufficient number of image pairs contained in the Training Data Set No.2 used for training the likelihood values (only 20). However, the obtained recognition rate is very high.

The decisive factor for the success of this method is the stochastic fusion of evidence. This can be inferred from Fig. 5, where typical examples are shown for two image pair sequences and both recognition methods. The curves in each



(a) Method using (3) (b) Method using (5)

(c) Method using (3) (d) Method using (5)

Fig. 5. Evolution of posterior probabilities

graph show the posterior probability variations in all classification channels. It can be observed that the posterior probabilities evolve as new image pairs are processed. The thick lines represent the posterior probability curves of the channel which was in charge of identifying the person of the input image pair sequence. From these results we can verify that the posterior probability of the channel whose target person ID correctly provides the ID of the person depicted in the input image sequence (shown as thick lines) tends to converge to 1.0 after some initial period of fluctuations, whereas the posterior probabilities of all other channels tend to drift toward 0. This behavior could be verified for all test sequences.

On average, a stable state is reached after about 20 input image pairs. This "settling time", of course, depends on a number of factors, such as the quality of the expert classifiers used or the threshold value against which the integral values are judged with respect to deciding whether a stable state was reached.

In the remainder of this section we present the results of additional experiments with which we attempt to analyze which system components contribute most to the success of this face recognition method.

4.5 Results for Expert Classifiers Only

First we tested the face recognition capability of the NN-based expert classifiers alone, using the following procedure: All images of *Recognition Data Set B* (containing 50 image pairs) were used as input to each of the 30 NNs. The recognition rate for a given NN was computed from the number of images R_a that were correctly classified (i.e. e=1) when images of the target person of that NN were presented, divided by the number of images in the sequence (here: 50). The total recognition rate was computed by summing all correctly classified images from each classification channel and dividing them by the total number of all images included in all sequences (here: 1500). Likewise, the rejection rate was computed by summing all correctly classified images (i.e. e=0) when images of persons other than the target person were presented to the NNs and dividing the sum by the total number of all such images included in all sequences (here: 43500).

The obtained recognition rates were **78.1**% for light intensity images and **75.0**% for thermal images, and the rejection rates were **89.53**% for light intensity images and **86.6**% for thermal images. These rates are well below 100%, but also well above the randomness level of 50%.

4.6 Contribution of Light Intensity – Thermal Image Fusion

In order to investigate how much "light intensity – thermal" image fusion contributes to the success of the method, we deactivated the recursive integration of evidence obtained from successive image pairs in (3) and (5), leaving only the fusion of $e^{h.L}$ and $e^{h.I}$ active. We presented the image pairs of *Recognition Data Set B* as input, computed the posterior probabilities for each image pair, and counted the number of correct decisions. We obtained a recognition rate

of **59.8**% when (3) was used and **51.5**% when (5) was used. The drop in the recognition rate, as compared with results for expert classifiers only, is probably due to classifier-image pairings that do not match very well.

4.7 Contribution of Temporal Integration of Image Sequences

In order to investigate how much the temporal integration of image sequences contributes to the success of the method, we deactivated the multi-signal (i.e. intensity-thermal images) fusion feature in Bayes formula and computed the posterior probabilities separately for light intensity and thermal images. Phase 2 system training was also carried out separately for light intensity and thermal images. Recognition was carried out as described in Section 3.3 The obtained recognition rates for *light intensity* images were as follows:

90.0% for (3) and *Recognition Data Set A*;
46.4% for (5) and *Recognition Data Set A*;
76.6% for (3) and *Recognition Data Set B*;
96.6% for (5) and *Recognition Data Set B*.

The obtained recognition rates for *thermal* images were as follows:

73.3% for (3) and *Recognition Data Set A*;
13.3% for (5) and *Recognition Data Set A*;
83.3% for (3) and *Recognition Data Set B*;
90.0% for (5) and *Recognition Data Set B*.

In this case, there are recognition rates which are significantly lower than those for expert classifiers only, but also significantly higher ones have been obtained. Temporal integration obviously introduces a tendency to push the results towards the extremes due to the feedback-effect included in the recursive Bayes formulation. It should be noted that the deactivation of the second signal obviously leads to overall lower recognition results in comparison with the proposed method.

5 Conclusions

The experimental results for the proposed face recognition system which is based on recursive stochastic fusion of two kinds of signals and the results from expert classifier sets indicate that very high recognition rates and rejection rates can be achieved, even though the respective rates for the expert classifiers' alone are not very high. Using long enough image pair sequences in conjunction with reasonably good expert classifiers should always allow us to achieve close to perfect recognition rates, as the information from new images is accumulating and gradually pushing the decision toward the correct person ID. The greatest contribution to the success of this method is due to the temporal fusion of information in image sequences, but the addition of a second image signal (here the thermal signal) also contributes significantly to that success. The second

signal has a stabilizing effect on the recognition results, and it should be noted that close to perfect recognition results could only be achieved when the second signal was activated. We expect even better results from the use of better expert classifiers which we are currently testing in our laboratory.

References

1. P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, The FERET Evaluation Methodology for Face-Recognition Algorithms, IEEE Trans. on Pattern Recognition and Machine Intelligence, Vol.22, No.10 (2000) 1090–1104
2. A. K. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross, Biometrics: A Grand Challenge, Proc. of the 17th International Conference on Pattern Recognition, Vol.2, Cambridge UK, (2004) 935-942
3. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, Multimodal Biometric Authentication Using Quality Signals in Mobile Communications, Proc. of 12th International Conference on Image Analysis and Processing, IEEE Computer Society Press, Los Alamitos (2003) 2–11
4. M. Hild, K. Yoshida and M. Hashimoto, Pose-Invariant Face-Head Identification Using a Bank of Neural Networks and the 3-D Neck Reference Point, Applications of Artificial Neural Networks in Image Processing VIII, N. M. Nasrabadi, A. K. Katsaggelos, Editors, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 5015 (2003) 47–54
5. F. Roli, J. Kittler, G. Fumera and D. Muntoni, An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems, Proc. of Third International Workshop on Multiple Classifier Systems MCS 2002, F. Roli and J. Kittler (Eds) : Lecture Notes in Computer Science Vol. 2364, Springer-Verlag, Berlin (2002) 325–335
6. W. Zhang, S. Shan, W. Gao, Y. Chang, B. Cao and P. Yang, Information Fusion in Face Identification, Proc. of the 17th International Conference on Pattern Recognition, Vol.3, Cambridge UK, (2004) 950-953
7. G. Friedrich and Y. Yeshurun, Seeing People in the Dark: Face Recognition in Infrared Images, Proc. of Second International Workshop on Biologically Motivated Computer Vision, BMCV2002, H. H. Bülthoff, S.-W. Lee, T. A. Poggio and C. Wallraven, (Eds.) : Lecture Notes in Computer Science Vol. 2525, Springer-Verlag, Berlin (2002) 348–359
8. S. Gong, S. J. McKenna and A. Psarrou, Dynamic Vision – From Images to Face Recognition, Imperial College Press, London UK, (2000) chapt. 10

Principal Deformations of Fingerprints

Sergey Novikov and Oleg Ushmaev

Biolink Technologies, Inc.
{SNovikov,OUshmaev}@BiolinkUSA.com
www.BioLinkUSA.com

Abstract. We studied natural relative deformations of fingerprints using the methods of elasticity theory [1,2]. As shown by experiments, the registration of deformations results in almost 3 times improvement for direct overlap matching. Principal components of deformations (eigen deformations) are obtained here from the statistics of genuine matches of fingerprints from several public available databases [3,4]. Energy and cross-compatibility analysis for eigen deformations bases carried out on different datasets are adduced in couple with the examples of implementations where dimensionality reduction in the representation of elastic deformations yields significant advantage.

1 Introduction

As known, the elastic deformation (ED) is the basic distortion factor that negatively affects the performance of fingerprint verification [5-8]. In spite of existence of developed theory of elastic deformations, it is rarely applied to the real-time systems due to computational complexity.

There are different approaches to registration of elastic deformations. One of the first approaches was introduced by D.J. Burr [9], and used the concept of "rubber masks". The way suggested by A.M. Bazen and S.H. Gerez [10, 11] is based on the thin-plate spline (TPS) models, firstly applied to biological objects by F.L. Bookstein [12]. This method requires determining correspondent points in two compared images (matching point) and it suffers from the lack of precision in case of few matching points. Modifications of TPS (approximate thin-plate splines and radial based function splines) were introduced by M. Fornefett, K. Rohr and H. Stiehl [13],[14]. They consider deformations of biological tissues. But this way also requires many matching points (more then 100) what is virtually impossible in fingerprint applications, because number of minutiae in fingerprint image rarely exceeds 50. This fact makes TPS and its variants hardly applicable to fingerprint deformations registration.

Very interesting empirical approach has been suggested by R. Cappelli, D. Maio and D. Maltoni [15]. They developed analytical model of fingerprint deformation, however not specifying the algorithm for its registration.

Solid state mechanics [1,2] defines ED in linear approximation as a solution of Navier linear PDE:

$$\mu\nabla^2\mathbf{u} + (\lambda+\mu)\nabla\mathrm{div}\mathbf{u} + \mathbf{F} = 0. \tag{1}$$

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 250–259, 2005.
© Springer-Verlag Berlin Heidelberg 2005

where $\mathbf{u}$ is the vector of displacement; $\mathbf{F}$ is the external force. Coefficients λ and μ are the Lame's elasticity constants. These parameters can be interpreted in the terms of Young's modulus E and Poisson's ratio ν

$$E = \frac{\mu\,(3\lambda + 2\mu)}{(\lambda + \mu)}, \tag{2}$$

$$\nu = \frac{\lambda}{2\,(\lambda + \mu)}. \tag{3}$$

In [16] we have proposed an algorithm for registration of ED, when a set of mated minutiae is known, using numerical solution of Navier PDE (1) by finite elements method (FEM). There we gave the examples of its implementation and statistical analysis of the distribution of deformation energy for the existing available fingerprint databases [3,4]. The obtained statistics of ED energy distribution allows the estimating of natural limits for possible deformations. In [17] we have proposed one more very simple method based on the convolution with point spread function, and we also estimated theoretical limitations for linear model as the values of some natural parameters such as mean local stretch and averaged discrepancy through image field in fingerprints intersection area. Here, firstly we demonstrate the ROC for direct overlap method before and after the registration of deformations, then we use principal components analysis (PCA) to reduce dimensionality in ED representation, and finally we propose a number of possible applications for the compact ED representations. The analysis has been performed using public available databases FVC2002 [3], the cross-compatibility of the obtained bases of eigen deformations being pointed out.

2 Performance Evaluation

For the evaluation of deformation registration performance we used best fit direct overlap of binary images. This matching method was selected as the most independent on specific minutiae extraction and comparison technique. However, to reduce computational expenses we used Biolink algorithm [18] for primary approximation. An example of direct overlapping of fingerprints before and after nonlinear deformations registration is demonstrated on Fig.1. As one can see, registration of deformations based on proposed model of fingerprint distortions visually corresponds to the real physical processes.

The performance was evaluated on public available FVC2002 databases DB1 and DB3. These fingerprint bases have different quality and size of stored fingerprints. DB1 contains of images of average size, captured by optical scanners; number of deformed fingerprints is relatively great. DB3 fingerprints were obtained from capacity devices. Their size and consequently deformations are sufficiently smaller than for DB1.

We solve equation (1) looking for a minimum of the functional [16] where the main information for deformation registration is pair-wise correspondence of

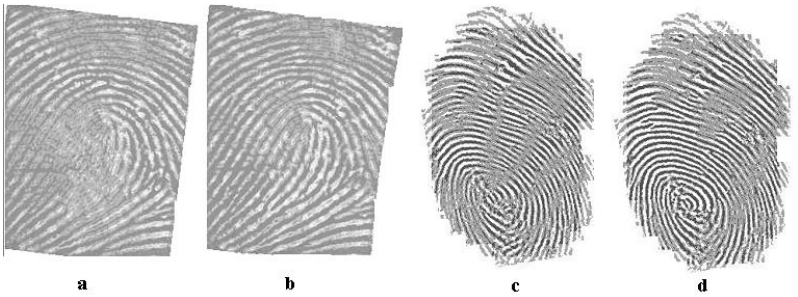


Fig. 1. Direct overlapping of fingerprints. Example 1: capacity scanner, moderate deformation (**a** – without, **b** – with registration of nonlinear deformations). Example 2: optical scanner, strong deformation (**c** – without, **d** – with registration of nonlinear deformations)

fingerprint minutiae. It means that nominally deformations might be registered in impostor matches as well. The average improvement of coefficients of image correlation and average distance between minutiae is shown in the table 1. In impostor matches the improvement of distance (discrepancy) between minutiae is rather irrelevant, because the algorithm of minutiae configuration matching uses great number of characteristics besides minutiae distance. The image correlation appears to be a more informative index of deformation registration performance [20].

Table 1. The average improvement of the basic parameters after ED registration (genuine/impostor)

Index	DB1	DB3
Binary Correlation (relative increase)	+24% / –41%	+7.8% / –55%
Binary Correlation (absolute increase)	+5.8 / –2.2	+1.7 / –3.7
Average distance between minutiae without registration of deformations (pixels)	4.9 / 11.8	4.8 / 7.6
Average distance between minutiae with registration of deformations (pixels)	2.5 / 5.2	2.6 / 5.1

The FAR and the FRR of fingerprint recognition by direct overlap with and without deformations are presented in the Fig.2. In the experiment carried out performance improvement from deformation registration is less than the method's capability, because the minutiae extraction and minutiae matching algorithms may add their errors.

As clear from ROC presented in the Fig.2, the registration of deformations brings sufficient recognition performance improvement, especially on DB1, where the deformation factor is considerable.

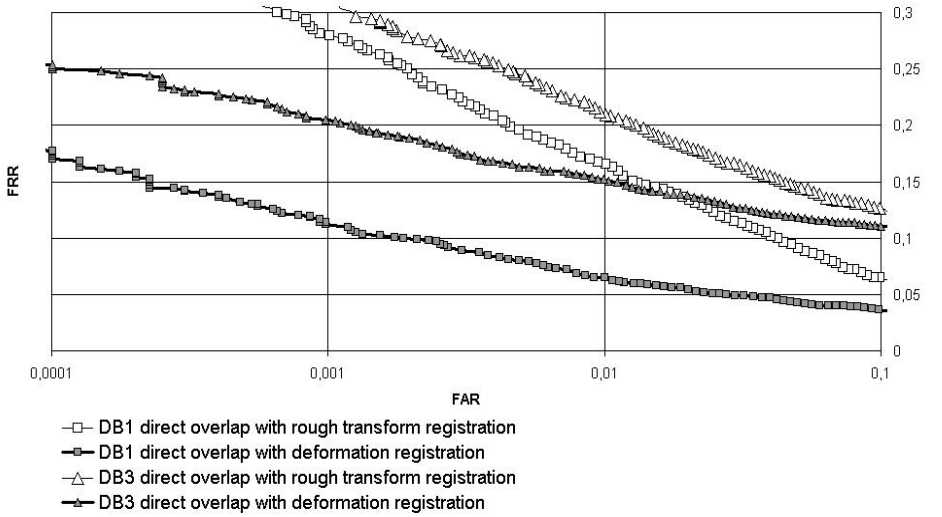


Fig. 2. Fingerprint recognition by direct overlap with deformation registration

3 Eigen Deformations and Reduction of Dimensionality

Further on we consider all deformations over image area of 300x400 pixels size with 500dpi resolution that corresponds to the averaged natural application area. Direct representation of deformations as displacement vector at each pixel has huge dimension and make any further work virtually impossible. Even assuming that we interpolate the local displacement function (deformation) by the nodes of a very sparse grid, say 16 pixels apart, it would require storing of 468 values of 2D vector components, that significantly complicates the analysis.

Following traditional principal components analysis (PCA) scheme [19], one might represent all the valued in nodes as one 936 dimensional vector, find eigen values and eigen vector of correlation matrix that has been calculated on genuine matches, and then select the first n eigen vectors correspondent to the first n eigen values taken in the descending order, so as to make the residual variance being no greater than, for example, 1%. As computational experiments had shown, we need approximately 30 principal components to cover 99% of overall variance, and the cumulative variance distribution has very slight slope, i.e. the effective dimensionality is rather great. Moreover, the first eigen deformation look somewhat weird. The reason is both in the loss of robustness (due to relatively small training set, each FVC2002 database provides us with 2800 genuine matches) and in linear relations between spatial locations being inadequate to represent the nature of fingerprints deformations.

To reduce the dimension and consequently increase robustness, we initially perform 2D DFT for both vertical and horizontal components of displacement function. Analysis showed that 92 components (46 per each displacement component) of lower frequencies cover approximately 93% of the entire image energy

and 98% of the image energy in the internal area (without 16 pixel wide margins where side effects dominate).

To obtain principal deformations, we finally performed PCA in the 92D spectral space. It must be mentioned, that obtained deformation was normalized with respect to rigid movement, i.e. integral shift and rotation were subtracted from displacement field.

As soon as spectral eigen vectors had been obtained, we perform the reverse Fourier transform to get the correspondent eigen vectors in the initial space. As a result, the first 20 eigen vectors covers more than 98% of overall variance in the initial space, while 4 first eigen vectors cover 85% of variance (Fig. 3).

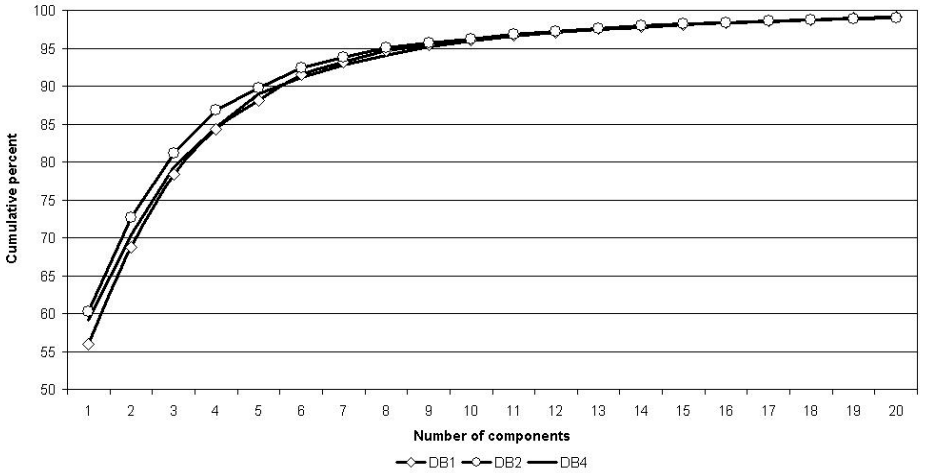


Fig. 3. Cumulative percent of variance of principal deformations for FVC2002 Databases

The 4 principal deformations for all 3 databases are presented in the Fig. 4. As one can see, the outlook of the first 4 components coincides for all 3 datasets used (Db1, Db2, Db4). At the same time their order varies what is normal because it depends on application style of current DB clients. The first component is a micro rotation and appears due to the impossibility of precise factorization for integral rotation and border effects. It should be pointed out, that the dispersion of the displacement through the image for this component is one order less to the dispersion for the next 3 components. Thus we have an unexpected fact: the influence of the main component in each pixel is far below the intensity of the significant distortion while integrally it dominates.

The remaining 3 principal components are absolutely natural: torsion and traction along two axes. The qualitative description of the main deformations for fingerprints has been given earlier [15], however here we obtained them as a result of the precise statistical experiment and as a set of orthonormal vectors (the principal components are othonormal in the spectral space because of PCA

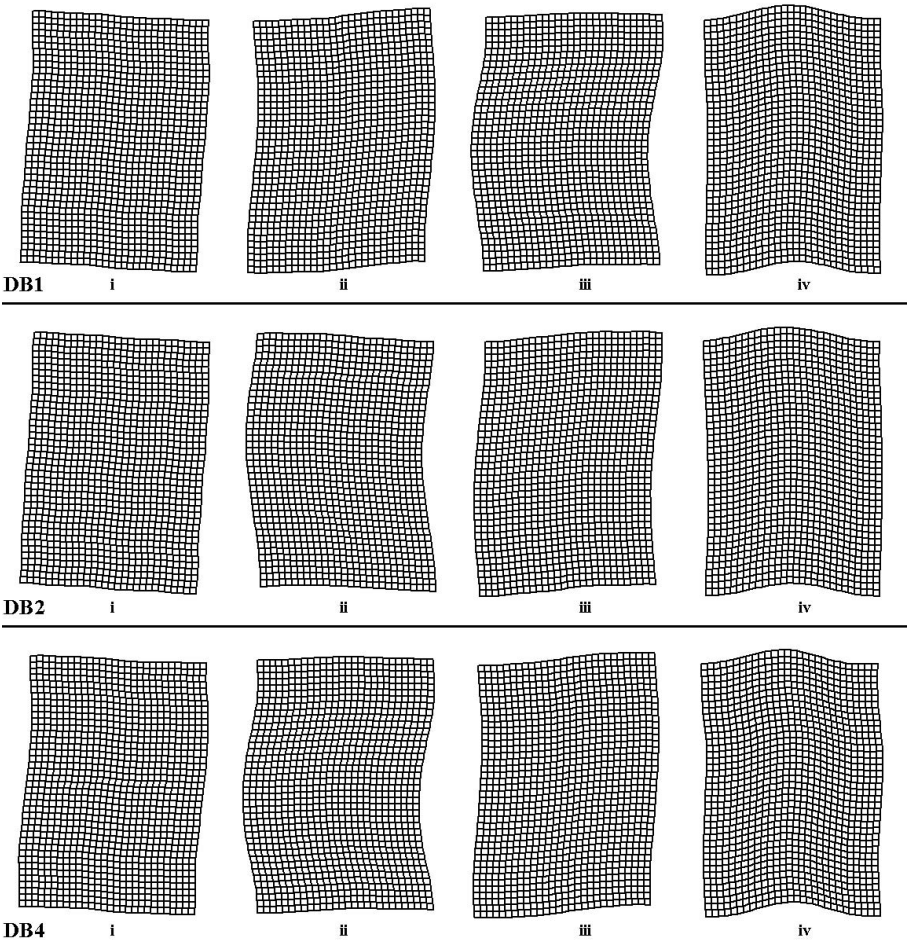


Fig. 4. Four principal deformations for FVC Databases

properties, at the same time obtained correspondent principal deformations are orthonormal in the initial space because the DFT is a unitary transformation).

Since the set of the initial n eigen deformations is orthonormal, any deformation D can be expanded in the eigen deformation set:

$$\mathrm{D} = \sum_{i=1}^{n} c_i \mathrm{PD}_i + \epsilon_n =_{\mathrm{def}} \mathrm{D}_n + \epsilon_n, \tag{4}$$

where coefficients are standard inner product $c_i =< \mathrm{D}, \mathrm{PD_i} >$ if displacement field is represented as vector. The discrepancy norm ϵ_n is equal to

$$||\epsilon_n||^2 = ||\mathbf{D}||^2 - \sum_{i=1}^{n} c_i^2 \tag{5}$$

and not zero as far as principal deformations set is not full basis (its size is not greater than 92 while the possible dimension of deformation space in our experiment is 24000).

To support the visual equivalence (or cross-compatibility) of principal components obtained from different databases, we carried out the following experiment. Deformations obtained from genuine matches for each FVC2002 database were approximated by three sets of principal deformations, correspondent to each base. In that case we get three different approximations for each deformation. As the indices of quality of approximation the following distances between initial deformation and restored deformation (D_n) are used:

1. Mean absolute difference depending on number n of principal components involved
2. Mean relative area of exact approximation where deformation restored with pixel wise precision
3. Mean relative discrepancy ($||\epsilon_n||^2/||\mathbf{D}||^2$)

The results presented in the Fig.5 reveal similar approximation characteristic of principal deformations obtained from different bases confirming that character of principal deformations is independent on the database while quantative specification of deformation sufficiently varies among the databases. The DB1 images tends to be most deformed (mean displacement is about 0.5 mm, while "immobile", in sense of [15], area is less then 45% of the entire image) while DB4 synthetic fingerprints having almost the same size as DB1 ones approximately 2 times less deformed.

4 Possible Implementations

The reducing in dimensionality enables or facilitates the solution of many applied problems. Let us enlist some of them.

1. **Synthetic databases generation**
   Although deformation character for synthetic database Db4 is almost identical to the other FVC databases and main deformations (up to slight asymmetry) almost coincide, it is hard to imagine how much effort its creation required from the authors. Now, having principal components and their standard deviations (PCA eigen values), it is very easily to generate the set of deformations with statistical features being equal to the natural ones. It may be done by a direct sum of principal deformations multiplied by random values that are distributed with corresponding standard deviations and zero means.
2. **Simulating of multiple applications at enrollment phase**
   Since in modern biometric systems it is regarded as improper to ask a client for multiple applications while being enrolled, the additional information may be obtained by artificial applications being simulated in a way similar to 4.1.

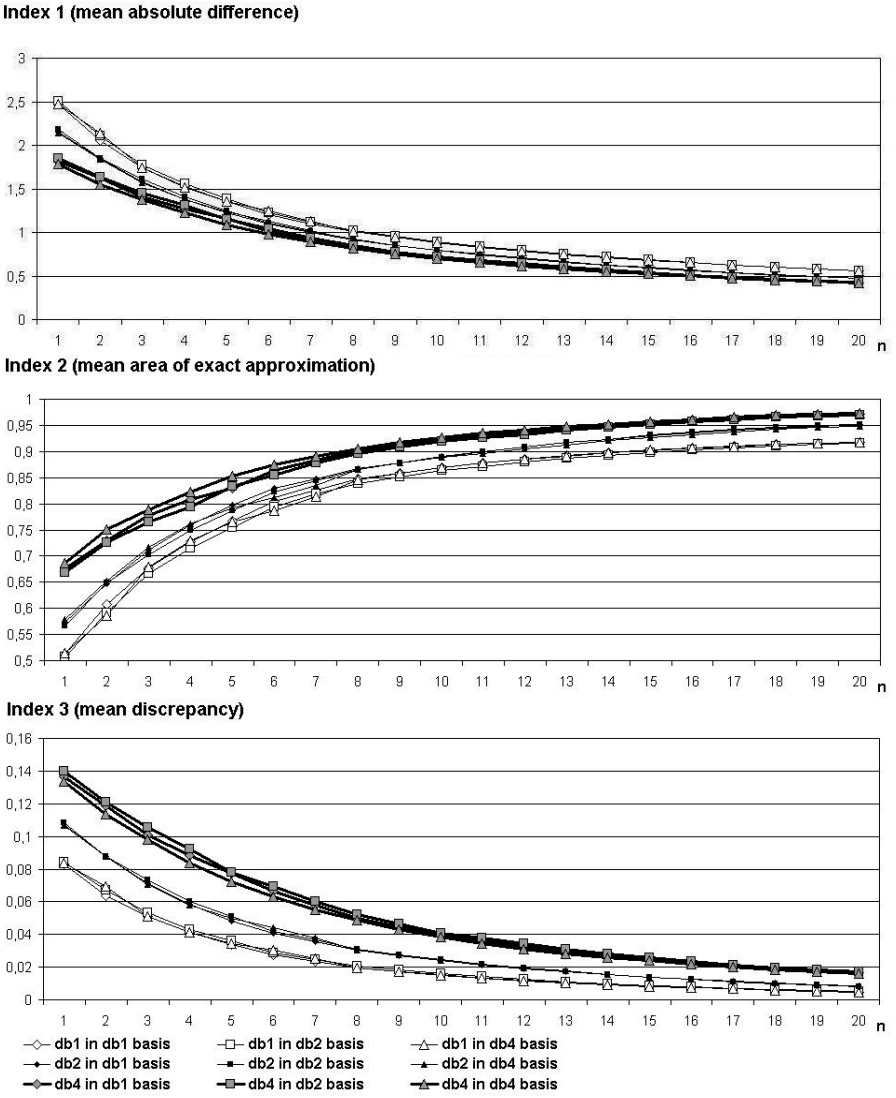


Fig. 5. Quality of deformation approximation

3. **On-line template improvement**

An interesting approach has been proposed in [11] to find the most effective deformation, i.e. actually to approximate the initial (in some sense "undeformed") state of the fingerprint that already had been enrolled. For this purpose the collection and analysis of the deformations registered during each genuine match is required. The solution becomes certainly more effective after the abrupt reduction of dimensionality using suggested eigen deformation technique.

5 Conclusion

We have obtained principal components of plain fingerprint deformations based on the statistics from several public available databases. The initial 4 principal deformations provide about 80% precision of representation and have an illustrative interpretation: micro movement that compensates matching inaccuracy, torsion and orthogonal tractions, 8 eigen vectors give 90% precision, and 20 – up to 96%. After 10th component we have practically only analytic basis (harmonics).

We must underline that intrinsic dimension of real *plain* fingerprint deformation is not greater than 10–15 because the 10–15 principal components describe deformation to high degree of accuracy. Residual part is almost noise component that appears due to minutiae extraction inaccuracy.

It has been demonstrated that the bases obtained on different datasets correlate almost absolutely. We gave a number of applications where the reducing in dimensionality is crucial.

References

1. Shames, I.H. and Pitarresi, J.M., Introduction to Solid Mechanics, Upper Saddle River, NJ, 2000.
2. Landau L.D., Lifshits E.M., "Theory of Elasticity: Course of Theoretical Physics", Butterworth-Heinemann Edition, 1995.
3. FVC2002, the Second International Competition for Fingerprint Verification Algorithms (FVC2000), bias.csr.unibo.it/fvc2002/.
4. First International Competition for Fingerprint Verification Algorithms (FVC2000), bias.csr.unibo.it/fvc2000/.
5. S. Pankanti, S. Prabhakar and A.K. Jain, "On the Individuality of Fingerprints", IEEE Trans. PAMI, 2002, 24(8), pp. 1010-1025.
6. Wilson C.L., Watson C.I., Garris M.D., and Hicklin A., "Studies of Fingerprint Matching Using the NIST Verification Test Bed (VTB)" // available at ftp://sequoyah.nist.gov/pub/nist_internal_reports/ir_7020.pdf
7. Lee H.C. and Gaenssley R.E., Advances in Fingerprint Technology, Elsevier, New York, 1991.
8. Halici U., Jain L.C., Erol A., "Introduction to Fingerprint Recognition", Intelligent Biometric Techniques in Fingerprint and Face Recognition, CRC Press, 1999.
9. Burr D.J., "A Dynamic Model for Image Registration" Computer Graphics and Image Processing Vol. 15 pp. 102-112, 1981
10. Bazen A.M., Gerez S.H., "Thin-Plate Spline Modelling of Elastic Deformation in Fingerprints", Proceedings of 3rd IEEE Benelux Signal Processing Symposium, 2002.
11. A. Ross, S. Dass and A. K. Jain, "Estimating Fingerprint Deformation", Proc. of International Conference on Biometric Authentication (ICBA), (Hong Kong), LNCS vol. 3072, pp. 249-255, Springer Publishers, July 2004.
12. Bookstein F.L., "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations", IEEE Trans. PAMI, 1989, 11(6), pp. 567-585.

13. M. Fornefett, K. Rohr and H.S. Stiehl, “Elastic Medical Image Registration Using Surface Landmarks with Automatic Finding of Correspondences”, In A. Horsch and T. Lehmann, editors Proc. Workshop Bildverarbeitung fur die Medizinl, Informatik actuell, Munchen, Germany, Springer-Verlag Berlin Heidelberg, 2000, pp. 48-52.
14. M. Fornefett, K. Rohr and H.S. Stiehl, “Radial Basis Functions with Compact Support for Elastic Registration of Medical Images”, Image and Vision Computing, 19 (1-2), 2001, pp. 87-96.
15. Raffaele Cappelli, Dario Maio, Davide Maltoni, “Modelling Plastic Distortion in Fingerprint Images”, ICAPR2001, pp. 369-376.
16. Ushmaev O., Novikov S., “Registration of Elastic Deformations of Fingerprint Images with Automatic Finding of Correspondences”, Proc. MMUA03, Santa Barbara, CA, 2003, pp. 196-201.
17. Sergey Novikov and Oleg Ushmaev, ”Registration and Modelling of Elastic Deformations of Fingerprints”. Biometric Authentication (ECCV 2004 International Workshop, BioAW2004), Prague, Chezh Republic, May2004, Proceedings. Springer, Eds. Davide Maltoni, Anil K. Jain. pp. 80-88.
18. U.S. Patent No. 6 282 304.
19. R.O. Duda, P.E. Hart, “Pattern Classification and Scene Analysis”, John Wiley & Sons, Inc. 1973.
20. O.Ushmaev, S.Novikov, ”Efficiency of Elastic Deformation Registration for Fingerprint Identification”. 7th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-7-2004). St. Petersburg, October 18-23, 2004. Conference Proceedings (Vol. I-III), Volume III, St. Petersburg, SPbETU 2004, pp.833-836.

Fingerprint Mosaicking by Rolling and Sliding

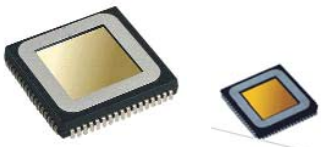
Kyoungtaek Choi, Hee-seung Choi, and Jaihie Kim

Department of Electrical and Electronic Engineering,Yonsei University
Biometrics Engineering Research Center, Seoul, Korea
maninquestion@yonsei.ac.kr

Abstract. In this paper, we propose a new scheme that a user enrolls his fingerprint images sequentially captured by rolling and sliding his finger, thus continuously contacting on the sensor. We also developed an image-fusion algorithm to mosaic the images obtained by the enrollment scheme. Conventional fusion algorithms for fingerprint images are based on large-sized sensors, and they are easily failed to combine images if there are not enough common areas among images. Our enrollment scheme assures that the common area between two sequential images is large enough to be combined even with a small-sized sensor. Experimental results show that average combined images are 1.91 times larger than a single image, and success rate for combining is 2.3 times higher than a conventional dab approach.

1 Introduction

Nowdays small sized sensors have been spread more and more for fingerprint recognition. One advantage of small sensors (e.g., solid-state sensors) is that they can be used with many applications (e.g., laptops, cellular phones). However, information about the fingerprint is limited due to the small physical size of the sensing area , as shown in Fig. 1.



(a) (b)

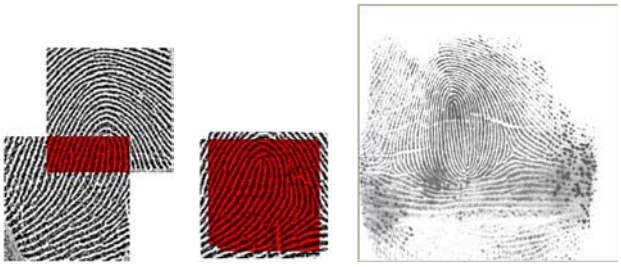
Fig. 1. Fingerprint Sensors: (a) a large sensor(sensing area 16×16mm), (b) a small sensor(sensing area 6.5×6.5mm)

The small overlap between the template impression and the query impression due to the small sensing area, produces inferior recognition performance, for example, a higher rate of false rejects. To overcome this problem, some researchers have explored the field of fingerprint fusion. The fingerprint fusion algorithm can

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 260–269, 2005.
© Springer-Verlag Berlin Heidelberg 2005

be categorized largely into two types. The first type fuses feature sets from several fingerprint images; the second type produces a mosaicked fingerprint image with several images. At the feature level, the fusion algorithm is very simple, and also, its recognition performance is better than that of a non-fusion system. However, it is difficult to apply it to the systems which use features different from those used in the feature level fusion algorithm. It is also difficult to add new features to the system. Several researchers have studied fusion systems at an image level. Jain and Lee captured several impressions by dabbing a finger on a small sensor and made a mosaicked image by using a rigid transform [1],[2]. Ratha captured sequential impressions by rolling a finger on a large sensor. The larger sensor can cover a whole fingerprint so we can mosaic sequential impressions easily by stitching without calculating the transform among the fingerprints [3].

The multiple impressions captured by the dab approach (as used by Jain and Lee) are very hard to be mosaicked when the overlap between two images is very small, as shown in Fig. 2(a). In addition, the dab approach has little effect on two impressions obtained from a similar portion of a finger, as shown in Fig. 2(b). Otherwise, the rolling approach (suggested by Ratha) is able to acquire a whole fingerprint, but it requires a large sensor, as shown in Fig. 2(c). It therefore cannot be applied to systems that use a small sensor.



(a) (b) (c)

Fig. 2. Mosaicked images of previous algorithms: (a) small overlap area, (b) large overlap area, (c) rolled image with a large sensor

To capture a whole fingerprint with a small sensor we present a new enrollment scheme and propose a new mosaicking algorithm which mosaics sequential images captured by our enrollment scheme.

Our paper is organized as follows. In Section 2, we describe our enrollment scheme and our system flow chart. In Section 3, we describe an image selection method that rejects a low-quality image from a sequence and a simple stitching method. In Section 4, we describe the image mosaicking process that calculates the local transform parameters between sequential images and warps one image to the other. The experimental results are shown in Section 5. Finally, conclusions appear in Section 6.

2 Fingerprint Enrollment and System Flow Charts

To capture a whole fingerprint image with a small sensor, we present a new enrollment scheme as shown in Fig. 3. The user puts the left side of his finger on a small sensor, as shown in Fig. 3(a). The user rolls the finger on the sensor until the right part of the captured image contains the foreground region of a fingerprint, as shown in Fig. 3(b). The user then slides the finger on the sensor horizontally to acquire the part of the fingerprint that has not yet been captured, as shown in Fig. 3(c) and 3(d). Finally the user rolls the finger again to capture the right side of the finger, as shown in Fig. 3(e). Using this enrollment scheme, we are able to capture sequential images of a fingerprint and mosaic the sequence to produce a whole fingerprint. By using our enrollment scheme, we can obtain a wider fingerprint area with a small sensor than by using the dab approach.

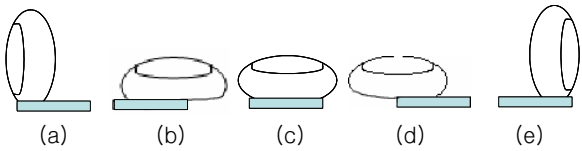


Fig. 3. New Enrollment Scheme



Fig. 4. The images captured by sliding vertically: (a) samples captured by sliding vertically, (b) the mosaicked image

We have explained how to acquire the horizontal region of a finger, as shown in Fig. 3. We can also acquire the vertical region of a finger by sliding a finger on the sensor vertically, as shown in Fig. 4. Furthermore, we can acquire a whole fingerprint by combining the horizontal region with the vertical region of a finger. The algorithm used to mosaic images captured by rolling and sliding a finger on the sensor horizontally is very similar to the algorithm that mosaics images

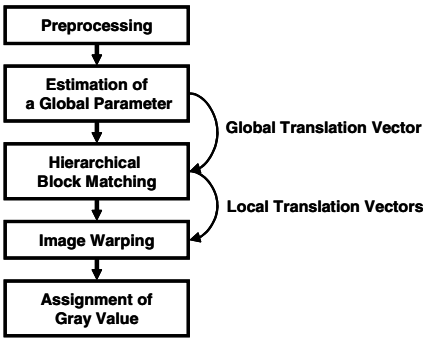


Fig. 5. The system flow chart

captured by sliding the finger vertically. In this paper, we will explain only the former algorithm.

Fig. 5 shows the flow chart of our algorithm. In the preprocessing procedure, we segment foreground (fingerprint) and background areas in each frame and reject the motion blurred image. If the translation between the previous frame and the current frame doesn't occur, our algorithm stitches the current frame to the previous one without alignment. If the translation does occur, the algorithm aligns the current frame to the previous one. To mosaic two images, our algorithm uses the global transform parameter to align one image to the other (roughly) and then uses the local transform parameters to align the local image blocks of one image to those of the other image. The corresponding points are considered to be the center points of each block. Finally, we warp one image to the other with these corresponding points and mosaic the two images. Each procedure of our algorithm is explained in greater detail in the following sections.

3 Preprocessing

In the preprocessing procedure, we first segment foreground and background areas in an image using the block variance of the image. After the segmentation, we find the mean and variance of the foreground area and normalize the image, as suggested in [4].

In our enrollment scheme, it is possible that a few images become blurred. Blurring occurs when a user slides his finger very fast on the sensor. These motion-blurred images are rejected by thresholding the median value of the tenengrad of each image [5]. (The tenengrad refers to the magnitude of the gradient of an image.) The median value of the tenengrad is computed as

$$T = med\left(\frac{1}{N \times N}\sum_{i,j \in Fb_k}^{N \times N} G_x(i,j)^2 + G_y(i,j)^2\right) \tag{1}$$

The gradients G_x and G_y of an image are calculated by the sobel gradient algorithm. and then the median value of the tenengrad is calculated in the fore-

ground area blocks. Fb_k means the foreground area and the size(N×N) of each block is 8*8 pixels. Fig. 6 shows the median tenengrad values of each frame. In Fig. 6 the tenengrads of motion-blurred images at frame number 17,79,92 are below the threshold (150) that is defined by using 2400 samples. While the image at frame number 33 has a very high tenengrad value. After rejecting motion-blurred images, we check if the translation between the current frame and the previous one has occurred. If there has been no translation, we stitch the current frame to the previous one to expand a foreground area by using the center method as proposed in [3]. To check the occurrence of the translation, we check whether the SAD (Sum of Absolute Difference) in the common foreground region between two frames is below the threshold or not. If the translation has occurred, we mosaic two frames by using the warping method. In the following section, our image mosaicking procedure is explained in more detail.

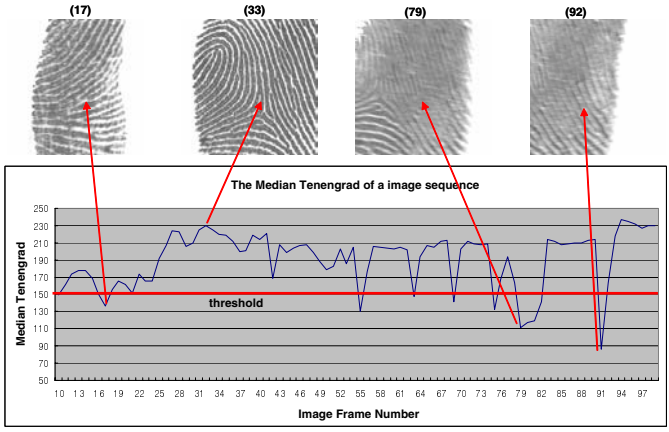


Fig. 6. The Median tenengrad of a image sequence

4 Image Mosaicking

The image mosaicking procedure is divided into three parts. The first part involves searching for a global translation vector, the second part involves hierarchical block matching to find a local translation vector of each image block, and the third part involves the warping and assigning of gray-values to pixels which are found at the boundary between the two frames.

4.1 Searching for a Global Translation Vector

We do not consider the rotation parameter between two frames because users slide their fingers horizontally on the sensor. We use the block-matching algorithm proposed by Chen *et al* to find a global translation vector and local ones

[6]. Chen's algorithm allows the translation vector to be searched in a global minimum (like the full-search algorithm). Processing time can be reduced by about 1/10 when using Chen's algorithm. Even though many algorithms are faster than Chen's algorithm, most of them do not guarantee the global minimum solution. Furthermore, these searching algorithms are more likely to be trapped in a local minimum in fingerprint images than in other images, because the local patterns of a fingerprint image are very similar to each other.

4.2 Hierarchical Block Matching

After finding a global translation vector, we have to find the local translation vectors hierarchically, as shown in Fig. 7. When the user slides the finger on the sensor, plastic distortion caused by rubbing is inevitable. This makes it hard to align one image to the other exactly when using a global translation vector. To solve this problem, we divide an image into several blocks and find the local translation vector of each block. We then warp one image to the other with these local translation vectors. To find these local translation vectors, we align two frames (roughly) with a global translation vector and set the common foreground area between two frames. The common area is divided into four sub-blocks and each sub-block is divided into four high-level blocks until the smallest block size becomes 16×16 pixels, as shown in Fig. 7. The size of each block is a multiple of 2, so if the size of the common area is not a multiple of 2, the sub-blocks overlap. The smaller the sub-block size, the larger the probability of incorrect searching of the sub-block translation vector. This incorrect searching can be due to image noise, plastic distortion and simple patterns in a small area of a fingerprint image. To find the vectors correctly we implemented a regularization step on the Bayesian theory. The translation vector of the sub-block i in level $l+1$ is computed as

$$t_{l+1,i} = \arg\max_{t^k_{l+1,i}} \left(P\left(t^k_{l+1,i} / t_{l,i/2} \right) \right) = \arg\max_{t^k_{l+1,i}} \left(P\left(t_{l,i/2} / t^k_{l+1,i} \right) \cdot P(t^k_{l+1,i}) \right) \tag{2}$$

$$P\left(t_{l,i/2} / t^k_{l+1,i} \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{\frac{-\left\| t_{l,i/2} - t^k_{l+1,i} \right\|^2}{2\sigma^2}} \tag{3}$$

$$P(t^k_{l+1,i}) = \frac{1}{N(S_{l+1}) - 1} \left(1 - \frac{SAD(t^k_{l+1,i})}{\sum_{t^x_{l+1,i} \in S_{l+1}} SAD(t^x_{l+1,i})} \right) \tag{4}$$

We assume that the A-posterior PDF (Probability Density Function) of the translation vector is Gaussian and find the MAP (Maximum A-posterior Probability) solution for the local translation vector. In Eq. 3 $t^k_{l+1,i}$ is the kth translation vector of the block i in level $l+1$ and the A-posterior probability in the translation vector $t_{l,i/2}$ of the parent block $i/2$ in level l, is computed as Eq. 3. The prior probability of the translation vector $t^k_{l+1,i}$ is $P(t^k_{l+1,i})$ and the size of

the searching area for the translation vector is $N(S_{l+1})$. We find the translation vector of each sub-block in each level hierarchically through Eq. 2, as shown in Fig. 7. We define the plastic distortion of each block as the difference between the local translation vector and the global one. The center points of each smallest sub-block are the corresponding points between two images, which can be used in the image warping procedure.

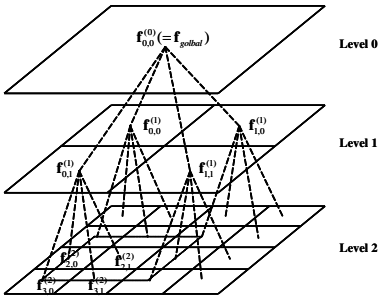


Fig. 7. Hierarchical structure for searching the local translation vector

4.3 Warping and Gray-Value Assignment

After the block translation vectors are estimated, local distortions are compensated for using the point-based warping technique. Every center point of every block is considered a corresponding point when image Q is warped. That is, points derived from a global translation vector are utilized as destination points while source points are defined as translated points by the local translation vectors, as shown in Fig. 8. In the warping procedure, we use the 2-pass mesh-warping algorithm [7]. This algorithm includes Fant's resampling algorithm and uses cublic spline as the interpolation method. The 2-pass mesh-warping algorithm is simple and well-suited to our algorithm because the center points of the sub-blocks have a lattice structure. After performing the warping procedure, we stitch the image Q to the image P by using the center method as suggested in [3]. Finally, we obtain a mosaicked image from images P and Q. We obtain a wide fingerprint image from an image sequence by applying our algorithm, as shown in Fig. 9.

5 Experimental Results

To collect the fingerprint images, we used 4 enrollment schemes as follows

1. An image for each finger enrolled by using the dab approach
2. A mosaicked image with images enrolled by using the dab approach [2].
3. A rolled image enrolled by rolling a finger on a large sensor [3].
4. A mosaicked image with sequential images enrolled by our enrollment scheme.

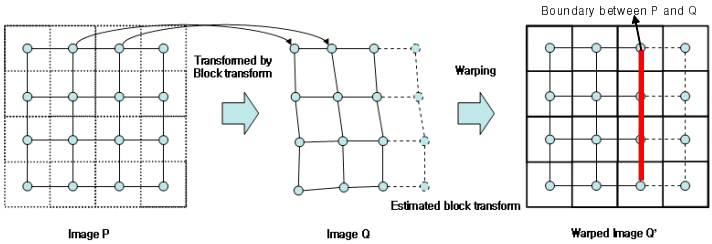


Fig. 8. Illustration of the warping scheme



Fig. 9. The mosaicked image with a fingerprint sequence

We enrolled 100 fingers through 4 enrollment schemes. For the 1st and the 2nd enrollment sheme we capture 1000 images (10 images per a finger) and for the 3rd one we capture 100 rolled images (an image per a finger) and for the last one we capture 100 sequences (a sequence per a finger). The number of images belong to a sequence becomes different by each person. We use the ACCO 1394 sensor whose image size is 600×600 with 500 dpi resolution [8]. For the 1st, the 2nd and the 4th enrollment shemes, we clipped the center region of an image whose size is 192×192. Instead of a small sensor, we used the ACCO 1394 sensor because its frame per second is enough to acquire sequential images enrolled by our enrollment scheme.

We compare the mosaicking success rate of Lee's algorithm [2] with ours as shown in Table. 1.

The success rate of Lee's algorithm varies according to the base image to which the algorithm align other images. Lee's algorithm has no method how to select the base image among several images, so we select an image among 10 images as a base image and align others to the image for Lee' s algorithm

Table 1. Mosaicking Success Rate

# of images	2	4	6	8	10	*our algorithm*
Random	0.48	0.279	0.197	0.147	0.48	0.88
Optimum	0.85	0.76	0.62	0.48	0.38	

Table 2. The size of foreground area and number of minutiae

Enrollment	One image	Lee's algorithm					Rolled	Ours
		2	4	6	8	10		
Size	37425	47067	58534	63692	66548	66869	229886	71639
# of minutiae	14	22	29	31	33	35	143	42

and execute this procedure 10 times per a finger by changing the base image. In Table. 1, *# of images* is the number of images successfully mosaicked and Random is the mosaicking success rate when we select a base image randomly and Optimum is the success rate when we select a base image to align others as well as possible. The success rate of our algorithm is independent of the base image, because our algorithm aligns the other images to the first enrolled image from a sequence. The mosaicking success rate of our algorithm is higher than Lee's algorithm because our enrollment scheme can assure that the common area between images exists when a user slides his finger properly. In our algorithm 12 sequences among 100 sequences fail to be mosaicked because of the low quality of images, the severe plastic distortion and blurred images. Especially if the number of blurred images from a sequence is too large, the sequence is hard to be mosaicked because the common area between images becomes small.

Table 1 shows that the images captured by the dab approach are hard to be mosaicked due to the small common area between images. To show that our enrollment scheme can acquire a wider area of a fingerprint than the dab approach, we measured the average size of the foreground area and the average number of minutiae from an image enrolled by the dab approach, the mosaicked image of Lee's algorithm, a rolled image and the mosaicked image of our algorithm as shown in Table. 2. In Table. 2 the size is the number of pixels of a foreground area. Table. 2 shows that the average foreground area of the images mosaicked by our algorithm is smaller than that of the rolled image but larger than that of the images mosaicked by Lee's algorithm.

6 Conclusions and Future Works

We have described a new enrollment scheme and a non-minutiae based mosaicking algorithm for the sequential fingerprint images. Our enrollment scheme has the larger foreground area than that of Lee's algorithm and also the mosaicking success rate of our algorithm is higher than Lee's algorithm because Lee's algorithm doesn't consider the plastic distortion and the common area between images enrolled by the dab approach doesn't exist or is too small to be mosaicked. The algorithms which align images by using minutiae or a core point, may show the low mosaicking success rate in case that captured images have small number of minutiae or no core point [1],[2].

We can aquire the wide foreground area of a fingerprint but the image qualities of our mosaicked images are worse than the rolled images because of the

accumulation of the registraion error, the severe plastic distortion and the motion blurring. In the future, we will try to solve those problems. Also to obtain the foreground area as large as that of a rolled image, we will try to combine the mosaicked image enrolled by rolling and sliding horizontally with the mosaicked image enrolled by sliding vertically.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through Biometrics Engineering Research Center at Yonsei University.

References

1. A.K. Jain and A. Ross, "Fingerprint mosaiking" *Proc. International Conference on Acoustic Speech and Signal Processing(ICASSP)*, vol. 4, pp.4064-4067, 2002
2. Dongjae Lee, Sanghoon Lee, Kyoungtaek Choi and Jaihie Kim, "Fingerprint fusion based on minutiae and ridge for enrollment" *LNCS on Audio-and Video-Based Biometric Person Authentication*, vol.2688, pp.478-485, Jun. 2003
3. N.K. Ratha, J.H. Connell and R.M Bolle, "Image mosaicing for rolled fingerprint construction" *Pattern Recognition, Proceedings. Fourteenth International Conference on*, vol.2 , pp.1651-1653, Aug. 1998
4. Lin Hong, Yifei Wan and A.K. Jain, "Fingerprint Image Enhancement Algorithm and Performance Evaluation" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.20, No. 8, pp.777-789, Aug. 1998
5. NK Chern, PA Neow and MH Ang Jr, "Practical issues in pixel-based autofocusing for machine vision" *Int. Conf. On Robotics and Automation*, pp.2791- 2796, 2001
6. Yong-Sheng Chen, Yi-Ping Hung and Chiou-Shann Fuh, "Fast block matching algorithm based on the winner-update strategy" *IEEE Transactions on Image Processing*, vol. 10, No. 8, pp.1212- 1222, Aug. 2001
7. George Wolberg, "Digital image warping" *IEEE Computer Society Press*, 1988
8. http://www.hbs-jena.com

Minutiae Matching Based Fingerprint Verification Using Delaunay Triangulation and Aligned-Edge-Guided Triangle Matching

Huimin Deng and Qiang Huo

Department of Computer Science, The University of Hong Kong, Hong Kong
{hmdeng,qhuo}@cs.hku.hk

Abstract. This paper presents a novel minutiae matching approach to fingerprint verification. Given an input or a template fingerprint image, minutiae are extracted first. Using Delaunay triangulation, each fingerprint is then represented as a special connected graph with each node being a minutia point and each edge connecting two minutiae. Such a graph is used to define the neighborhood of a minutia that facilitates a local-structure-based matching of two minutiae from input and template fingerprints respectively. The possible alignment of an edge in input graph and an edge in template graph can be identified efficiently. A global matching score between two fingerprints is finally calculated by using an aligned-edge-guided triangle matching procedure. The effectiveness of the proposed approach is confirmed by a benchmark test on FVC2000 and FVC2002 databases.

1 Introduction

Fingerprint matching is a difficult problem that has been studied for several decades. Among many approaches proposed, minutiae-based approach remains the most popular one [12]. For this type of approaches, minutiae (typically ridge endings or ridge bifurcations) are extracted first from given input and template fingerprint images. The fingerprint matching problem is then cast as a problem of matching two sets of planar point patterns. Again, many approaches have been proposed (e.g., [6, 7, 14, 15] and many references in [12]). Among them, we are particularly interested in those approaches that take advantage of strength offered by 1) using a local-structure-based matching for an efficient pre-alignment of two fingerprints or an early rejection of very different fingerprints, and 2) using a global minutiae matching strategy to consolidate the result of local matching and derive a global matching score of two fingerprints.

For example, in [7], the local structure is formed by the concerned minutia and its k-nearest neighbor minutiae (k=2 in practice). Local minutia matching consists of comparing two local minutiae structures characterized by attributes that are invariant with respect to global transformation such as translation, rotation, etc. The best matching minutiae pair is then selected and used for registering the two fingerprints. In the global matching stage, the remaining

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 270–278, 2005.
© Springer-Verlag Berlin Heidelberg 2005

aligned pairs are compared and a final score is computed by taking into account the different contributions from the above two matching stages. In [15], the local structure is defined more formally by using a graph notation (i.e., a *star*). All the minutiae within a pre-specified distance from the concerned minutia are treated as its neighbors. Some follow-up works of the above two approaches are reported in literature. For example, in [16], core points are used as reference points to speed up the initial local structure matching. In [9], more minutiae pairs are used as reference pairs to guide the consolidation step that improves robustness when the best-matching minutiae pair is unreliable.

Inspired by the above works, in this paper, we propose a new minutiae matching approach that also uses both local and global structures, but is different from the previous works in the following aspects:

- The neighborhood of a minutia is defined by the result of Delaunay triangulation of minutiae;
- In the local structure matching, instead of finding best-matching minutiae pair(s), we try to find the possible best-matching edge pairs;
- A global matching score between two fingerprints is calculated by using an aligned-edge-guided triangle matching procedure.

In the following section, we describe the details of our proposed approach. In Section 3, we report the benchmark test results on FVC2000 and FVC2002 databases. Finally, we conclude the paper in Section 4.

2 Our Approach

2.1 Minutiae Extraction

Given an input or a template fingerprint image, minutiae are extracted first. In literature, there are mainly two kinds of minutiae extraction approaches: one will go through enhancement, binarization and thinning, and minutiae extraction (e.g., [5, 6]); the other extracts the minutia directly from gray scale image (e.g., [8, 10]). It was reported in literature and observed in our preliminary experiments that the former approach may create some spurious minutiae, while the latter may miss some genuine minutiae. To obtain the main minutiae structure, we adopted an approach similar to the ones in [5, 6] for minutiae extraction. Each extracted minutia, M_k, is represented as a feature vector:

$$M_k = (x_k, y_k, \theta_k, t_k)^T$$

where x_k, y_k are its location coordinates, θ_k is its orientation defined as the local ridge orientation of the associated ridge, and t_k is the minutia type (0 for ridge ending, 1 for ridge bifurcation). The original ridge orientation is in the range of $[0, \pi)$, but is mapped to the range of $[0, 2\pi)$ as described in [7]. Ridge count between two minutiae in a fingerprint is recorded for future use.

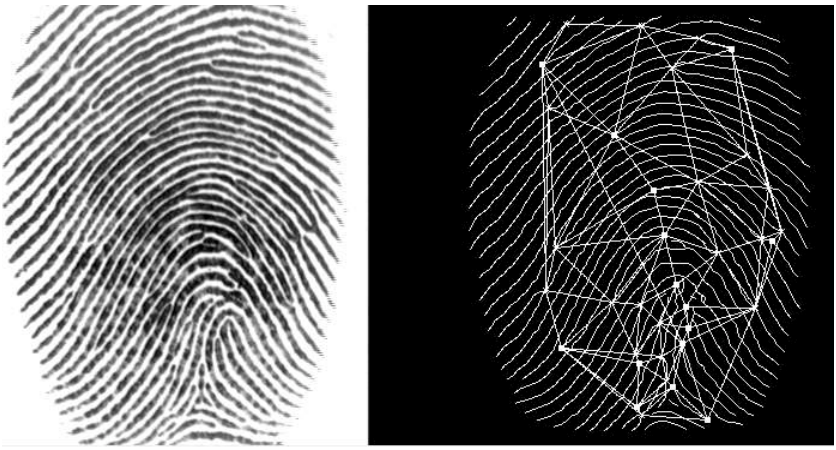


Fig. 1. A fingerprint image (left) and its representation using the Delaunay triangulation of the minutiae (right)

2.2 Delaunay Triangulation

After the above minutiae extraction step, we obtain two sets of minutiae for the input and template images respectively. Using Delaunay triangulation (e.g., [1, 2, 13]), each fingerprint can be represented as a special connected graph with each node being a minutia point and each edge connecting two minutiae. Figure 1 gives an example of a fingerprint image and its representation using the Delaunay triangulation of the minutiae. Delaunay triangulation has certain desirable properties, including 1) the Delaunay triangulation of a non-degenerate set of N minutiae points is unique, can be computed efficiently in $O(N \log N)$ time, and produces $O(N)$ triangles, 2) missing or spurious minutiae points affect the Delaunay triangulation only locally, and 3) the triangles obtained are not "skinny".

As described in the next subsection, such a graph representation can be used to define the neighborhood of a minutia that facilitates a local-structure-based matching of two minutiae from input and template fingerprints respectively. It also constrains the number of edges to be examined.

2.3 Using Local-Structure-Based Matching to Find Possible Alignment of Edge Pairs

Our local-structure-based matching algorithm is inspired by [7], but we use it in a different way. In our approach, the local structure is formed by the concerned minutia and its adjacent minutiae in Delaunay triangulation representation. An example is illustrated in Fig. 2. Given such a local structure, the relation between the minutia M_k and one of its neighboring minutia M_j can be characterized by the following feature vector:

$$V_{kj} = (d_{kj}, \theta_{kj}, \phi_{kj}, rc_{kj}, t_k, t_j)^T \tag{1}$$

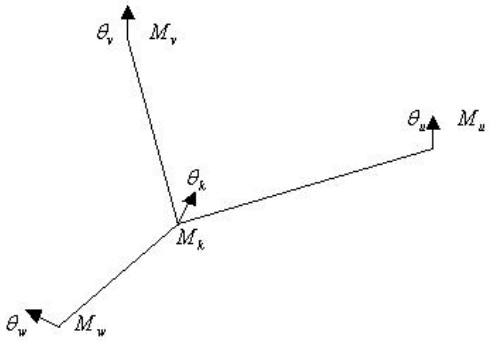


Fig. 2. The local structure of a minutia M_k

where $d_{kj} = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}$ is the distance between minutiae M_k and M_j, $\theta_{kj} = \theta_k - \theta_j$ is the direction difference between the orientation angles θ_k and θ_j of M_k and M_j, $\phi_{kj} = \arctan(\frac{y_k - y_j}{x_k - x_j}) - \theta_k$ is the direction difference between the orientation angle θ_k of M_k and the direction of the edge connecting M_k to M_j, rc_{kj} is the ridge count between M_k and M_j, t_k and t_j are the minutiae types of M_k and M_j respectively. Note that the above definition of d_{kj}, θ_{kj}, ϕ_{kj}, rc_{kj} applies to any directed edge connecting a minutia M_k to another minutia M_j, and will be used in the next subsection.

Let's use M_p^I and M_q^T to denote the minutiae in input and template images respectively, and use N_p^I and N_q^T to denote the corresponding numbers of neighboring minutiae. We define a similarity measure between an edge $\overrightarrow{pi}$ in the input fingerprint and an edge $\overrightarrow{qj}$ in the template fingerprint as follows:

$$SE(\overrightarrow{pi}, \overrightarrow{qj}) = \begin{cases} \frac{TH_1 - W^T|V_{pi} - V_{qj}|}{TH_1} & \text{if } W^T|V_{pi} - V_{qj}| < TH_1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where W is a weighting vector, TH_1 is a threshold, M_i^I and M_j^T are neighboring minutiae of M_p^I and M_q^T respectively. Given two local structures with central minutiae of M_p^I and M_q^T, we can calculate a $N_p^I \times N_q^T$ similarity score matrix $SE(\cdot, \cdot)$. By following the same strategy in [7], $SE(\cdot, \cdot)$ is modified according to the following rule:

$$\begin{aligned} &If\ \ \exists k\{SE(\overrightarrow{pk}, \overrightarrow{qj}) > SE(\overrightarrow{pi}, \overrightarrow{qj})\}\ or\ \exists k\{SE(\overrightarrow{pi}, \overrightarrow{qk}) > SE(\overrightarrow{pi}, \overrightarrow{qj})\}, \\ &\quad then,\ SE(\overrightarrow{pi}, \overrightarrow{qj}) = 0;\ Otherwise,\ no\ change\ is\ made. \end{aligned}$$

The matching score of two local structures with central minutiae of M_p^I and M_q^T is then defined as follows:

$$SL(M_p^I, M_q^T) = \frac{\sum_{i=1}^{N_p^I} \sum_{j=1}^{N_q^T} SE(\overrightarrow{pi}, \overrightarrow{qj})}{\min\{N_p^I, N_q^T\}} \quad . \tag{3}$$

This score will be used in the following procedure to find the possible matching edges in input and template fingerprints respectively:

Step 1: Initialization of Edge-Pair Set:
If the number of extracted minutiae in a fingerprint image is below a threshold TH_{num}, all the possible edges connecting two minutiae are considered; otherwise, only those edges in Delaunay triangulation are considered. Furthermore, remove those edges with a length less than a threshold TH_{edge}. Consequently, two sets of initial edges, E_I and E_T are formed for input and template fingerprints respectively. Therefore, the initial set of possible aligned edge pairs is $EP = E_I \times E_T$.

Step 2: Edge-Pair Pruning:
Remove edge pair $\{(\overrightarrow{pq}, \overrightarrow{ab}) | \overrightarrow{pq} \in E_I, \overrightarrow{ab} \in E_T\}$ from EP, if any of the following six conditions is satisfied:

$$|d_{pq} - d_{ab}| > TH_d \tag{4}$$
$$|\theta_{pq} - \theta_{ab}| > TH_\theta \tag{5}$$
$$|\phi_{pq} - \phi_{ab}| > TH_\phi \tag{6}$$
$$|rc_{pq} - rc_{ab}| > TH_{rc} \tag{7}$$
$$SL(M_p^I, M_a^T) < TH_{SL} \tag{8}$$
$$SL(M_q^I, M_b^T) < TH_{SL} \tag{9}$$

where TH_d, TH_θ, TH_ϕ, TH_{rc}, TH_{SL} are thresholds to be set empirically.

Step 3: Termination:
The remaining edge pairs in EP are considered as possible aligned edges.

2.4 Aligned-Edge-Guided Triangle Matching

Using each pair of possible aligned edges in EP as reference edges, a matching score between input and template fingerprints can be calculated by using the following *Aligned-Edge-Guided Triangle Matching* procedure:

Step 1: Sort other minutiae in each fingerprint in ascending order of their angles subtended by the reference edge with the edges connecting the initial minutia point in the reference edge to the minutiae concerned. An example is illustrated in Fig. 3. In this figure, we use $\overrightarrow{AB}$ to denote the reference edge in the input fingerprint, and $\overrightarrow{A'B'}$ to denote the reference edge in the template fingerprint. Minutiae in both fingerprints are sorted with respect to the reference edges as $\{C, D, \cdots\}$ and $\{C', D', \cdots\}$ respectively.

Step 2: Connect other minutiae in each fingerprint with the minutiae of the reference edge to form triangles as illustrated in Fig. 3.

Step 3: For each unexamined pair of triangles, say, $\triangle ABC$ and $\triangle A'B'C'$ as illustrated in Fig. 3, do the following:

- If $|\angle ABC - \angle A'B'C'| < TH_{ang}$, set the matching score of two triangles, $ST(\triangle ABC, \triangle A'B'C')$ to zero, i.e., $ST(\triangle ABC, \triangle A'B'C') = 0$, and go to **Step 4**; Otherwise,

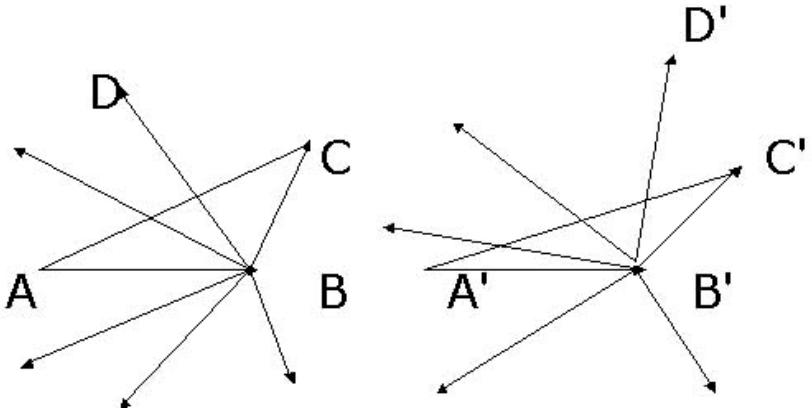


Fig. 3. Aligned-edge-guided triangle matching

– If $|d_{AC} - d_{A'C'}| \leq TH_d$ and $|\theta_{AC} - \theta_{A'C'}| \leq TH_\theta$ and $|\phi_{AC} - \phi_{A'C'}| \leq TH_\phi$ and $|rc_{AC} - rc_{A'C'}| \leq TH_{rc}$ and $|d_{BC} - d_{B'C'}| \leq TH_d$ and $|\theta_{BC} - \theta_{B'C'}| \leq TH_\theta$ and $|\phi_{BC} - \phi_{B'C'}| \leq TH_\phi$ and $|rc_{BC} - rc_{B'C'}| \leq TH_{rc}$, set

$$ST(\triangle ABC, \triangle A'B'C') = 0.5 + 0.5SL(C, C'),$$

and go to **Step 4**; Otherwise, set $ST(\triangle ABC, \triangle A'B'C') = 0$, and go to **Step 4**;

Step 4: If all possible pairs of triangles have been examined, go to **Step 5**; Otherwise, go to **Step 3**.

Step 5: If $\exists P\{ST(\triangle ABC, \triangle A'B'P) > ST(\triangle ABC, \triangle A'B'C')\}$ or $\exists P\{ST(\triangle ABP, \triangle A'B'C') > ST(\triangle ABC, \triangle A'B'C')\}$, set $ST(\triangle ABC, \triangle A'B'C') = 0$.

Step 6: The matching score between input and template fingerprints based on the aligned edge pair $(\overrightarrow{AB}, \overrightarrow{A'B'})$ is calculated as follows:

$$SG(\overrightarrow{AB}, \overrightarrow{A'B'}) = \frac{1}{MN}\{2 + \sum_{\{P \neq A,B\}} \sum_{\{P' \neq A',B'\}} ST(\triangle ABP, \triangle A'B'P')\} \quad (10)$$

where M and N are the total number of minutiae in the input and template fingerprints respectively.

Once the above *Aligned-Edge-Guided Triangle Matching* procedure has been applied to all possible pair of aligned edges in EP, we can calculate the final matching score of the input and template fingerprints as follows:

$$Matching\ Score(input, template) = \max_{\{(\overrightarrow{AB}, \overrightarrow{A'B'}) \in EP\}} SG(\overrightarrow{AB}, \overrightarrow{A'B'}) \ . \quad (11)$$

Table 1. A summary of benchmark testing performance in terms of equal error rate (EER in %) of our algorithm on FVC2000 and FVC2002 databases, and its comparison with that achieved by the best performing algorithms from other academic institutions in FVC2000 and FVC2002 competitions

Benchmark Databases		EER (in %)	
		Our Algorithm	Reference Algorithm
FVC2000	DB1	2.80	7.60
	DB2	2.75	2.75
	DB3	7.46	5.36
	DB4	2.86	5.04
FVC2002	DB1	1.82	2.36
	DB2	1.52	2.35
	DB3	4.94	6.62
	DB4	2.29	3.70

3 Experimental Results

We have performed a benchmark test of our proposed fingerprint matching algorithm on FVC2000 [3, 11] and FVC2002 [4] databases by following exactly the protocol specified by the FVC2000 and FVC2002 organizers. Table 1 summarizes the equal error rates (EER in %) achieved by our algorithm on different sub-corpora of the above two databases. Tables 2 and 3 summarize a snapshot of the false acceptance rate (FAR in %) and false rejection rate (FRR in %) at several operational points achieved by our algorithm on FVC2000 and FVC2002 respectively. Note that we used the following single setting of control parameters for all the experiments:

- In Eq. (2), $TH_1 = 36$, $W = (1, 0.3 * 180/\pi, 0.3 * 180/\pi, 3, 6, 6)^T$;
- In the procedure of finding the possible matching edges described in Section 2.3, $TH_{num} = 20$, $TH_{edge} = 15$, $TH_d = 8$, $TH_\theta = \pi/6$, $TH_\phi = \pi/6$, $TH_{rc} = 3$, $TH_{SL} = 0.2$;
- In the procedure of *Aligned-Edge-Guided Triangle Matching* described in Section 2.4, $TH_{ang} = \pi/6$. TH_d, TH_θ, TH_ϕ, TH_{rc}, TH_{SL} are set as the above.

For comparison, in Table 1, we also quote the best performance achieved on the same databases by research groups from academic institutions who participated FVC2000 and/or FVC2002 competitions [3, 4]. The performance of our algorithm is very encouraging. The "Response Time" for matching an input and a template fingerprint, excluding time for minutiae extraction, is, on the average, approximately 0.079 seconds on a Pentium III 933MHz notebook running Windows XP OS under a normal work load.

4 Summary

We have proposed a novel minutiae matching approach to fingerprint verification. Given an input or a template fingerprint image, minutiae are extracted

Table 2. A summary of the false acceptance rate (FAR in %) and false rejection rate (FRR in %) at several operational points achieved by our algorithm on FVC2000

DB1	FAR	0.38	0.87	2.70	8.59	26.77
	FRR	5.93	4.50	2.89	1.93	0.89
DB2	FAR	0.20	0.69	1.94	5.82	18.55
	FRR	5.75	4.43	3.46	2.57	1.29
DB3	FAR	1.98	3.33	6.53	11.90	28.93
	FRR	11.93	10.32	8.39	6.50	3.96
DB4	FAR	0.38	1.17	2.54	5.03	11.23
	FRR	6.07	4.29	3.14	2.29	1.32

Table 3. A summary of the false acceptance rate (FAR in %) and false rejection rate (FRR in %) at several operational points achieved by our algorithm on FVC2002

DB1	FAR	0.14	0.38	2.18	6.22	28.44
	FRR	5.18	3.50	1.45	1.39	0.54
DB2	FAR	0.02	0.38	1.80	8.34	34.81
	FRR	4.18	2.39	1.25	0.68	0.18
DB3	FAR	0.36	1.43	4.12	20.30	60.38
	FRR	10.93	7.82	5.75	2.89	1.18
DB4	FAR	0.24	0.75	1.72	7.03	29.72
	FRR	5.64	4.00	2.86	1.36	0.64

first. Using Delaunay triangulation, each fingerprint is then represented as a special connected graph with each node being a minutia point and each edge connecting two minutiae. Such a graph is used to define the neighborhood of a minutia that facilitates a local-structure-based matching of two minutiae from input and template fingerprints respectively. The possible alignment of an edge in input graph and an edge in template graph can be identified efficiently. A global matching score between two fingerprints is finally calculated by using an aligned-edge-guided triangle matching procedure. The effectiveness of the proposed approach is confirmed by a benchmark test on FVC2000 and FVC2002 databases. Our future works include 1) improving minutiae extraction algorithm, especially for low quality fingerprint images, 2) refining our matching algorithm to speed up the matching process, 3) improving the robustness of our matching algorithms, 4) investigating the sensitivity of different settings of control parameters on matching performance, 5) investigating in depth the pros and cons of our approach *versus* other approaches, 6) conduct benchmark testing on more databases.

References

1. G. Bebis, T. Deaconu, and M. Georgiopoulos, "Fingerprint identification using Delaunay triangulation," in *Proc. of Int. Conf. on Information Intelligence and Systems* 1999, pp.452-459.

2. A. V. Ceguerra and I. Koprinska, "Integrating local and global features in automatic fingerprint verification," in *Proc. of ICPR-2002*, 2002, pp.III-347-350.
3. FVC2000, http://bias.csr.unibo.it/fvc2000/
4. FVC2002, http://bias.csr.unibo.it/fvc2002/
5. L. Hong, Y. Wan, and A. Jain, "Fingerprint image enhancement: algorithm and performance evaluation," *IEEE Trans. on PAMI*, Vol. 20, No. 8, pp.777-789, 1998.
6. A. K. Jain, L. Hong and R. Bolle, "On-line fingerprint verification," *IEEE Trans. on PAMI*, Vol. 19, No. 4, pp.302-314, 1997.
7. X. Jiang and W.-Y. Yau, "Fingerprint minutiae matching based on the local and global structures," *Proc. of ICPR-2000*, 2000, pp.1038-1041.
8. X. Jiang, W.-Y. Yau, and W. Ser, "Detecting the fingerprint minutiae by adaptive tracing the gray-level ridge," *Pattern Recognition*, Vol. 34, No. 5, pp.999-1013, 2001.
9. D. Lee, K. Choi, and J. Kim, "A robust fingerprint matching algorithm using local alignment," in *Proc. of ICPR-2002*, 2002, pp.III-803-806.
10. D. Maio and D. Maltoni, "Direct gray-scale minutiae detection in fingerprints," *IEEE Trans. on PAMI*, Vol. 19, No. 1, pp.27-40, 1997.
11. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "FVC2000: fingerprint verification competition," *IEEE Trans. on PAMI*, Vol. 24, No. 3, pp.402-412, 2002.
12. D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, Springer, 2003
13. G. Parziale and A. Niel, "A fingerprint matching using minutiae triangulation," *Proc. of First International Conference on Biometric Authentication*, Hong Kong, 2004, pp.241-248.
14. N. K. Ratha, K. Karu, S. Chen, and A. K. Jain, "A real-time matching system for large fingerprint databases," *IEEE Trans. on PAMI*, Vol. 18, No. 8, pp.799-813, 1996.
15. N. K. Ratha, V. D. Pandit, R. M. Bolle, and V. Vaish, "Robust fingerprint authentication using local structural similarity," in *Proc. of 5th IEEE Workshop on Applications of Computer Vision*, 2000, pp.29-34.
16. W. Zhang and Y. Wang, "Core-based structure matching algorithm of fingerprint verification," in *Proc. of ICPR-2002*, 2002, pp.I-70-74.

An Asymmetric Fingerprint Matching Algorithm for Java Card™

Stefano Bistarelli[1,2], Francesco Santini[2], and Anna Vaccarelli[2]

[1] Dipartimento di Scienze, Università "G. D'Annunzio" di Chieti-Pescara, Italy
bista@sci.unich.it
[2] Istituto di Informatica e Telematica, CNR, Pisa, Italy
{stefano.bistarelli,francesco.santini,anna.vaccarelli}
@iit.cnr.it

Abstract. A novel fingerprint matching algorithm is proposed in this paper. The algorithm is based on the minutiae local structures, that are invariant with respect to global transformations like translation and rotation. The match algorithm has been implemented inside a smartcard over the Java Card™ platform, meeting the individual's need for information privacy and the overall authentication procedure security. The main characteristic of the algorithm is to have an asymmetric behaviour, in respect to the execution time, between correct positive and negative matches. The performances in terms of authentication reliability and speed have been tested on some databases from the Fingerprint Verification Competition 2002 (FVC2002). Moreover, our procedure has shown better reliability results when compared with related Java Card™ algorithms.

1 Introduction

The term "biometrics" is commonly used today to refer to the authentication of a person by analyzing the physical characteristics (like fingerprints) or behaviour characteristics (like voice or gait). Fingerprint matching is one of the most diffused biometric techniques used in automatic personal identification or verification, because of its strong reliability and its low implementation cost.

Performing a biometric verification inside a smart card is notoriously difficult, since the templates tend to eat-up a large part of the card's memory, while the biometric verification algorithms are almost beyond the processing capabilities of standard processors. With *Match On Card* (MOC) technology the fingerprint template is stored within the card, unavailable to the external applications and the outside world. In addition the matching decision is securely authenticated internally by the smartcard itself: in this way, the card only trusts itself for eventually unblocking stored sensitive information, such as digital certificates or private keys for digital signature. Our verification MOC algorithm has been developed to work in this very strictly bounded environment.

The algorithm is based on some minutiae characteristics (ridge pattern micro characteristics) and more precisely on their local structure information, so there is no need to pre-align the processing fingerprint templates, that would be a difficult task to implement inside a smartcard. Moreover it shows an asymmetric execution time between correct positive matches (same fingerprint) and correct negative matches (two different

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 279–288, 2005.
© Springer-Verlag Berlin Heidelberg 2005

fingers), and this because the match procedure stops immediately when few minutiae pairs result in a positive match. If this check doesn't succeed, for example if the two fingers are different or if the two acquisitions of the same finger are very disturbed, the procedure is fully executed lasting longer.

2 Background

The most evident structural characteristic of a fingerprint is the pattern of interleaved ridges and valleys that often run in parallel; at local level, other important features called *minutiae* refer to ridge discontinuities. Most frequently the minutiae types can be individuated by terminations, where a ridge line ends, and bifurcations, where a ridge bifurcates forming a "Y". The minutiae can be used in fingerprint matching since they represent unique details of the ridge flow and are considered as a proof of identity.

The template, in its generic definition, is a mathematical representation of the fingerprint "uniqueness" to be used later during the matching phase: the template acquired during enrollment is defined as the *reference template* and it is in some way associated with the system user identity, while the template acquired during the verification phase is defined as the *candidate template*.

Matching the templates represents an extremely difficult problem because of the variability in different impressions of the same fingers; most important affecting factors introduced during image acquisition are the *displacement* and the *rotation* depending on the different positioning of the finger on the acquisition sensor, *non-linear distortions* due to the skin plasticity and *partial overlap*, since a part of the fingerprint can fall outside of the acquisition area and therefore different samples of the same finger could correspond only on a smaller area.

The algorithms used to resolve fingerprint matching can be classified [1] in three main branches: *correlation-based* [6], where the match is performed by superimposing two fingerprint images and computing the correlation between corresponding pixels (in this case the template is directly the finger image); *minutiae-based*, whose theory is fundamentally the same as for manual fingerprint examiners, and *ridge feature-based* [5], where the fingerprints are compared in terms of ridge pattern features other than the minutiae or pixels intensity, like texture information or sweat pores.

Focusing on the minutiae based algorithms, the match procedure essentially consists in finding the maximum number of corresponding minutiae between two templates; this problem can be addressed also as a more general *point pattern matching* problem [13]. We can subdivide minutiae matching class into two more branches: *global minutiae matching* [3] requires a first fingerprint alignment phase that subsequently permits to match the aligned templates. In *local minutiae matching* [4] two fingerprints are instead compared according to their local minutiae structures, which are characterized by attributes invariant with respect to global transformations such as translations or rotations. Local matching supplies simplicity, low computational complexity and higher distortion tolerance, while a global matching grants a high distinctiveness.

Regarding smartcard related work, in [10] is described a very simple $O(n^2)$ matching algorithm, where n is the minutiae number in one template. For a given minutia in the reference template, the algorithm finds all the minutiae in the candidate template

for which the distance between position coordinates and the difference in orientation angles are below the predefined thresholds. If more than one can be matched with the same reference minutia, this conflict is resolved by choosing the geometrically nearest.

One specific algorithm for fingerprint matching on the Java Card™ platform, using a feature extraction environment similar to ours, is described in [11]; it uses two distinct algorithms on different feature types (hybrid matcher) and at the end the overall score is calculated as a linear combination of the two independent sub-scores. The first algorithm is based on the minutiae features and a graph structure is built starting from the core point position in the fingerprint, then visiting the neighbor minutiae; the matching procedure has been inspired from the point-pattern matching algorithm in [12] and its purpose is to find a *spanning ordered tree* touching as many nodes as possible in the two graphs. The second algorithm is a ridge feature-based and has been implemented as described in [5]; this algorithm is very fast and can be easily implemented on a smartcard, since the match consists only in finding the euclidean distance between two feature vectors.

3 Our Matching Algorithm

3.1 Features

In our algorithm implementation we have adopted the NIST Fingerprint Image Software (NFIS) [2], an open source toolkit which includes the MINDTCT minutiae data extractor used to extract the minutiae from a given fingerprint image. We have used this information to derive additional features directly used in our matching algorithm; these features are computed for each minutia in respect to its neighbors, and so each neighbor is described by the following four features (see also Fig. 1):

- the euclidean distance between the central minutia and its neighbor minutia (segment *D* in Fig. 1); referred to as Ed in the rest of the paper.
- the angle between segment *D* and the central minutia ridge direction (angle α in Fig. 1); latterly referred to as Dra.
- the difference angle between central minutia and neighbor ridge orientation angle ($\theta_1 - \theta_2$ in Fig. 1); latterly referred to as Oda.
- the ridge count between central and neighbor minutiae: given two points a and b, the ridge count between them is the number of ridges intersected by the segment $\overline{ab}$ (in Fig. 1 ridge count value is 1); latterly referred to as Rc.

Choosing the maximum number of the neighbors is very important for the system reliability performances (but in contrast with the matching speed), and so we have decided to increase this number from the default MINDTCT value (5) to the new value of 8. We have also modified the MINDTCT C source code to consider only the neighbors with a minimum reliability threshold: the modified MINDTCT finds for every minutia its eight nearest neighbors in respect to the euclidean distance, with a good reliability estimation given by a predefined threshold value. If the number of neighbors for a minutia found in this way is low (i.e. less than 5), then the neighbors are searched again with a lower reliability threshold (the reliability evaluation is found by MINDTCT); we have introduced all these changes to build a "good" neighborhood with more information, enough to face the possible lack of some minutiae in the template.

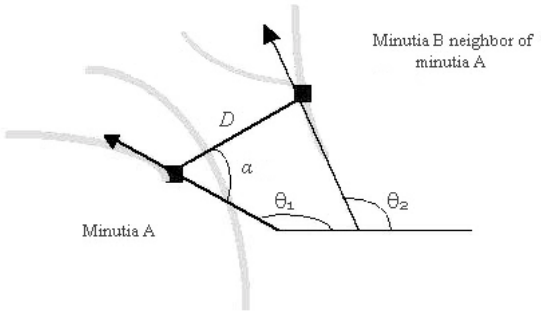


Fig. 1. Features graphical description

3.2 Algorithm Description

Our proposed matching algorithm computes how much the neighborhood of a minutia in the candidate template is similar to the neighborhood of each minutia in the reference template; after this scan, the two most similar minutiae are matched and then discarded from subsequent scan phases concerning other different minutiae of the candidate template. All these similarity measures are summed together during the process, and at the end the algorithm can decide if the two templates match by applying a threshold on this global score. Our procedure is based on the minutiae local structures (Section 2).

But as said before, matching on smartcard environment is bounded by the low computational complexity due to the hardware simplicity (CPU limitations first of all), and thus waiting for a complete minutiae match could lead to a waiting time too long for the user. In our algorithm we solve this problem by stopping the computation as soon as it is possible to assert, with satisfactory confidence, that the considered templates belong to the same fingerprint. To realize this improvement, our algorithm stops as soon as it finds some minutiae pairs (i.e. a number between 2 and 5) matching with a very good similarity measure, or even promptly when only the last examined minutiae pair has a matching value less than a very rigourous threshold; otherwise, if these two conditions are not fulfilled, the algorithm explores all the minutiae pairings space. This relaxation showed a very good security performance in our tests and provided an evident speed improvement in the matching decisions regarding positive matches (Section 4). The delay for unsuccessful matches scanning all the minutiae list is not of much interest, because it is clearly more important to gain a high execution speed while verifying the true card-owner identity than quickly rejecting an impostor!

As input, the matching procedure receives both the neighbor features information for the one by one candidate minutia to be matched, and the entire reference template. The algorithm scans sequentially the minutiae of the reference template until a good match for the input minutia is found (reference 1 in Fig 2). Both candidate and reference minutiae lists are stored according to the increasing minutia reliability value: in this way we try to stop the procedure more quickly by scanning a reduced portion of the template minutiae lists, since a minutia with a high reliability in a given template, if not cut away by partial overlapping (Section 2), will have probably a high reliability also in other templates obtained from the same finger. So the stopping conditions can be met earlier than in a casual disposition of the minutiae in the list. Moreover, it is obviously better

to prematurely stop the procedure with few but "good" minutiae than with low quality ones. The chosen matching minutia in the reference template is then marked as "already matched" and it is not considered in the successive iterations.

To compute the dissimilarity between two minutiae in different templates, the algorithm uses the information about the neighbor features and executes the following four steps in sequence (4 in Fig 2):

1. To find the difference in absolute value between corresponding features: $EdDiff$ $=\mid Ed_1 - Ed_2 \mid$, $rcDiff =\mid Rc_1 - Rc_2 \mid$, $draDiff =\mid Dra_1 - Dra_2 \mid$ and $odDiff =\mid Oda_1 - Oda_2 \mid$.
2. To check that every feature difference value is below the corresponding acceptance threshold; if only one difference value exceeds the relative threshold, the two neighbors cannot correspond in the two neighborhoods ($edDiff$ must not be greater than $edDiffThr$, $rcDiff$ than $rcThr$, $edDiff$ than $draThr$ and $odDiff$ than $odThr$). The set of the four feature difference thresholds can be globally defined as the features *bounding box*, which makes the algorithm tolerant to small non-linear distortions.
3. To multiply each feature difference for the relative weight value: $edWghtDiff = edDiff * edWght$, $rcWghDiff = rcDiff * rcWght$, $odWghtDiff = odDiff * odWght$ and $draWghtDiff = draDiff * draWght$. The different weight values are necessary to attribute more importance to the features that match better, for example the euclidean distance. To obtain each weight value, we have also divided by the respective feature difference bounding box threshold, since we want these differences to be normalized and homogenous.
4. To sum together all the four weighted differences to represent the global dissimilarity between the two neighbors: $NeighDissimilarity = edWghtDiff + rcWghtDiff + draWghtDiff + odWghtDiff$.

Following these steps, the algorithm finds for the first neighbor (in the casual neighborhood order) of the reference minutia, the most similar neighbor in the input minutia among those satisfying the bounding box checks; the most similar is the one for which the algorithm finds the lowest *NeighDissimilarity* value. The chosen most similar neighbor in the reference minutia is then marked and not considered while matching other neighbors. The obtained *NeighDissimilarity* value is then added to the global similarity score between the minutiae, *MinDissimilarity*. The procedure is repeated exactly for all the other neighbors (excluding the already marked ones, 3 in Fig 2) or until the required minimum number *N* (i.e. 4) of neighbors is matched. At the end of the two neighborhoods scanning (at the end of the *for*, 2 in Fig 2), if the procedure has found less than *N* matching neighbor pairs between the two minutiae (6 in Fig 2), then these two minutiae are not considered as matching because their neighborhoods agree on too few evidences to be a reliable matching minutiae pair, even if the *NeighDissimilarity* value is very low. At the same time, this procedure stops immediately as we match the previous *N* threshold value of neighbors (5 in Fig 2), because we have seen that stopping before the whole neighborhood scan is sufficient to grant a good reliability and, meanwhile, the match time is considerably speeded up.

The *MinDissimilarity* score between the minutiae is finally divided by the number of matched neighbor pairs and then added to the global dissimilarity value between the

```
{MINUTIAE MATCHING PROCEDURE}

   - Input:* one candidate template minutia m1;
            * minutiae list of the reference template;

1  For each minutia m2 in reference template not yet matched{
2    For each neighbor n2 of minutia m2 {
       - MinDiff = upperLimit;
       - ChosenNbr= null;
3      For each not already matched neighbor n1 of m1 {
4        - Executes the four steps between the n1-n2
           corresponding features (directly processes next n1
           if the bounding box rejects the controls);
         If (NeighDissimilarity < MinDiff) {
           - MinDiff = NeighDissimilarity;
           - ChosenNbr = n1; }
       }
       If (ChosenNbr != null) {
         - ChosenNbr is marked as "matched";
         - MinDissimilarity += MinDiff;
         - number of matched neighbors NM= NM + 1; }
5      If (NM > N)
         - m1 and m2 are "matched": break from this For;
     }
6    If (NM < N)
       - Continue with the next minutia m2
     else {
7      - m1 and m2 are "matched": TemplDissimilarity+=
         (MinDissimilarity \ NM);
       - break from this For;
     }
   }
   - m1_m2_MatchCost = MinDissimilarity \ NM;
   If (m1 and m2 are "matched") {
     - MinutiaeNMatched++;
     - Mark reference minutia m2 as "matched";
8    If (m1_m2_MatchCost < VeryOptValue)
       - STOP: the match is accepted;
     If (m1_m2_MatchCost < OptValue)
       - OptMinNumber++;
9    If (OptMinNumber == OptNumberThreshold)
       - STOP: the match is accepted;
   }
   - Process another minutia m1 if no stopping condition
     has occurred or if m1 and m2 are not "matched";
```

Fig. 2. Matching core function; text reference is in the first column

candidate and reference templates (7 in Fig 2): the *TemplDissimilarity*; the same algorithm is then executed for the next candidate template minutia in reliability order. When all of the input minutiae have been processed, this global *TemplDissimilarity* value on templates is divided by the number of matched minutiae *MinutiaeNMatched*, finding in this way the mean. A comparison between a match threshold and this mean value can consequently be used to decide if the two templates belong to the same fingerprint (if the mean is below the threshold): lower *TemplDissimilarity* expresses more affinity.

That, therefore, is the full algorithm description, but as said before, the matching procedure will probably end before the complete minutiae list of the candidate template has been processed: if at the end of the minutiae matching routine the dissimilarity value between two matched minutiae is “very good”, that is below a tightening threshold *OptValue*, the counter *OptMinNumber* is incremented and as soon as it reaches a predefined constant value corresponding to the threshold *OptNumberThreshold*, the whole match-

ing procedure can be stopped with a positive result (8 in Fig 2). The algorithm can be positively stopped also as soon as it finds only one minutiae pair with an "exceptionally good" *MinDissimilarity* value below the *VeryOptValue* threshold (9 in Fig 2), which is intended to be much stricter than the previous *OptValue*.

The described algorithm complexity is $O(n^2)$, where n is the number of the minutiae in a single template, even if in practice, the approach of stopping the computation with few minutiae shows a significant speed improvement.

3.3 Algorithm Implementation

The fingerprint matching algorithm described in Sec. 3.2 has been fully developed in Java Card™ using the Java Card™ 2.1.2 API and finally deployed on Cyberflex Access 32Kb Java Card™ with the Cyberflex Access SDK (version 4.3). The chosen smartcard has 32Kbyte of EEPROM, about 1Kbyte of RAM memory distributed between the transaction mirror, stack and transient space, 8 bit CPU at up to 7.5Mhz external clock frequency and the transmission protocol used is the *T=0* at 9600 bit/sec.

The algorithm has been developed by implementing the Java Card™ Biometric API [9] realized by Java Card Forum™ (JCF): this application programming interface ensures the interoperability of many biometric technologies with Java Card™ and allows multiple independent applications on a card to access the biometric functionalities (like verification); this is ideal to secure the digital signature, storing and updating account information, personal data (i.e. health information) and even monetary value. Clearly, our application manages even the enrollment and match requests coming from the external PC applications through several *Card Acceptance Device* (CAD) sessions.

Java Card™ technology [7] adapts the Java™ platform for the use on smartcards, smart buttons or other simple devices, like USB tokens. This adaptation produces a global reduction of the platform functionalities and its result is a substantial decreasing of the expressive capacity. Benefits and drawbacks are identical to those of its "mother technology": high portability and programming/developing quickness, but also a reduced execution speed due to the additional bytecode interpretation layer.

Due to the environment constraints like the EEPROM space, we have limited the maximum number of minutiae forming the template to the 20 most reliable, and the neighbor feature values have been sampled to be then stored in the low capacity Java Card™ data types as *byte* type (the maximum minutia occupation is 40 *byte*).

4 Performance Results

To measure the performance, we used the *Finger Verification Competition 2002* [8] edition (FVC2002) fingerprint databases, which, as we know, is the only public benchmark (together with the other editions of the same contest, FVC2000 and FVC2004[1]) allowing the developers to unambiguously compare their algorithms. In particular, we have adopted the two databases respectively collected using the optical sensor "FX2000" by

[1] We have used the databases from FVC2002 since the FCV2004 collections were granted to us only a few weeks ago. For the permission to use these collections, we acknowledge Raffaele Cappelli, Dario Maio and Davide Maltoni from the Biometric Systems Lab (University of Bologna)

Biometrika and the optical sensor "TouchView II" by Identix; each of the databases is 100 fingers wide and 8 impressions per finger deep. Moreover, we have analyzed our algorithm in respect to the one described in [11], using the proprietary image database provided to us by the authors[2] and made up of about 550 samples collected using the FX2000 optical scanner model; henceforth referred to as the "Hybrid Database".

The most important biometric systems evaluation parameters are the *False Acceptance Rate* (FAR), *False Rejection Rate* (FRR), the *Equal-Error Rate* (EER), which denotes the error rate for which FAR and FRR are identical, and the *Receiver Operating Characteristic* (ROC) curve. Other interesting performance indicators can be derived to show the algorithm's behaviour for applications that need high security: for example, the FAR100 (lowest achievable FRR for a FAR $\leq 1\%$), FAR1000 (lowest FRR for FAR $\leq 0.1\%$) and ZeroFAR (the lowest FRR for FAR = 0%). Another important factor to be considered, especially for MOC algorithms, is clearly the average matching time.

The test distribution between positive and negative matches can greatly influence the declared performances, so we decided to run the same tests as the FVC2002 competition [8] between the same fingerprint images: 2,800 iterations to find FRR and 4,950 to find FAR. The same test distribution criteria of FVC2002 were adopted also for the Hybrid Database (1,485 FAR tests and 2,449 for FRR). We also kept the same algorithm parameters configuration during the tests of all these three image collections. However, better results could be obtained by suitably adapting the parameters to each database. In Table 1 we present the obtained reliability results, where "-" means that the score is not achievable with the particular parameter configuration used.

Table 1. Overall performance results of our algorithm

Fingerprint database	EER	FAR100	FAR1000	ZeroFAR
FVC2002 FX2000	8.5%	10.6%	12.5%	-
FVC2002 TouchView II	8.5%	10.6%	12.3%	-
Hybrid Database	0.48%	0.44%	0.53%	0.57%

Fig. 3 shows the FAR-FRR and the ROC curves for the tests on the FVC2002 FX2000 database. The strange shape of the graph lines (in respect to classic ones) comes from the decision to stop the algorithm even with few but "good" minutiae pairs. This decision is independent from the final matching score and so setting the match threshold to a low or a high value does not correspondingly results in a FAR or a FRR of 0% and 100%. For example, the FRR curve in Fig. 3 starts with a match threshold equal to 0 from about 14% and not 100%, since most of the matches has been however accepted using one of the stopping conditions described in Sec. 3.2. Moreover, these conditions introduce only less than 5% errors in the total of the false matches accepted.

For our main parameters configuration and our purposes we were essentially interested in giving the best FAR1000 performance, but we have tested other configurations that can improve EER to about 7% or take ZeroFAR to 15.6%.

In Fig. 4 we report, for two databases, the algorithm match time distribution with respect to the correct FRR tests. We can observe that an on card matching time of about

[2] For the permission to use their fingerprint database, we thank Tommaso Cucinotta and Riccardo Brigo from ReTiS Lab of Sant'Anna School of Advanced Studies (Pisa)

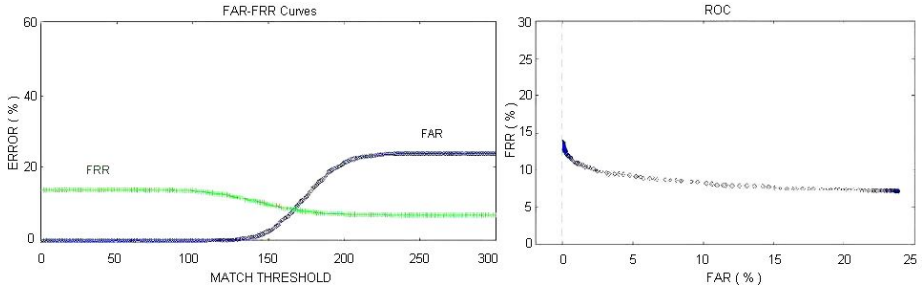


Fig. 3. FAR-FRR curves and ROC for FVC2002 FX2000 database

1-8 seconds is obtained for nearly all of the matches (about 90% for the Hybrid Db.). We have noticed that the minimum time of 1 second has been achieved frequently in these tests and can be obtained even more often using a good enrollment fingerprint image, since in this case the two stopping conditions of Sec. 3.2 can be met very often. The maximum time is instead about 45 seconds, but this result is obtained only when the two acquisitions belong to different fingerprints (not interesting for our purposes) or they are very disturbed: this prevents the algorithm for quickly stopping without exploring all the minutiae pairings (image quality affects the average match time).

All the tests have been run on a PC with Java™ , but using the exact same Java Card™ code downloaded in the card, since the second language is a subset of the first; in this way the same security performances are fully achievable even on the smartcard. We derived the matching time for the card application from the average time needed to match one single minutia on the card (measured directly in this environment), and multiplying for the minutiae number needed to stop the match (calculated in PC tests).

We have also compared our work on the Hybrid Database provided by the authors of [11], developed to be executed in a similar Java Card™ environment. Results show that we nearly halved the EER percentage of 0.8% achieved by that algorithm, obtaining a value of 0.48% (Table 1). Our algorithm is better also for the matching execution time: 1-8 seconds for nearly all of the matches (ours), against 11-12 seconds in [11].

It is important to point out that using a good quality enrollment image considerably improves the overall security performances: FAR1000 value can be reduced to about 5-6% as measured from other tests, mitigating also the partial overlapping problem. The hypothesis of having a good quality template is not too restrictive and it is easily applicable, since the enrollment phase is accomplished only one time at the release of the smartcard and the quality of the enrolling image can be easily checked. Therefore, using a good image at enrollment improves both reliability and speed performances.

5 Conclusion

In this paper we have proposed a new fingerprint matching algorithm tolerant to typical problems such as rotation, translation and ridge deformation. Our procedure achieves a very good speed performance for the Java Card™ platform restrictions: 1-8 seconds for most of the positive match tests. The high reliability, as determined from our analysis, can be further greatly improved using a good enrollment image, thus scoring a

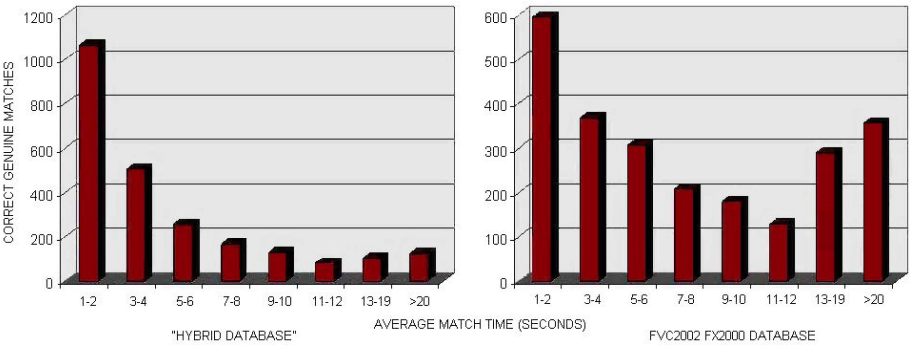


Fig. 4. Average smartcard match time on correct FRR tests, sampled in time intervals

FAR1000 result of about 5-6%, which makes the algorithm implementation feasible in the live-scan applications for identity verification (like a MOC system). Our procedure is stopped as soon as the two templates are considered to belong to the same finger, and so the algorithm stops before in correct FRR tests and later in correct FAR ones, showing an asymmetric behaviour.

References

1. D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, *Handbook of Fingerprint Recognition*, Springer, 2003, ISBN 0-387-95431-7.
2. *User's Guide to NIST Fingerprint Image Software (NFIS)*, NISTIR 6813, National Insitute of Standards and Technology.
3. J.H.Wegstein, *An Automated Fingerprint Identification System*, U.S. Government Publication, Washington, 1982.
4. N. K. Ratha, R. M. Bolle, V. D. Pandit, V. Vaish, *Robust Fingerprint Authentication Using Local Structural Similarity*, IEEE 2000.
5. A. K. Jain, S. Prabhakar, L. Hong and S. Pankanti, *Filterbank-based Fingerprint Matching*, IEEE Transactions on Image Processing, Vol. 9, No.5, pp. 846-859, 2000.
6. T. Hatano , T. Adachi, S. Shigematsu, H. Morimura, S. Onishi, Y. Okazaki, H. Kyuragi, *A Fingerprint Verification Algorithm Using the Differential Matching Rate*, ICPR02, III volume: pp. 799-802, 2002.
7. C. Enrique Ortiz, *An Introduction to Java Card™ Technology*, Part 1-2-3, 2003.
8. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman and A. K. Jain, *FVC2002: Second Fingerprint Verification Competition*, Proc. of International Conference on Pattern Recognition, pp. 811-814, Quebec City, August 11-15, 2002.
9. *Java Card™ Biometric API White Paper (Working Document)*, Version 1.1, NIST/Biometric Consortium, 2002.
10. Y.S.Moon, H.C. Ho, K.L. Ng, *A Secure Card System with Biometric Capability*, IEEE Conference on Electrical and Computer Eng., Volume 1, pp 261-266, 1999.
11. T. Cucinotta, R. Brigo, M. Di Natale, *Hybrid Fingerprint Matching on Programmable SmartCards*, TrustBus 2004, Springer LNCS volume 3184/2004 p. 232.
12. P. B. van Wamelen, Z. Li, S. S. Iyengar, *A Fast Algorithm for the Point Pattern Matching Problem*, 2000.
13. S. Bistarelli, G. Boffi, F. Rossi, *Computer Algebra for Fingerprint Matching*, Proc. International Workshop CASA'2003, Springer LNCS vol. 2657 2003.

Scenario Based Performance Optimisation in Face Verification Using Smart Cards

Thirimachos Bourlai, Kieron Messer, and Josef Kittler

Centre of Vision, Speech and Signal Processing
School of Electronics and Physical Sciences
University of Surrey
Guildford GU2 7XH, UK
{t.bourlai,k.messer,j.kittler}@surrey.ac.uk

Abstract. We discuss the effect of an optimisation strategy to be applied to image data in a smart card based face verification system. Accordingly, we propose a system architecture considering the trade-off between performance versus the improvement of memory and bandwidth management. In order to establish the system limitations, studies were performed on the XM2VTS and FERET databases demonstrating that, spatial and grey level resolution as well as JPEG compression settings for face representation can be optimised from the point of view of verification error. We show that the use of a fixed precision data type does not affect system performance very much but can speed up the verification process. Since the optimisation framework of such a system is very complicated, the search space was simplified by applying some heuristics to the problem. In the adopted suboptimal search strategy one parameter is optimised at a time. The optimisation of one stage in the sequence was carried out for the parameters of the subsequent stages. Different results were achieved on different databases, indicating that the selection of different optimum parameters for system evaluation may call for different optimum operating points.

1 Introduction

Designing an automatic personal identity verification system based on facial images is a challenging task [10, 11]. In a conventional architecture of a face verification system, the biometric template is stored in a database on the server where the verification is also carried out. Although acceptable for some applications, this mode of operation raises many privacy and security issues, which compromise user acceptability. To alleviate these problems, a favoured system setup was proposed in [2] where the biometric template is stored on a smart card together with the verification algorithm. In this novel distributed architecture (see *Figure 1*) the decision making is carried out on the smart card itself. No user data ever leaves the card for a verification making the system more secure and user friendly.

However, due to the severe constraints and limitations that small computing platforms often impose (such as low computational power, small storage capacity and poor communication bandwidth of the smart card), special considerations of the system design issues have to be made:

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 289–300, 2005.
© Springer-Verlag Berlin Heidelberg 2005

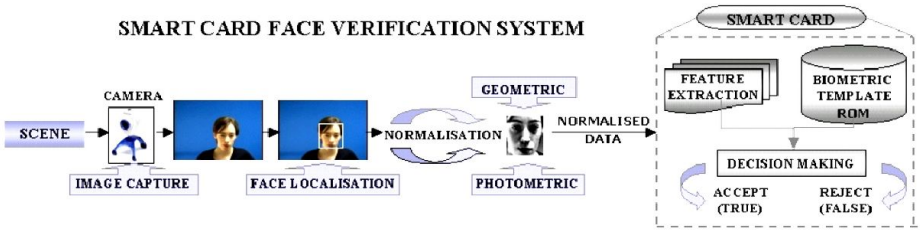


Fig. 1. Proposed smart card face verification system

- The verification algorithm needs to be computationally simple.
- The size of the biometric template to be stored on the card needs to small.
- The amount of probe image data transferred to the smart card needs to be limited.
- Smart cards do not yet have floating point co-processors so the number of mathematical operations need to be limited and number bit representation reduced.

In order to identify an optimised trade off between the computational complexity of the system (server or smart card) and the system performance (as measured by the verification error), many experiments have been performed. In our previous work [3], we determined the lowest number of bits needed for the template stored on the card; the minimum amount of data a probe image can be represented with and whether the performance is affected by using fixed point arithmetic. This was done independently for each of these parameters. We demonstrated that in general the computational complexity of our face verification system could be safely reduced without an corresponding increase in the verification error.

In [2] it has already been demonstrated that CS-LDA [5] is a computationally simple face verification algorithm that can be effectively used on smart cards. CS-LDA also achieves very low verification error rates. However, to enable the technique to run on smaller and cheaper cards the matching algorithm can be speeded up by reducing the bit resolution of the mathematical operations. In our experiments it has been found that one can significantly reduce the bit resolution without error degradation.

In practice, the storage and communication requirements of the system can be decreased. Therefore, the combined reduction of both spatial and grey level resolutions was investigated for the normalised face images, along with the use of the baseline JPEG compression scheme on both template and probe images [1]. Again it has been found that, in most cases examined, the spatial resolution can be significantly reduced and high compression rates can be used without any considerable increase in the observed error rates.

In this paper we revisit the performance versus computational complexity trade-off concerns and establish a optimisation methodology for combining and selecting all of these system parameters jointly. Since the optimisation framework of such a system is very complicated in terms of performance and computational cost, the search space was simplified by applying some heuristics to the problem. In the adopted suboptimal search strategy one parameter is optimised at a time. The optimisation of one stage in the sequence was carried out for the parameters of the subsequent stages. We also demonstrate that these optimum parameters change for each operational scenario as

different optimum operating points were identified when using different configurations and images from the XM2VTS and FERET datasets.

The rest of the paper is organised as follows. In the next section the database and protocols used in our experiments will be described. In Section 3 the experiments and results are presented before finally, some conclusions are made.

2 Databases and Protocols

For the purpose of this study, XM2VTS and FERET face databases were used in the experiments. XM2VTS is a multi-modal database consisting of face images, video sequences and speech recordings taken from 295 subjects at one month intervals. In this database, the data acquisition was distributed over a long period of time that resulted in significant variability of appearance of clients, e.g. changes of hair style, facial hair, shape and presence or absence of glasses (see *Figure 2*). The XM2VTS [6] database contains 4 sessions. During each session two head rotation and and "speaking" shots (subjects are looking just below the camera while reading a phonetically balanced sentence) were taken. From the "speaking" shot a single image with a closed mouth was chosen. Two shots at each session, with and without glasses, were acquired for people regularly wearing glasses.



Fig. 2. Sample images from XM2VTS database

For the purpose of personal verification, a standard protocol for performance assessment has been defined. This is the Lausanne protocol[4], which splits randomly all subjects into a client and impostor groups. The client group contains 200 subjects and the impostor group is divided into 25 evaluation and 70 test impostors. Eight images from 4 sessions are used. 200 subjects were used for training, that results in a total of 600/800 face images for configuration C1/C2.

From the sets containing the face images, training, evaluation and test set is built. There exist two different protocol configurations that differ by selecting particular shots of people into the training, evaluation and test sets. The training set is used to construct

client models; the evaluation set produces client and impostor access scores (used to compute a client-specific or global threshold that determines acceptance or rejection of a person); and the test set is selected to simulate realistic authentication tests where impostor's identity is unknown to the system. According to the Lausanne protocol the threshold is set to satisfy certain performance levels on the evaluation set. Finally, the performance measures of the verification system are the FA and FR rate as explained above. The XM2VTS protocol is an example of a closed-set verification protocol where the population of clients is fixed; system design can be tuned to the clients in the set.

FERET is a program ran from 1993 through 1997, that was sponsored by the Department of Defence's Counterdrug Technology Development Program through the Defence Advanced Research Products Agency (DARPA). The primary aim was to develop automatic face recognition capabilities that could be employed to assist security, intelligence and law enforcement personnel in the performance of their duties. The FERET image database [7] was assembled to support government monitored testing and evaluation of face recognition algorithms using standardised tests and procedures. The final database consists of 14051 eight-bit grey scale images of human heads with views ranging from frontal to left and right profiles.

The images were collected in 15 session, acquired in a semi-controlled environment but using the same physical setup in each photography session to maintain a degree of consistency. However, because the equipment had to be reassembled for each session, there was variation from session to session (see *Figure 3*). Images of an individual were acquired in sets of 5 to 11 images. Two frontal views were taken labelled **fa** and **fb**, where we have a different facial expression. For 200 sets of images, a third frontal image was taken, labelled as **fc**, using a different camera and different lighting. The rest of the images were collected at various aspects between right and left profile. Simple variations to the database were added by the photographers by taking a second set of images for which the subjects were asked to put on their glasses and/or pull their hair back. In some cases a second set of images of a person was taken on a later date (*duplicate set*). Such a set includes variations in pose, scale, illumination and expression of a face. The total number of clients that results in a total of 3570 face frontal images (used for training) is 1201 subjects.



Fig. 3. Sample of frontal images from FERET database [7]

3 Experiments and Results

Our face verification system has been evaluated via a set of experiments using both the XM2VTS and FERET data sets in a total of four different testing configurations:

- XM2VTS C1 Configuration I for the XM2VTS database.
- XM2VTS C2 Configuration II for the XM2VTS database.
- FE-XM2VTS C1 Configuration I for the XM2VTS database but with FERET data set used to generate the initial statistical model.
- FE-XM2VTS C2 Configuration II for the XM2VTS database but with FERET data set used to generate the initial statistical model.

The first two test protocols represent a closed-set protocol where all the enrolled clients are known to the system. The second two protocols represent an open set protocol - clients are not known to the system prior to enrolment. The system performance levels of the verification system were measured in terms of half-total error rate (HTER) on the test set of each protocol obtained using the EER threshold determined from the ROC curve computed on an independent evaluation set. Both ROC curves on the evaluation as well as on the test set would produce additional information about the system behaviour. However, because of the lack of space, these curves could not be included.

3.1 Optimisation Framework

The optimisation framework of our smart card face verification system in terms of performance and computational cost is very complicated because it includes a large number of degrees of freedom. It consists of 4 different testing configurations, 8 grey level resolutions (8bpp down to 1bpp), 16 spatial resolutions (110x102 down to 8x7), n-bit precision fixed point numbers ($n = [1-16]$), and finally 4 different operational stages where JPEG compression is applied using 19 different quality factors ranging from 5 to 100. Obviously, global optimisation requires an exhaustive search of 622592 experiments, which renders the effort not feasible. This is without including the pre-processing parameters that (for instance) a different filtering technique would introduce i.e. mask size and variance, or the number of PCA components necessary for achieving an optimum performance. In order to simplify the search space and to find a reasonable solution to such a problem some heuristics were applied. In the adopted suboptimal search strategy one parameter is optimised at a time. The optimisation of one stage in the sequence was carried out for the parameters of the subsequent stages set out as follows:

- The optimum grey level and spatial resolution for each dataset scenario was identified without applying any compression.
- Joint optimisation of JPEG compression quality factor and operational scenario (per testing configuration) was performed under the condition of fixed spatial resolution of 55x51 and grey level resolution kept to maximum (8bpp).
- The determined optimal JPEG compression operational scenario was applied and the optimal compression quality factor was identified under the condition of using the optimal spatial resolution and grey level resolution kept to maximum (8bpp).
- The optimum n-bit precision fixed point number was identified to lie within the range $n = [5 - 14]$ independently of the other parameters. This parameter was applied at the end of the previous stage (avoiding also to introduce statistical errors to all stages).

By adopting such a strategy, interesting results were obtained and in general different operating points were defined for the different testing configurations. But most importantly, the number of the experiments performed was limited to 1200 (approximately 519 times less computational effort).

3.2 Grey Scale Resolution

In this experiment we investigated the effect on performance by altering the grey-scale pixel resolution of the (55x51) normalised probe images in the training set. Since an 8 bit camera is used, the initial performance of our system was measured by using 8 bits per pixel for each face image. Then, the grey-scale resolution was reduced by a factor of 1bpp each time before building the PCA and LDA model. Both protocols in XM2VTS database were tested when this dataset was used exclusively and when FERET was used for training and XM2VTS for testing.

Table 1. Results obtained on XM2VTS when grey-scale resolution was reduced. (FE-XM2VTS = FERET was used for training and XM2VTS for testing)

Precision	XM2VTS C1	XM2VTS C2	FE-XM2VTS C1	FE-XM2VTS C2
8bpp	**0.04588**	**0.02644**	**0.06816**	**0.04028**
7bpp	0.06149	0.03859	0.08804	0.04888
6bpp	0.1773	0.14641	0.188	0.1645
5bpp	0.3835	0.3438	0.3282	0.3396
4bpp	0.4718	0.44	0.4203	0.4436
3bpp	0.5042	0.4868	0.459	0.4826
2bpp	0.5106	0.5002	0.4808	0.495
1bpp	0.5108	0.50735	0.4888	0.4982

From the results obtained (see table 1), it was concluded that the use of 8-bit grey-scale pixel resolution yields the best overall performance. However, system behaviour suggests that performance could be further improved if more than 8bpp are used.

3.3 Spatial Resolution

In this experiment the optimum spatial resolution for each dataset and configuration was obtained. The initial raw face images of both XM2VTS and FERET datasets were geometrically and photometrically normalised to a spatial resolution that was varied from 110x102 down to 8x7. The grey-scale resolution was kept at 8 bpp. Table 2 shows a summary of the results obtained.

For the XM2VTS database the image size can be reduced from 110x102 to 18x16 (CI) and to 40x37 (CII) for each configuration respectively. Over 38(5) times less data need to be sent to the smart card. Therefore, the computation load for the template matching on the smart card is significantly reduced while the performance is slightly improved. Comparable results are observed when FERET was used for training and XM2VTS (CI) for testing. However, in configuration CII an optimum operating point was obtained at almost the highest resolution, suggesting the need for another strategy to be adopted in order to minimised data transfer and processor load.

Table 2. Results obtained on XM2VTS when different spatial resolution was used. (FE-XM2VTS = FERET was used for training and XM2VTS for testing)

IMAGE SIZE	XM2VTS C1	XM2VTS C2	FE-XM2VTS C1	FE-XM2VTS C2
8x7	0.08025	0.08863	0.10091	0.11121
10x8	0.05482	0.06218	0.07652	0.07703
13x11	0.05296	0.04642	0.07053	0.06392
15x13	0.04662	0.03802	0.07319	0.05735
18x16	**0.03977**	0.03208	0.06524	0.04980
20x18	0.04340	0.03114	0.06811	0.04535
25x23	0.04190	0.02575	**0.06014**	0.04675
30x28	0.04025	0.02758	0.06248	0.04121
40x37	0.04225	**0.02214**	0.06281	0.04304
55x51	0.04588	0.02644	0.06816	0.04028
61x57	0.04409	0.02465	0.06828	0.03891
70x65	0.04494	0.02343	0.06951	0.03805
80x75	0.04777	0.02535	0.06788	0.03472
90x85	0.04711	0.02522	0.06994	0.03367
100x93	0.04680	0.02524	0.07248	**0.03177**
110x102	0.04647	0.02350	0.07437	0.03559

3.4 Fixed Point Arithmetic

In the absence of a floating point co-processor on the smart card, the use of the built-in simulated floating point unit will result in an increase of the overall computational cost on the card. By using n-bit precision data types on the server instead, we are able to use integers on the smart card, which can be extremely advantageous in terms of computational speed. Therefore, in this experiment the trade-off between performance and bit precision for the verification function parameters was investigated when using fixed point arithmetic for authentication. These parameters are the client specific LDA transformation matrix $\mathbf{a}$, the client mean vector μ_i and the global mean $\mu = \sum_{j=1}^{N} \mathbf{z}_i$, where N is the size of the training set and $\mathbf{z}_i$ are the training images. The basic idea behind that was to change the precision of the CSLDA transformation that is actually sent on the smart card for on-card verification based on the distance metric d_c given in the following equation:

$$d_c = \left| \mathbf{a}_i^T \mathbf{z} - \mathbf{a}_i^T \mu_i \right| \tag{1}$$

where $\mathbf{z}$ is the probe image and μ_i is the client mean.

Based on the results acquired (see *Figure 4*), the optimum n-bit precision was identified to lie within the range $n = [5 - 14]$. Specifically, 10-bit precision is the optimum one for the reference resolution 55x51 on both datasets. Note that fixed point numbers introduce some statistical errors, which do not necessarily affect negatively the performance results.

3.5 Compression

In this experiment, the optimum JPEG compression parameters were identified on the optimum operational stage obtained. Further to being an international standard, JPEG

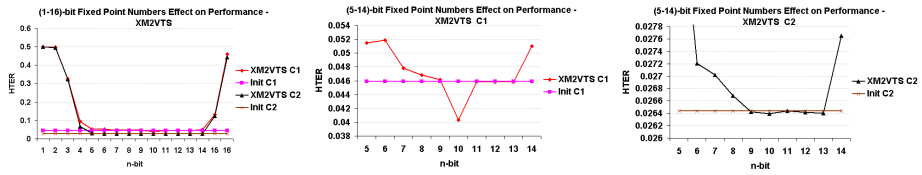


Fig. 4. Results obtained on the XM2VTS when fixed point arithmetic was applied. (a) Overall results, (b)/(c) (5-14)-bit on CI/CII respectively

was selected because it can make a claim to high visual fidelity, satisfactory error resilience and generality (that is, the ability to efficiently compress different types of imagery across a wide range of bit rates). Moreover, JPEG boasts a very low computational complexity, especially compared to methods like JPEG2000. Among many JPEG compression schemes, the baseline mode was used; it is very easy to implement and port on a small platform while it still achieves high compression/decompression speed. In order to prove that JPEG was the most appropriate selection for our experiments, we considered the evaluation study on the comparison between different still image coding standards [8]. The results of this work show that the choice of the 'best' standard is application dependent [9] and in cases where the interest lies in a low complexity, lossy compression scheme, JPEG provides a satisfactory solution.

For this experiment, a spatial resolution of 55x51 was used in order to study the effect of using JPEG compression at four different operational stages.

1. On probe images of all experimental sets, training, evaluation and testing set;
2. On probe images of only evaluation and testing set. This was deliberately chosen because it would be interesting to witness the effect of compression on the overall performance only in the case where probe images are sent to the smart card and training remains unaffected;
3. On templates;
4. On both probes (training and testing) and templates.

Different quality settings for the compressor were used. Image quality was traded-off against file size by adjusting those settings. In all cases, the range of the quality factor was modified from 5 to 100. In order to identify the optimum JPEG scenario, the system was evaluated based on the comparison between the initial HTER and the average HTER for all quality values (5-100) in each JPEG scenario. However, we also took into consideration the consistency of the good performance results in each scenario. Based on the results given in *Figure 5* we identified JPEG scenario one as the optimum in the case of XM2VTS and scenario two in the case where FERET was used for system training. Some example cases are also given in *Figure 6*.

3.6 Suboptimal Search Strategy Results

The final results obtained are brought together in table 3. There we can see the initial performance results and the results acquired after applying the new optimum parameters step by step. It is demonstrated that the use of such a combined parameter strategy

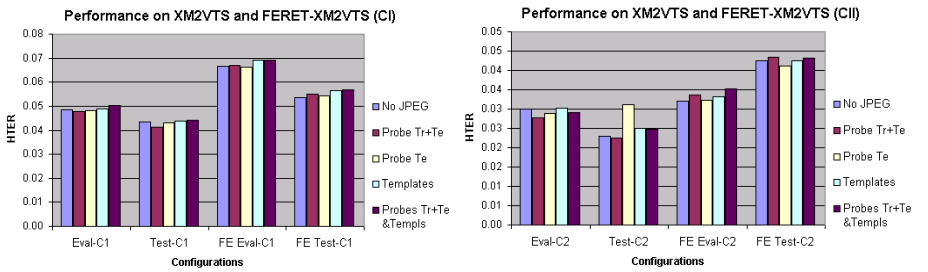


Fig. 5. Performance evaluation in XM2VTS and FERET-XM2VTS of all four JPEG scenarios

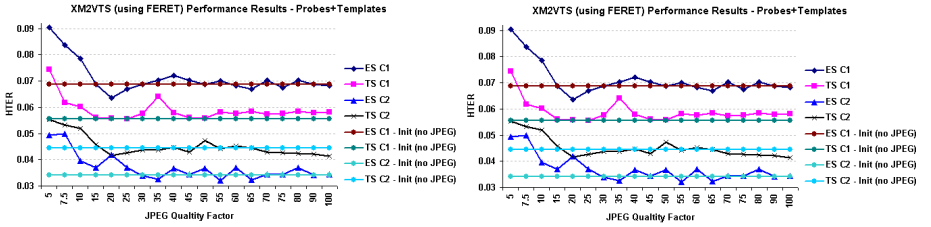


Fig. 6. Example cases of system behaviour on the different databases used. (a) JPEG scenario 3 on XM2VTS and (b) JPEG scenario 4 on FERET-XM2VTS

does not result in performance degradation and can become extremely advantageous when high resolution images are identified as the optimum ones in terms of system performance.

Interesting results were obtained by applying the proposed combined strategy. XM2VTS CI behaved better in low resolution images where the initial performance was improved by 13.4% only by fine tuning the resolution. It is obvious a higher-resolution image can tolerate more compression and that JPEG does not work well with extremely low resolutions where the byte file size of a JPEG compressed image increases due to the overhead of the JPEG file format. (A JPEG image has a specific internal structure that is common to all JPEG files. Part of this internal structure is known as the file ”header”, which basically contains information about the image file: its size, dimensions, etc.). In such an extreme case, by using fixed point arithmetic without JPEG we can achieve both performance improvement and system acceleration. Better overall results were achieved in configuration II (it has a bigger training set the CI and is more representative of a real system), where the optimum spatial resolution was identified to be a much higher one (40x37). By combining an 11-bit precision and JPEG compression on a relatively medium resolution, the performance increases by about 20% with an additional increase of system speed, both by the use of fixed point numbers and by decreasing about the size of the probe images sent to the card by a factor of three. An additional advantage is the improved overall memory management within the system through the compression of the probe images on the training set.

An expected degradation of performance was introduced by using FERET for training and XM2VTS for testing. Configuration one (CI) produced relatively the same results as before. However, in configuration two (CII), the optimum resolution was almost

Table 3. Best cases in both databases when the combined strategy was used (PROT = Protocol, PR = Probes, TE = Templates, Tr = Train, Te = Test, QUAL = quality, IBFS/CBFS P/T = Initial/Compressed Byte File Size for Probes/Templates, RESOL = resolution, FPN(n) = n-bit fixed point number)

DATABASES	PROT	Case	QUAL	IBFS P/T	CBFS P/T	RESOL	FPN(n)	HTER
XM2VTS	CI	-	-	-	-	55x51	-	**0.04588**
XM2VTS	CI	-	-	-	-	18x16	-	0.03977
XM2VTS	CI	PR/Tr/Te	65	288	429	18x16	-	0.04292
XM2VTS	CI	PR/Tr/Te	65	288	429	18x16	10	0.04267
XM2VTS	CII	-	-	-	-	55x51	-	**0.02644**
XM2VTS	CII	-	-	-	-	40x37	-	0.02213
XM2VTS	CII	PR/Tr/Te	20	1480	532	40x37	-	0.02128
XM2VTS	CII	PR/Tr/Te	20	1480	532	40x37	11	0.02127
FERET-XM2VTS	CI	-	-	-	-	55x51	-	**0.06816**
FERET-XM2VTS	CI	-	-	-	-	25x23	-	0.06014
FERET-XM2VTS	CI	PR/Te	35	575	475	25x23	-	0.05964
FERET-XM2VTS	CI	PR/Te	35	575	475	25x23	10	0.05964
FERET-XM2VTS	CII	-	-	-	-	55x51	-	**0.04028**
FERET-XM2VTS	CII	-	-	-	-	100x93	-	0.03177
FERET-XM2VTS	CII	PR/Te	45	9300	1579	100x93	-	0.03136
FERET-XM2VTS	CII	PR/Te	45	9300	1579	100x93	13	0.03137

the highest one (100x93) of the investigated range and the overall benefit of the JPEG compression efficiency to the system was highlighted. In both configurations, the overall trend remained similar to the one we had when XM2VTS was exclusively used. Particularly, in the case of FERET-XM2VTS and CI the performance increased by 12.5% and in CII by more than 22%. The overall results are summarised in table 3 and some particular examples are provided in *Figure 7*.

4 Conclusions

A suboptimal optimisation strategy for a smart card face verification system has been proposed. The effect of different parameters on the system performance was investigated in such a way that the search space was simplified. The benefits of optimisation can be further appreciated when fusion methods are to be incorporated onto the smart card and therefore an increased number of biometric templates have to be stored on the card.

The system was evaluated on different datasets and configurations hoping to achieve good and consistent results across all testing configurations for the same parameter setting. However, it transpired that each testing configuration required different parameter setting under the exception of grey level resolution and fixed point number representation. Since an 8 bit camera is used, system behaviour suggests that maximum performance is achieved at 8bpp and could possibly be further improved if higher resolution was available. The 10-bit fixed point number representation would provide optimal set-

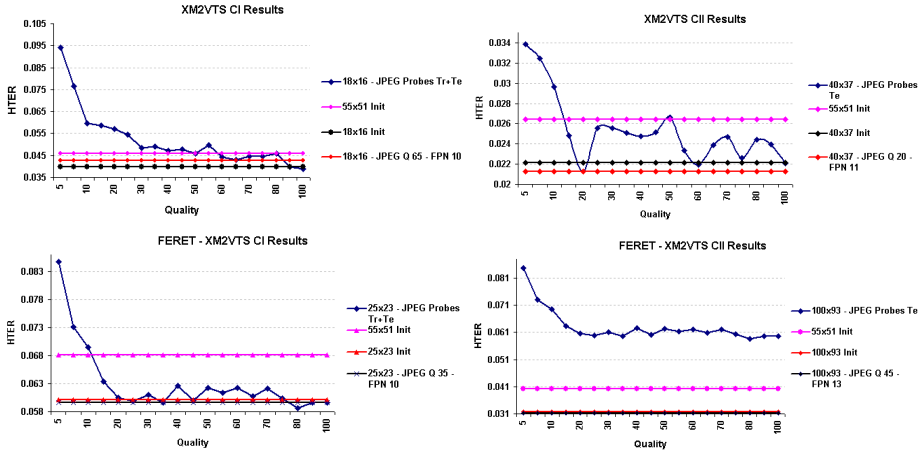


Fig. 7. The effect on XM2VTS (a),(b) top and FERET (c),(d) bottom, face databases when using the combines strategies

ting for all testing configurations, while speeding up the verification process. However, the optimum spatial resolution, JPEG compression quality factor as well as JPEG operational scenario differ from one experimental condition to another. Note that a quality threshold has been identified, below which, not only the performance can degrade but the amount of data to be stored can actually increase due to the overhead of the JPEG file format. Above that, there is a surprisingly wide quality range where compression does not seem adversely to affect performance, and for specific scenarios it may even improve system performance. Generally speaking, when operating at the limit of the quality settings good performance can be achieved, as well as gain in memory size and transfer speed.

An interesting example of such a strategy is when XM2VTS (CII) is used. In this case, the system speed is doubled and performance is improved by more than 16% only by the selection of an optimum resolution. Another 4% in performance can be gained by using JPEG while increasing system speed about three times more. Finally, the use of 11-bit precision, does not degrade performance as well as offering a significant relief on its complexity.

Acknowledgements

The authors would like to acknowledge the support received from OmniPerception Ltd, EU Media TrusteS and Biosecure projects, and EPSRC under grant GR/S46543/01(P).

References

1. T. Bourlai, J. Kittler, and K. Messer, ‘Jpeg compression effects on a smart card face verification system’, *Submitted for acceptance in IAPR Conference on Machine Vision Applications*, (16-18 May 2005).

2. T. Bourlai, K. Messer, and J. Kittler, 'Face verification system architecture using smart cards', *In the Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, **1**, 793–796, (23-26 August 2004).
3. T. Bourlai, K. Messer, and J. Kittler, 'Performance versus computational complexity trade-off in face verification', *In the Proceedings of the International Conference on Biometric Authentication, ICBA 2004*, 169–177, (15-17 July 2004).
4. J. Leuttin and G. Maître, 'Evaluation protocol for the extended m2vts database (xm2vts)', *IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence*, (July 1998).
5. Y.P. Li, J. Kittler, and J. Matas, 'Face verification using client specific fisher faces', *In J. T. Kent and R. G. Aykroyd, editors, Proc. Int. conf. on The Statistics of Directions, Shapes and Images*, 63–66, (September 2000).
6. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, 'Xm2vtsdb: The extended m2vts database', *AVBPA*, 72–77, (March 1999).
7. P. J. Phillips, H. J. Moon, S. A. Rizvi, and P. J. Rauss, 'The feret evaluation methodology for face-recognition algorithms', *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **22**(10), 1090–1104, (October 2000).
8. D. Santa-Crus and T. Ebrahimi, 'A study of jpeg 2000 still image coding versus other standards', *In Proc. Of the X European Signal Processing Conference, Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne*, 673–676, (September 2002).
9. M. Rabbani, 'The jpeg 2000 still-image compresion standard', *Eastman Kodak Research Labs, Diego Santa Cruz*, (2003).
10. M. Turk and A. Pentland, 'Eigenfaces for recognition', *J. Cognitive NeuroscienceIEEE Transactions on Pattern Analysis and Machine Intelligence*, **3**(1), 71–86, (1991).
11. W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, 'Face recognition: A literature survey', *UMD CfAR Technical Report CAR-TR-948*, (2000).

Characterization, Similarity Score and Uniqueness Associated with Perspiration Pattern

Aditya Abhyankar[1] and Stephanie Schuckers[1,2]

[1] Department of Electrical and Computer Engineering, Clarkson University, Potsdam NY 13676, USA
[2] Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, West Virginia 26506, USA

Abstract. Vulnerabilities in biometric systems including spoofing has emerged as an important issue. The focus of this work is on characterization of 'perspiration pattern' in a time-series of fingerprint images for liveness detection. By using information in the high pass bands of the images the similarity score for the two images is calculated to determine the uniqueness of the perspiration pattern. In this wavelet-based approach, the perspiration pattern is characterized by its energy distribution in the decomposed wavelet sub bands. We develop a similarity matching technique that is based on quantifying marginal distribution of the wavelet coefficients. The similarity match technique is based on Kullback-Leibler distance, which is used to decide 'uniqueness' associated with the perspiration pattern. Experimental results show good separation resolution in similarity scores of inter (43 subjects) and intra (12 subjects over 5 months) class comparisons. This may be considered as a robust liveness test for biometric devices.

1 Introduction

Biometrics are gaining popularity over conventional identifiers for doing personal identification. Unfortunately, with increased technological advancement, spoofing into a fingerprint identification system has become easier. Among various fingerprint security issues, it is of particular interest to check whether source of input signal is a live genuine finger, in order to make the system intelligent enough to be able to differentiate it from a signal originating from a spoof or a cadaver. This security test added as supplement to the authentication is termed as "liveness" detection [1], [2]. It has been demonstrated that perspiration can be used as a measure of "liveness" detection in case of fingerprint matching systems [3],[4],[5],[6],[7]. Unlike cadaver or spoof fingers, live fingers demonstrate a distinctive spatial moisture pattern, when in physical contact with the capturing surface of the fingerprint scanner. This is demonstrated in Figure (1). In this paper, it is shown that this pattern, called as 'perspiration pattern', may be unique in itself across individuals. Testing of this hypothesis is performed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 301–309, 2005.
© Springer-Verlag Berlin Heidelberg 2005

using similarity measurements in the wavelet sub-bands. Kullback-Leibler distance on wavelet sub bands is applied and the similarity score obtained is used to demonstrate "uniqueness" associated with the perspiration pattern [8].



Fig. 1. The fingerprints captured as a time sequence. The left figure is captured at zeroth second, while the right is captured after five seconds. Perspiration is observed as time progresses in live fingers

All natural images are by default, non-stationary, and contain several sub-images which can be exploited by varying the scale of analysis. Without extracting the exact positioning of these sub-images, direct matching of the images is not efficient. Wavelets are used to do this extraction.

In this paper, a simple sub-image decomposition using Daubechies orthogonal filters is designed. Sub-image decomposition parameters are computed to match high frequency components of the images, in order to enhance similarity score calculations. If the difference parameters are close to zero, two images are similar. Histograms of all the sub-images are calculated before performing matching.

2 Data Collection

The data required for this work was collected at Clarkson University and West Virginia University. Protocols for data collection from the subjects were followed that were approved by the Clarkson University and West Virginia University Institutional Review Board (IRB) (HS#14517 and HS#15322).

Data previously collected in our lab is used to test interclass similarity scores. This data set is diverse as far as age, sex, and ethnicity is concerned. This data set is comprised of different age groups (11 people between ages 20-30 years, 9 people between 30-40, 7 people between 40-50, and 6 people greater than 50), ethnicities (Asian-Indian, Caucasian, Middle Eastern), and approximately equal

Table 1. Data set: Distribution

Live	Capacitive DC Precise Biometric	Electro-optical Ethentica	Optical Secugen	Time span
Inter-class	31	30	30	
Intra-class	12	12	12	5 months

numbers of men and women. The data was collected using three different fingerprint scanners, with different underlying technologies. Three scanners from three different companies namely Secugen (model FDU01), Ethentica (model Ethenticator USB 2500)and Precise Biometrics (model PS100) with optical, electro-optical and capacitive DC technique of capturing fingerprint respectively, were chosen. Following table summarizes this data.

For analyzing intra class similarity data was collected from 12 subjects, over a period of 5 months. The subjects were asked to give their fingerprint samples at three visits and every time time-series capture was obtained three times. So, for 12 classes 9 intra class patterns were obtained.

3 Perspiration Pattern Extraction Algorithm

The algorithm is given in detail in [5],[6],[7]. The algorithm uses Daubechies wavelet-based approach to decompose 0 second image and second image at either 2 or 5 seconds. Maxima energy extraction is done after initial image enhancement for both the images. Multiresolution analysis is used to process LP information, while wavelet packets are used to extract information form HP bands. Only those coefficients are retained which experience more than 40% change in the

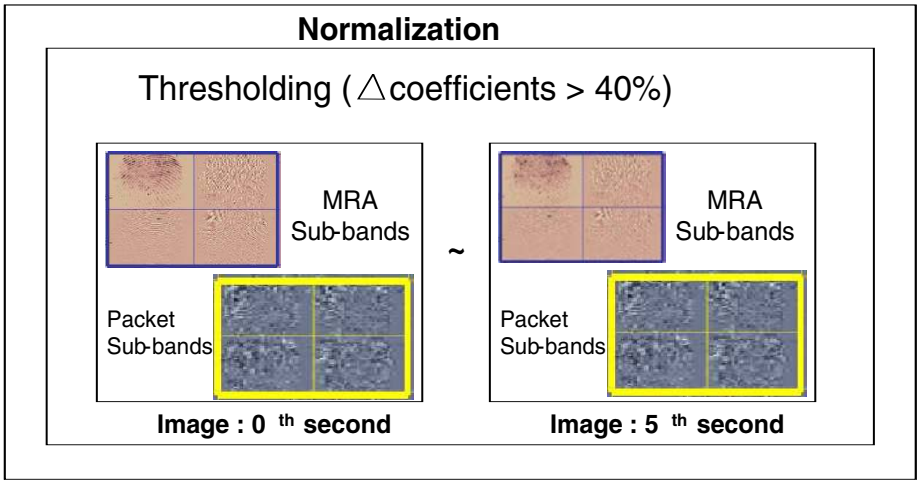


Fig. 2. Algorithm to extract perspiration pattern. Four sub-bands each form Multiresolution analysis and wavelet packets are shown separately

energy content. Normalization is done by the energy content of the later image to account for pressure variations. The entire algorithm in snap shot is given in Figure (2).

Total energy associated with the changing coefficients, normalized by total energy of the second image is used as a measure to decide the "liveness" associated with the scanner. It is given by following formula,

Formula: "liveness measure"

$$e\% = \left(\frac{\sum \text{energy of sub bands of the thresholded difference image}}{\sum \text{energy of sub bands of the image captured after five seconds}}\right) \times 100 \tag{1}$$

The retained coefficients relate directly to the perspiration pattern as shown in the reconstructed images in Figure (3). The following section describes the sub-level decomposition using wavelets. The scales selected for the algorithm are 2 and 3 respectively for mulriresolution analysis and wavelet packet analysis, respectively. The basis is formed by ψ as well as ϕ.

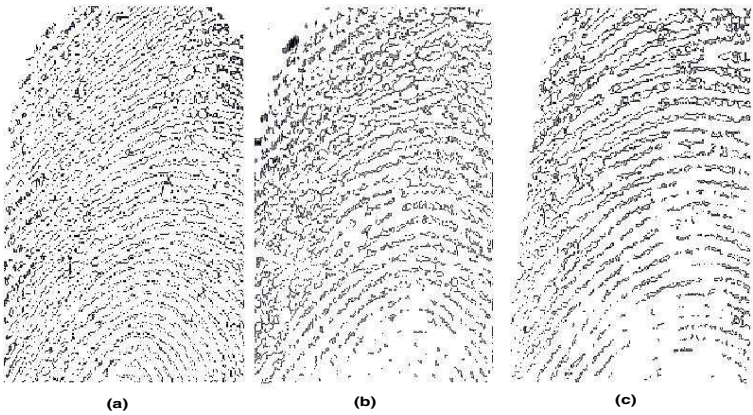


Fig. 3. (a)-(b)-(c) show perspiration patterns for three different subjects. Although these patterns are directly related to the energy changes, they are not normalized and hence are not exactly the same energy measures as given in formula (1)

4 Sub-level Decomposition of Images and Pattern Characterization

Assuming existence of nested sequences of subspaces $\{V_j\}_{j=-\infty}^{\infty}$ the selected set of Daubechies scaling functions is $\{\phi(x-k)\}_{k\in Z}$ is an orthonormal basis,i.e.,

$$\int_{-\infty}^{\infty} \phi(x-k)\phi(x-k')dx = \begin{cases} 0, k \neq k', \ k, k' \in Z \\ 1, k = k' \end{cases} \tag{2}$$

For the vector space V_j spanned by the discrete scaling functions $\{\phi(2^j x-k)\}$, $f_j(x) \in V_j$ [9],

$$f_j(x) = \sum_k \alpha_{j,k} 2^{\frac{j}{2}} \phi(2^j x - k) \tag{3}$$

$$\alpha_{j,k} = \int_{-\infty}^{\infty} f(x) 2^{\frac{j}{2}} \phi(2^j x - k) dx \tag{4}$$

and the set $\{\phi(2^j x - k)\}$ constitutes the basis.

Now $V_j \in V_{j+1}$, and V_{j+1} has better refinement than V_j. This "difference" is a subset of V_{j+1} spanned by the discrete wavelets of the subspace W_j, $g_j(x) \in W_j$,

$$g_j(x) = \sum_k \beta_{j,k} 2^{\frac{j}{2}} \psi(2^j x - k) \tag{5}$$

$$\beta_{j,k} = \int_{-\infty}^{\infty} f(x) 2^{\frac{j}{2}} \psi(2^j x - k) \tag{6}$$

The basis $\{\psi(2^j x - k)\}$ of this vector space W_j are always orthogonal to the scaling functions $\{\phi(2^j x - k)\}$ of V_j on $(-\infty, \infty)$,

$$\int_{-\infty}^{\infty} 2^{\frac{j}{2}} \phi(2^j x - k) 2^{\frac{j}{2}} \psi(2^j x - k') dx = \delta_{k,k'} \tag{7}$$

4.1 Orthogonal Wavelet Filters

The wavelet filters $LP_{k,l}$,$LP_{k,l}$,$HP_{m,l}$ and $HP_{m,l}$ form an orthogonal filter set if they satisfy following conditions,

$$< LP_{k,l}, LP_{k',l} >= \delta_{k,k'} \tag{8}$$

$$< HP_{m,l}, HP_{m',l} >= \delta_{m,m'} \tag{9}$$

$$< LP_{k,l}, HP_{m,l} >= 0 \tag{10}$$

where, $LP_{k,l}$ and $HP_{m,l}$ stand for low pass and high pass decomposition filters, respectively; while,$LP_{k,l}$ and $HP_{m,l}$ are reconstruction filters, respectively. Low pass filters come form ϕ functions, while high pass filters come form ψ functions.

If the base image is denoted by $IM^1_{i,j}$, then applying both low as well as high filters in horizontal and vertical directions would result in four sub-bands, namely LL, HL, LH and HH. They can be written as,

$$LL^0_{k,k'} = \sum_{i,j} \widetilde{LP_{k,i}} \widetilde{LP_{k',j}} IM^1_{i,j} \tag{11}$$

$$LH^0_{k,m} = \sum_{i,j} \widetilde{LP_{k,i}} \widetilde{HP_{m,j}} IM^1_{i,j} \tag{12}$$

$$HL^0_{m,k} = \sum_{i,j} \widetilde{HP_{m,i}} \widetilde{LP_{k,j}} IM^1_{i,j} \tag{13}$$

$$HH^0_{m,m'} = \sum_{i,j} \widetilde{HP_{m,i}} \widetilde{HP_{m',j}} IM^1_{i,j} \tag{14}$$

Here, $LL^0_{k,k'}$ is the low frequency component, $LH^0_{k,m}, HL^0_{m,k}, HH^0_{m,m'}$ are the high frequency components in horizontal, vertical and diagonal directions, respectively, after decomposing the image. Since, the filters used are orthogonal filters, as they follow properties in equation (2), the original image can be reconstructed from these sub-bands using following reconstruction formulae [10],

$$IM^1_{i,j} = \sum_{i,j} \widetilde{LP_{k,i}} \widetilde{LP_{k',j}} IM^1_{i,j} + \sum_{i,j} \widetilde{LP_{k,i}} \widetilde{HP_{m,j}} IM^1_{i,j} + \sum_{i,j} \widetilde{HP_{m,i}} \widetilde{LP_{k,j}} IM^1_{i,j} + \sum_{i,j} \widetilde{HP_{m,i}} \widetilde{HP_{m',j}} IM^1_{i,j}$$

This formula, after extending it to the respective levels of decomposition for MRA and packet analysis, can be used to visualize the perspiration pattern, after doing the thresholding. By the end of the process we will have 1 LL band and 7 high pass bands, coming from MRA and packet analysis. Sub-band alignment is performed as the complete algorithm itself is sub-band oriented. The MRA uses LL band and computes $IM^{-1}_{i,j}$, while packet analysis gives $IM^{-2}_{i,j}$, where the basis is adaptive and is calculated using Shannon entropy [7]. Thus, for MRA we get our output from $V_{-2} \oplus W_{-2} \oplus W_{-1}$ and for packet analysis we get our output from $V_{-3} \oplus W_{-3} \oplus W_{-2} \oplus W_{-1}$. Essentially, after thresholding, the method measures energy at the output of filter banks as extracted perspiration pattern. The main idea behind the algorithm is that energy distribution in energy domain identifies a pattern [7].

5 Similarity Measurement

The energy distribution described above is the pattern being considered to analyze similarity between inter and intra class perspiration patterns, score based on the 'Kullback-Leibler' distance between two images is calculated [8]. The Kullback-Leibler distance is essentially a relative entropy between two densities d_1 and d_2, and is given as [11],

$$D(d_1||d_2) = \int f \log \frac{d_1}{d_2} \tag{15}$$

where, densities d_1 and d_2 represents two images under consideration. This is an attempt to characterize the perspiration pattern via marginal distributions of their wavelet sub-band coefficients [12].

Generalized Gaussian density (GGD) is defined as,

$$p(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(\frac{|x|}{\alpha})^\beta} \tag{16}$$

where, $\Gamma(.)$ is the Gamma function, so,

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, z > 0$$

Here, α is variance and β is inversely proportional to the decreasing peak frequency. A good probability density function (PDF) approximation for the marginal density of coefficients at a particular sub-band (2 for MRA and 3 for packets for this algorithm), is achieved by adaptively varying α and β of GGD. Marginal distributions give better representation of of perspiration pattern than the wavelet sub-band energies.

Using equations (15) and (16), closed form of the KLD is given as [12],

$$D(p(.;\alpha_1,\beta_1)||p(.;\alpha_2,\beta_2)) = \log(\frac{\beta_1\alpha_2\Gamma(1/\beta_2)}{\beta_2\alpha_1\Gamma(1/\beta_1)}) + (\frac{\alpha_1}{\alpha_2})\frac{\Gamma((\beta_2+1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1} \tag{17}$$

The similarity measurement between two wavelet sub-bands can be computed very efficiently using the model parameters.

The overall distance between the images can be given as follows,

$$D(IM_1, IM_2) = \sum_j D(p(.;\alpha_1^j,\beta_1^j)||p(.;\alpha_2^j,\beta_2^j)) \tag{18}$$

This is because of the scalable nature of the wavelet transform. The wavelet coefficient in different sub-bands are independent, and so to find the overall score, individual KLDs are summed up. Here, j is the sub-band level [13],[12].

6 Experimental Results

The experimental calculations are done on the data set mentioned in section (2). Statistical independence of the perspiration patterns is directly proportional to the degrees of freedom used for doing similarity analysis. Moreover more degrees of freedom results in a more complex system. We continued to use 10000 maximum energy extracted points by the algorithm in [5]. 820 interclass and 72 intra-class combinations were exploited. As, for the intra-class scores, as the data collection was performed over 5 months, consistency factor of the perspiration pattern is also studied. No environmental conditions are tested intentionally.

The wavelet coefficients obtained from the algorithm [5], were used for further processing. The similarity score between the images is calculated using equation (17). This form of KLD is easy to implement and is found better than other close techniques like Bhattacharya coefficient [11]. A smaller similarity score indicates a better match. The in-band matching is done for individual bands, and then the scores are added and normalized. No further thresholding is implemented as the coefficients are already thresholded.

Results are shown in Figure (4). The normalized rates are plotted on the y-axis and normalized similarity scores are plotted on x-axis. Inter class distribution is observed to be very much random, while intra class distribution is observed to be very much similar. Hardly any variation in the intra-class distribution is observed, thus confirming the consistency in the perspiration pattern of the 12 subjects over the period of 5 months. Distinct separation between the two classes is seen.

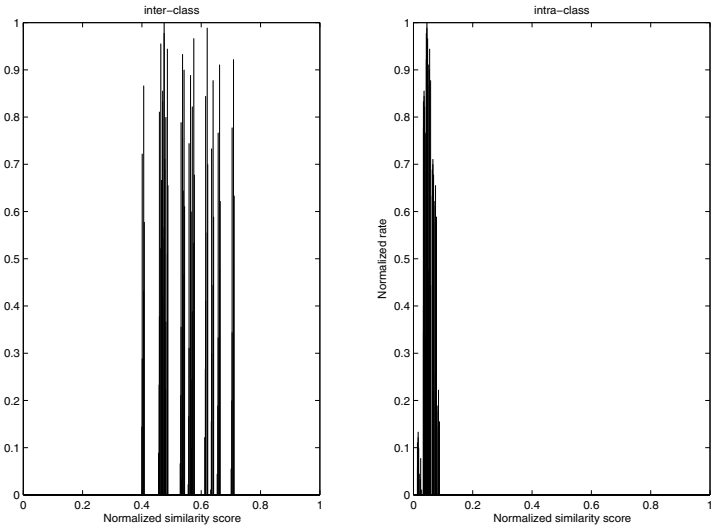


Fig. 4. Results of characterization of perspiration pattern. Similarity scores for inter and intra-class patterns. The similarity score is normalized. The intra class can be seen very similar, and inter class can be seen distributed around 0.6, and hence very much random

7 Conclusion

The perspiration pattern was observed to be 'unique', for the limited data set. The pattern also showed good consistency within the same class when monitored over 5 months. It is required to explore this matter with much wider data set, for different environmental conditions, to see the exact relation between sweat pores and this pattern. Previous work in liveness using perspiration pattern used general features across all the subjects to specify liveness. By expecting a specific liveness pattern from an individual, the liveness algorithm may be more robust to attacks. Beyond its possibility as a biometric alone or more promisingly in conjunction with fingerprint, it can certainly play a significant role in 'liveness' detection associated with the fingerprint scanners.

References

1. Davide Maltonie, Dario Maio, Anil K. Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag New York, Inc., 2003.
2. N. Ratha. Enhancing security and privacy in biometrics-based authentication systems. *IBM systems journal*, 40:614-6134, 2001.
3. Reza Derakshani, Stephanie Schuckers, Larry Hornak, and Lawrence Gorman. Determination of vitality from a non-invasive biomedical measurement for use in fingerprint scanners. *Pattern Recognition Journal*, 36(2), 2003.
4. Stephanie Schuckers. Spoofing and anti-spoofing measures. In *Information Security Technical Report*, volume 7, pages 56-62, 2002.

5. Aditya Abhyankar and Stephanie Schuckers. Wavelet-based approach to detecting liveness in fingerprint scanners. *Proceedings of the SPIE Defense and Security Symposium, Biometric Technology for Human Identification*, April 2004.
6. Stephanie Schuckers and Aditya Abhyankar. Detecting liveness in fingerprint scanners using wavelets: Results of the test dataset. *Proceedings of the Biometric Authentication Workshop, ECCV,*, May 2004.
7. Aditya Abhyankar. *A Wavelet-based approach to detecting liveness in fingerprint scanners*. Master's thesis, 2003.
8. S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat*, 22:7986 1951.
9. Ingrid Daubechies. *Ten Lectures on Wavelets*. Society of Industrial and Applied Mathematics, 1998.
10. Ingrid Daubechies, Yves Meyer, Pierre Gilles Lemerie-Rieusset, Philippe Techamitchian, Gregory Beylkin, Ronald Coifman, M. Victor Wickerhauser, and David Donoho. Wavelet transform and orthonormal wavelet bases. In *Different Perspectives on Wavelets*, volume 47, pages 1-33, San Antonio, Texas, Jan. 1993.
11. Don H. Johnson and Sinan Sinanovi. Symmetrizing the kullback-leibler distance. *IEEE Trans. Image Proc*, March 2001.
12. M. N. Do and M. Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Trans. Image Proc*, Dec 1999.
13. T. Chang and C.-C. J. Kuo. Texture analysis and classification with tree-structure wavelet transform. *IEEE Trans. Image Proc*, 2(4) 1993. This article was processed using the LATEX macro package with LLNCS style

Fuzzy Vault for Fingerprints

Umut Uludag[1], Sharath Pankanti[2], and Anil K. Jain[1]

[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824
{uludagum,jain}@cse.msu.edu
[2] Exploratory Computer Vision Group, IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598
sharat@us.ibm.com

Abstract. Biometrics-based user authentication has several advantages over traditional password-based systems for standalone authentication applications, such as secure cellular phone access. This is also true for new authentication architectures known as *crypto-biometric* systems, where cryptography and biometrics are merged to achieve high security and user convenience at the same time. In this paper, we explore the realization of a previously proposed cryptographic construct, called *fuzzy vault*, with the fingerprint minutiae data. This construct aims to secure critical data (e.g., secret encryption key) with the fingerprint data in a way that only the authorized user can access the secret by providing the valid fingerprint. The results show that 128-bit AES keys can be secured with fingerprint minutiae data using the proposed system.

1 Introduction

In traditional cryptography one or more keys are used to convert the plain text (data to be encrypted) to cipher text (encrypted data): the encrypting key(s) maps the plain text to essentially a sequence of random bits, that can only be mapped back to the plain text using the appropriate decrypting key(s). Without the knowledge of the correct decrypting keys, the conversion of cipher text to the plain text is *infeasible* considering time and cost limitations [1]. Hence, the cipher text is *secured*: even if an attacker obtains the cipher text, she cannot extract useful information from it. Here, the plain text can be any data that needs to be stored or transmitted securely: financial transactions, email communication, secret cryptographic keys, etc. Current cryptographic algorithms (e.g., Advanced Encryption Standard (AES) [2], Data Encryption Standard (DES) [1], RSA [1]) have a very high proven security but they suffer from the key management problem: all these algorithms fully depend on the assumption that the keys will be kept in absolute secrecy. If the secret key is compromised, the security provided by them immediately falls apart. Another limitation of the these algorithms is that they require the keys to be very long and random for higher security, e.g., 128 bits for AES [2], which makes it impossible for users to memorize the keys. As a result, the cryptographic keys are stored securely (e.g., in a computer or on a smart card) and released based on some alternative authentication mechanism. If this authentication succeeds, keys can be used in encryption/decryption procedures.

The most popular authentication mechanism used for key release is based on passwords, which are again cryptographic key-like strings but simple enough for users to

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 310–319, 2005.
© Springer-Verlag Berlin Heidelberg 2005

users to remember. Hence, the plain text protected by a cryptographic algorithm is only as secure as the password (weakest link) that releases the correct decrypting keys. Simple passwords compromise security, but complex passwords are difficult to remember and expensive to maintain. Further, passwords are unable to provide non-repudiation: a subject may deny releasing the key using password authentication, claiming that her password was stolen. Many of these limitations of password-based key release can be eliminated by incorporating biometric authentication. Biometric authentication [3] refers to verifying individuals based on their physiological and behavioral traits. It is inherently more reliable than password-based authentication as biometric characteristics cannot be lost or forgotten. Further, biometric characteristics are difficult to copy, share, and distribute, and require the person being authenticated to be present at the time and point of authentication. Thus, biometrics-based authentication is a potential candidate to replace password-based authentication, either for providing complete authentication mechanism or for securing the traditional cryptographic keys.

A biometric system and a cryptographic system can be merged in one of the following two modes: (i) In biometrics-based key release, the biometric matching is decoupled from the cryptographic part. Biometric matching operates on the traditional biometric templates: if they match, cryptographic key is released from its secure location, e.g., a smart card or a server. Here, biometrics effectively acts as a wrapper mechanism in cryptographic domain. (ii) In biometrics-based key generation, biometrics and cryptography are merged together at a much deeper level. Biometric matching can effectively take place within cryptographic domain, hence there is no separate matching operation that can be attacked; positive biometric matching *extracts* the secret key from the conglomerate (key/biometric template) data. An example of the biometric-based key generation, called *fuzzy vault,* was proposed by Juels and Sudan [4]. This cryptographic construct, as explained in later sections, has the characteristics that make it suitable for applications that combine biometric authentication and cryptography: the advantages of cryptography (e.g., proven security) and fingerprint-based authentication (e.g., user convenience, non-repudiation) can be utilized in such systems.

2 Background

Tuyls et al. [5] assume that a noise-free template X of a biometric identifier is available at the enrollment time and use this to enroll a secret S to generate a helper data W. If each dimension of the multidimensional template is quantized at q resolution levels, the process of obtaining W is akin to finding residuals that must be added to X to fit to odd or even grid quantum depending upon whether the corresponding S bit is 0 or 1. At decryption time, the noise-prone biometric template Y is used to decrypt W to obtain a decrypted message S' which is approximately the same as S. In each dimension, the process of decryption guesses whether a particular bit of secret S is 0 or 1 depending upon whether the sum of Y and W resides in even or odd quantum of the corresponding dimension. It is hoped that the relatively few errors in S' can be corrected using error-correction techniques. The proposed technique assumes that the biometric representations are completely aligned and that noise in each dimension is relatively small compared to the quantization magnitude Q. Due to variability in the

biometric identifier, different W may be generated for the same message S. The authors prove that very little information is revealed from W by appropriately tuning the quantization scheme with respect to the measurement noise.

Juels and Sudan's fuzzy vault scheme [4] is an improvement upon the previous work by Juels and Wattenberg [6]. In [4], Alice can place a secret κ (e.g., secret encryption key) in a vault and lock (secure) it using an unordered set A. Here, unordered set means that the relative positions of set elements do not change the characteristics of the set: e.g., the set {-2, -1, 3} conveys the same information as {3, -1, -2}. Bob, using an unordered set B, can unlock the vault (access κ) only if B overlaps with A to a great extent. The procedure for constructing the fuzzy vault is as follows: First, Alice selects a polynomial p of variable x that encodes κ (e.g., by fixing the coefficients of p according to κ). She computes the polynomial projections, $p(A)$, for the elements of A. She adds some randomly generated chaff points that do not lie on p, to arrive at the final point set R. When Bob tries to learn κ (i.e., find p), he uses his own unordered set B. If B overlaps with A substantially, he will be able to locate many points in R that lie on p. Using error-correction coding (e.g., Reed-Solomon [7]), it is assumed that he can reconstruct p (and hence κ). A simple numerical example for this process is as follows: Assume Alice selects the polynomial $p(x) = x^2 - 3x + 1$, where the coefficients (1, -3, 1) encode her secret κ. If her unordered set is $A = \{-1, -2, 3, 2\}$, she will obtain the polynomial projections as $\{(A, p(A))\} = \{(-1,5),(-2,11),(3,1),(2,-1)\}$. To this set, she adds two chaff points $C = \{(0,2),(1,0)\}$ that do not line on p, to find the final point set $R = \{(-1,5),(-2,11),(3,1),(2,-1),(0,2),(1,0)\}$. Now, if Bob can separate at least 3 points from R that lie on p, he can reconstruct p, hence decode the secret represented as the polynomial coefficients (1, -3, 1). Otherwise, he will end up with incorrect p, and he will not be able to access the secret κ.

The security of this scheme is based on the infeasibility of the polynomial reconstruction problem (i.e., if Bob does not locate many points which lie on p, he can not feasibly find the parameters of p, hence he cannot access κ). The scheme can tolerate some differences between the entities (unordered sets A and B) that lock and unlock the vault, so Juels and Sudan named their scheme *fuzzy vault*. This fuzziness can come from the variability of biometric data: even though the same biometric entity (e.g., right index finger) is analyzed during different acquisitions, the extracted biometric data will vary due to acquisition characteristics (e.g., placement of the finger on the sensor), sensor noise, etc. On the other hand, in traditional cryptography, if the keys are not exactly the same, the decryption operation will produce useless random data. Note that since the fuzzy vault can work with unordered sets (common in biometric templates, including fingerprint minutiae data), it is a promising candidate for biometric cryptosystems. Having said this, the fuzzy vault scheme requires prealigned biometric templates. Namely, the biometric data at the time of enrollment (locking) must be properly aligned with biometric data at the time of verification (unlocking). This is a very difficult problem due to different types of distortion that can occur in biometric data acquisition. Further, the number of feasible operating

points (where the vault operates with negligible complexity, e.g., conveyed via the number of required access attempts to reveal the secret, for a genuine user and with considerable complexity for an imposter user) for the fuzzy vault is limited: for example, the flexibility of a traditional biometric matcher (e.g., obtained by changing the system decision threshold) is not present.

Clancy et al. [8] proposed a fingerprint vault based on the fuzzy vault of Juels and Sudan [4]. Using multiple minutiae location sets per finger (based on 5 impressions of a finger), they first find the canonical positions of minutia, and use these as the elements of the set A. They added the maximum number of chaff points to find R that locks κ. However, their system inherently assumes that fingerprints (the one that locks the vault and the one that tries to unlock it) are pre-aligned. This is not a realistic assumption for fingerprint-based authentication schemes. Clancy et al. [8] simulated the error-correction step without actually implementing it. They found that 69-bit security (for False Accept Rate (FAR)) could be achieved with a False Reject Rate (FRR) of 20-30%. Note that the cited security translates to $2^{-69} \approx 1.7 \cdot 10^{-21}$ FAR. Further, FRR value suggests that a genuine user may need to present her finger multiple times to unlock the vault.

The design of a fuzzy vault (without the actual implementation) using minutiae-based lines was given in [9]. A more detailed survey of biometric cryptosystems can be found in [10].

3 Proposed Method

In this section we present our implementation of the fuzzy vault, operating on the fingerprint minutiae features. These features are defined as abrupt changes in the regular ridge structure on the fingertip, characterized by either ending or bifurcation of the ridges. Typically, they are represented as (x, y, θ) triplets, denoting their row indices (x), column indices (y) and angle of the associated ridge, respectively. These features are shown in Fig. 1, where two fingerprint images obtained from the same finger at different times, with overlaid minutiae are shown. The minutiae are found using the algorithm outlined in [11]. Note that the variability in the number and position of minutiae is evident in the two images.



(a) (b)

Fig. 1. Fingerprint images with overlaid minutiae: (a) template image (28 minutiae), (b) query image (26 minutiae)

Fig. 2 shows the block diagram of the proposed fingerprint fuzzy vault system. Fig. 3 shows the variables used in the system pictorially: the polynomial in Fig. 3(a) encodes the secret. It is evaluated at both genuine and chaff points in Fig. 3(b). Finally, the vault is the union of genuine and chaff points.

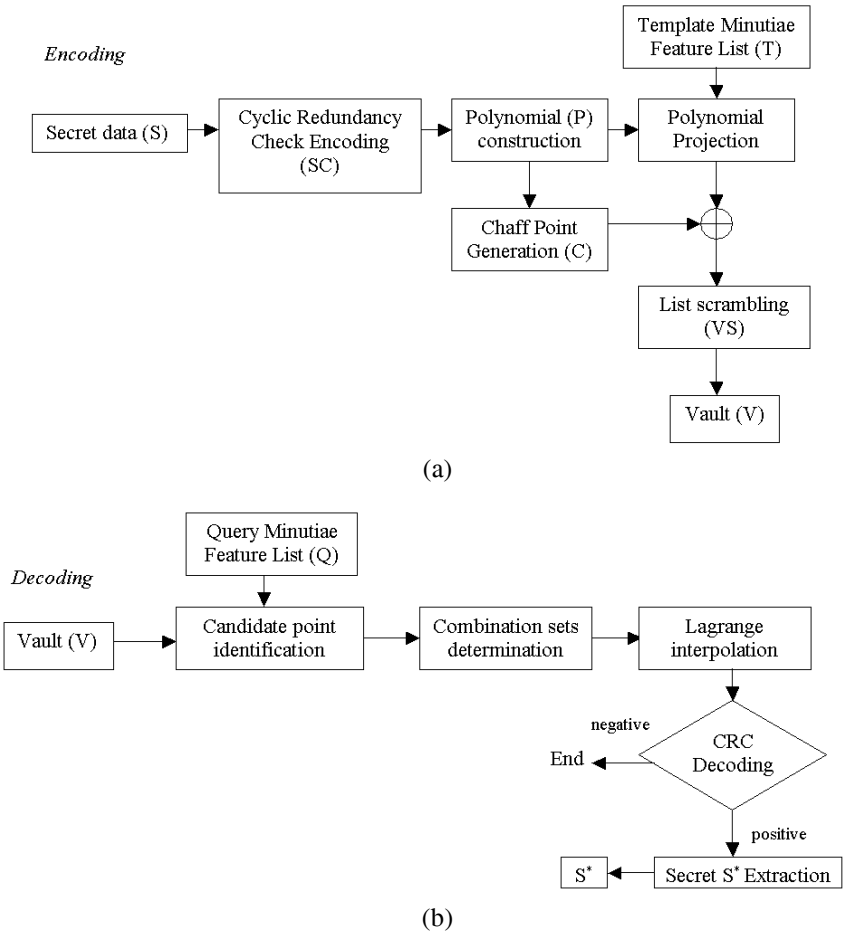


Fig. 2. Fuzzy fingerprint vault: (a) vault encoding, (b) vault decoding.

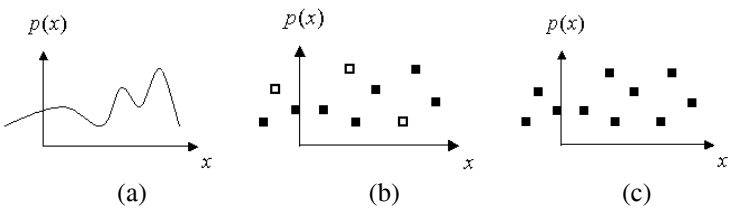


Fig. 3. System parameters: (a) polynomial, (b) evaluation of the polynomial (filled squares: genuine points, empty squares: chaff points), (c) final vault list

3.1 Encoding

Secret S is any data that needs to be protected, but the size of S that can be feasibly protected is limited by the capacity of the entity used for locking and unlocking the vault. Currently, we use x and y coordinates of minutiae points for locking/unlocking the vault. We first align the minutiae, namely compensate for the translation and rotation between template and query minutiae data. Encoding operation secures S with fingerprint minutiae data: if a query minutiae set similar to the template minutiae set is presented during decoding, it indicates the presence of an authorized person and S can be reconstructed accurately. Note that the vault operation is decoupled from any backend application (e.g., encryption/decryption using S): vault is only responsible for securing S with fingerprint data. The fingerprint template has the role of a key. Note that this is not the key in traditional cryptosystems (e.g., AES) per se: rather, it has the role of a key for a new cryptographic construct, namely the fuzzy vault. In the current implementation, S is generated as a 128-bit random bit stream. This can simulate securing AES symmetric encryption keys.

Our current decoding implementation does not include any error-correction scheme, as proposed by Juels and Sudan's [4], since there are serious difficulties to achieve error-correction with biometric data. Developing the necessary polynomial reconstruction via error-correction has not been demonstrated in the literature. Instead, our algorithm decodes many candidate secrets. To identify which one of these candidates is the actual secret, we need to put some structure into the secret S. By checking the validity of this structure during decoding, the algorithm can identify whether a given candidate secret is correct or not. Cyclic Redundancy Check (CRC) is a generalization of the simple parity bit checking. It is commonly used in communication channel applications for error detection where the errors are introduced due to channel noise. In our case using incorrect minutiae points during decoding will cause an incorrect polynomial reconstruction, resulting in errors. In the current implementation, we generate 16-bit CRC data from the secret S. Hence, the chance of a random error being undetected (i.e., failing to identify an incorrect decoding) is 2^{-16}. The 16-bit primitive polynomial, $g_{CRC}(a) = a^{16} + a^{15} + a^2 + 1$, we use for CRC generation is called "CRC-16" and is used for EBCDIC messages by IBM [12]. Appending the CRC bits to the original secret S (128-bits), we construct 144-bit data SC. From this point on, all operations take place in Galois fields with cardinality 65536, namely GF(2^{16}): we concatenate x and y coordinates of a minutiae (8-bits each) as $[x \mid y]$ to arrive at the 16-bit locking/unlocking data unit u. Note that to account for slight variations in minutiae data (due to nonlinear distortion), raw minutiae data are first quantized. Namely, each minutia is translated to lie in a square tessellation of the 2D image plane. For example, if the block size used in the tessellation is 7, any minutia that has x coordinate in the range [1, 7] is assumed to originate from x coordinate 4. This allows for ±3 pixel variations in the coordinates of template and query minutiae.

SC is used to find the coefficients of the polynomial p: 144-bit SC can be represented as a polynomial with 9 (144/16) coefficients in GF(2^{16}), with degree $D = 8$. Hence, $p(u) = c_8 u^8 + c_7 u^7 + \ldots + c_1 u + c_0$. Simply, SC is divided into non-overlapping

16-bit segments, and each segment is declared as a specific coefficient, c_i, $i = 0,1,2,...,8$. Note that this mapping method (from SC to c_i) should be known during decoding, where the inverse operation takes place: decoded coefficients (c_i^*) are mapped back to decoded secret SC*. Then, two sets composed of point pairs need to be generated. The first one, called genuine set G, is found by evaluating $p(u)$ on the template minutiae features (T). Starting with N template minutiae (if we have more than N minutia, we choose the first N sorted according to ascending u values), $u_1, u_2, ..., u_N$, we find $G = \{(u_1, p(u_1)), (u_2, p(u_2)), ..., (u_N, p(u_N))\}$. Note that the template minutiae are selected to be unique, namely, $u_i \neq u_k$, if $i \neq k$, $i = 1,2,...,N$, $k = 1,2,...,N$. The second set, called the chaff set C, determines the security of the system. Assuming we need to add M chaff points, we first generate M unique random points, $c_1, c_2, ..., c_M$ in the field GF(2^{16}), with the constraint that they do not overlap with $u_1, u_2, ..., u_N$, namely $c_j \neq u_i$, $j = 1,2,...,M$, $i = 1,2,...,N$. Then, we generate another set of M random points, $d_1, d_2, ..., d_M$, with the constraint that the pairs (c_j, d_j), $j = 1,2,...,M$ do not fall onto the polynomial $p(u)$. Chaff set C is then $C = \{(c_1, d_1), (c_2, d_2), ..., (c_M, d_M)\}$, where $d_j \neq p(c_j)$, $j = 1,2,...,M$. Union of these two sets, $G \cup C$, is finally passed through a list scrambler which randomizes the list, with the aim of removing any stray information that can be used to separate chaff points from genuine points. This results in vault set, $VS = \{(v_1, w_1), (v_2, w_2), ..., (v_{N+M}, w_{N+M})\}$. Along with VS, the polynomial degree D forms the final vault, V.

3.2 Decoding

Here, a user tries to unlock the vault V using the query minutiae features. Assuming that we have N (note that this number is the same as the number of genuine template minutiae in order to balance the complexity conveyed via the number of required access attempts to reveal the secret) query minutiae (Q), $u_1^*, u_2^*, ..., u_N^*$, the points to be used in polynomial reconstruction are found by comparing u_i^*, $i = 1,2,...,N$, with the abscissa values of the vault V, namely v_l, $l = 1,2,...,(M+N)$: if any u_i^*, $i = 1,2,...,N$ is equal to v_l, $l = 1,2,...,(M+N)$, the corresponding vault point (v_l, w_l) is added to the list of points to be used. Assume that this list has K points, where $K \leq N$. Now, for decoding a D-degree polynomial, $(D+1)$ unique projections are necessary. We find all possible combinations of $(D+1)$ points, among the list with size K. Hence, we end up with $C(K, D+1)$ combinations. For each of these combinations, we construct the Lagrange interpolating polynomial. For a specific combination set given as $L = \{(v_1, w_1), (v_2, w_2), ..., (v_{D+1}, w_{D+1})\}$, the corresponding polynomial is

$$p^*(u) = \frac{(u-v_2)(u-v_3)...(u-v_{D+1})}{(v_1-v_2)(v_1-v_3)...(v_1-v_{D+1})}w_1 + \frac{(u-v_1)(u-v_3)...(u-v_{D+1})}{(v_2-v_1)(v_2-v_3)...(v_2-v_{D+1})}w_2 + ...$$
$$... + \frac{(u-v_1)(u-v_2)...(u-v_D)}{(v_{D+1}-v_1)(v_{D+1}-v_2)...(v_{D+1}-v_D)}w_{D+1}$$

This calculation is done in GF(2^{16}) and yields $p^*(u) = c_8^* u^8 + c_7^* u^7 + ... + c_1^* u + c_0^*$. The coefficients are mapped back to the decoded secret SC*. For checking whether there are errors in this secret, we divide the polynomial corresponding to SC* with the CRC primitive polynomial, $g_{CRC}(a) = a^{16} + a^{15} + a^2 + 1$. Due to the definition of CRC, if the remainder is not zero, we are certain that there are errors. If the remainder is zero, with very high probability, there are no errors. For the latter case, SC* is segmented into 2 parts: the first 128-bits denote S* while the remaining 16-bits are CRC data. Finally, the system outputs S*. If the query minutiae list (Q) overlaps with template minutiae list (T) in at least $(D+1)$ points, for some combinations, the correct secret will be decoded, namely, S* = S will be obtained. This denotes the desired outcome when query and template fingerprints are from the same finger. Note that CRC is an error detection method, and it does not leak information that can be utilized by an imposter attacker (Bob). He cannot learn which one of the polynomial projections is wrong; hence he cannot separate genuine points from chaff points.

4 Experimental Results

We used the IBM-GTDB [10] fingerprint database (100 mated image pairs with 500 dpi resolution and approximately of size 300x400) for obtaining the results presented in this section. The minutiae coordinates are linearly mapped to 8-bit range (e.g., the values [0, 255]) for both row and column dimensions before using them in locking/unlocking the vaults. In this database a fingerprint expert has manually marked the minutiae in every image. Further, the expert also established the correspondence between minutiae in mating fingerprint images (two fingerprint images per finger). Using this database has many advantages since (i) the adverse effects of using an automatic (and possibly imperfect) minutiae extractor are eliminated, and (ii) minutiae correspondence has been established. Note that we specifically chose to use such a database because it allows us to establish the upper bound on the performance of the fuzzy fingerprint vault. In our fuzzy vault implementation, the pre-alignment of template and query minutiae sets is based on the correspondence marked by the expert. The translation and rotation parameters that minimize the location error (in the least squares sense) between the corresponding minutiae are found and the query minutiae sets are aligned with template minutiae sets. We randomly generated a 128-bit secret S, and after appending the 16 CRC bits, the resulting 144-bits are converted to the polynomial $p(u)$ as $p(u) = 60467u^8 + 63094u^7 + ... + 52482u^1 + 11995$. As explained in Section 3.1, 144-bit data is divided into 16-bit chunks, and each one of these chunks determines one of the 9 coefficients of $p(u)$. One hundred pairs of fingerprint images (of 100 users) from IBM-GTDB database are used for locking and unlocking the vaults with the following parameters: number of template and query

minutiae $N = 18$ (selected so that the number of candidate points will be enough for reconstruction), number of chaff points $M = 200$ (the effect of this number on the security of the method against attackers is given below), and the block size used for quantization of minutiae *x* and *y* coordinates is 7 pixels (determined experimentally; a larger value decreases the capacity of the vault, whereas a smaller value does not eliminate the minutiae variability). During decoding, 18 query minutiae selected $K = 12$ candidate points, on average. By evaluating the 9-element combinations of these candidate points, 79 of the 100 query fingerprints were able to successfully unlock the vault (that was locked by its mated fingerprint) and recover the secret S. The remaining 21 query fingerprints selected fewer than required number of genuine points (i.e., less than 9) during decoding, hence they were unable to unlock the vault. Hence, the False Reject Rate (FRR) of the proposed system is 0.21, for the cited system parameters (so, the Genuine Accept Rate is 0.79, i.e., 79%). The average number of point combinations required for a genuine user to unlock the vault was 201, which corresponds to 52 seconds of computation for a system with a 3.4 GHz processor. During decoding, many candidate secrets (201 on average) need to be extracted and evaluated, resulting in high time complexity. Further, the system is implemented in Matlab, contributing to high computational times. It is observed that, during decoding, CRC performed with 100% accuracy: it signaled an error *if and only if* there is an error in the decoded polynomial.

For evaluating the corresponding False Accept Rate (FAR), we tried to unlock the vaults with fingerprint templates that *were not* the mates of the templates that locked the vaults. Hence, we have 99 imposter unlocking attempts for a distinct finger, and totally 9900 (99x100) attempts for 100 users. Note that the unlocking templates are first aligned with the locking templates (to avoid giving an unfair disadvantage to imposter attempts), using the centers of mass of minutiae coordinates: the minutiae of unlocking template are translated in *x* and *y* dimensions till their center of mass coincides with the center of mass of locking templates. None of the 9900 attempts could unlock the vault. Hence, for this small database, experimental FAR is 0%.

We can also quantify the security of the system mathematically. Assume that we have an attacker who does not use real fingerprint data to unlock the vault; instead he tries to separate genuine points from chaff points in the vault using brute-force. The vault has 218 points (18 of them are genuine, remaining 200 are chaff); hence there are a total of $C(218,9) \approx 2.6 \bullet 10^{15}$ combinations with 9 elements. Only $C(18,9) = 48620$ of these combinations will reveal the secret (unlock the vault). So, it will take an average of $5.3 \bullet 10^{10}$ (=C(218,9)/C(18,9)) evaluations for an attacker to crack the vault; this corresponds to a computational time of 439 years for the previously used system with the 3.4 GHz processor.

5 Conclusions

After exploring the characteristics of a new cryptographic construct called fuzzy vault, we presented the results of its actual implementation using fingerprint minutiae data, without resorting to simulating the error-correction step. The vault performs as expected (namely, the genuine users can unlock the vault successfully, and the com-

plexity of attacks that can be launched by imposter users is high). It is shown that 128-bit AES keys can be feasibly secured using the proposed architecture. The limitations of our approach include high time complexity (due to the need for evaluating multiple point combinations during decoding). Currently, we are working on architectures for achieving automatic alignment within the fuzzy fingerprint vault.

References

1. W. Stallings, *Cryptography and Network Security: Principles and Practices*, 3. Ed., Prentice Hall, 2003.
2. NIST, Advanced Encryption Standard (AES), 2001. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf
3. A. Jain, R. Bolle, and S. Pankanti, Eds., *Biometrics: Personal Identification in Networked Society*, Kluwer, 1999.
4. A. Juels and M. Sudan, "A Fuzzy Vault Scheme", *Proc. IEEE Int'l. Symp. Inf. Theory*, A. Lapidoth and E. Teletar, Eds., pp. 408, 2002.
5. J.-P. Linnartz and P. Tuyls, "New Shielding Functions to Enhance Privacy and Prevent Misuse of Biometric Templates", *Proc. 4th Int'l Conf. Audio- and Video-based Biometric Person Authentication*, pp. 393-402, 2003.
6. A. Juels and M. Wattenberg, "A Fuzzy Commitment Scheme", In G. Tsudik, Ed., *Sixth ACM Conf. Computer and Comm. Security*, pp. 28-36, 1999.
7. S. Lin, *An Introduction to Error-Correcting Codes*, Prentice-Hall, 1970.
8. T. C. Clancy, N. Kiyavash, and D. J. Lin, "Secure Smartcard-Based Fingerprint Authentication", *Proc. ACM SIGMM 2003 Multim., Biom. Met. & App.*, pp. 45-52, 2003.
9. U. Uludag and A.K. Jain, "Fuzzy Fingerprint Vault", *Proc. Workshop: Biometrics: Challenges Arising from Theory to Practice*, pp. 13-16, 2004.
10. U. Uludag, S. Pankanti, S. Prabhakar and A. K. Jain, "Biometric Cryptosystems: Issues and Challenges", *Proc. IEEE*, vol. 92, no. 6, pp. 948-960, 2004.
11. A. K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An Identity Authentication System Using Fingerprints", *Proc. IEEE*, vol. 85, no. 9, pp. 1365-1388, 1997.
12. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2. Ed., Cambridge University Press, 1992.

Infrared Face Recognition by Using Blood Perfusion Data

Shi-Qian Wu[1], Wei Song[2], Li-Jun Jiang[1], Shou-Lie Xie[1],
Feng Pan[1], Wei-Yun Yau[1], and Surendra Ranganath[2]

[1] Institute for Infocomm Research, 119613, Singapore
{shiqian,ljjiang,slxie,efpan,wyyau}@i2r.a-star.edu.sg
http://www.i2r.a-star.edu.sg
[2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260
{eng10970,elesr}@nus.edu.sg

Abstract. This paper presents a blood perfusion model of human faces based on thermodynamics and thermal physiology. The target is to convert the facial temperature data which are liable to ambient temperature into consistent blood perfusion data in order to improve the performance of infrared (IR) face recognition. Our large number of experiments has demonstrated that the blood perfusion data are less sensitive to ambient temperature if the human bodies are in steady state, and the real data testing demonstrated that the performance by means of blood perfusion data is significantly superior to that via temperature data in terms of recognition rate.

1 Introduction

Over the last decades, researchers from multiple disciplines have endeavored to build machines capable of automatic face recognition. However, it has been shown in [1] that even the well-known commercial face recognition systems perform rather poor in less controlled situations. As a face acquired in visual images has significantly different appearance due to the large variations both in intrinsic (pose, expression, hairstyle etc) and extrinsic conditions (illumination, imaging system etc), it is difficult to find the unique characteristics for each face, and accordingly, it is not easy to develop a reliable system for face recognition by using visual images.

Recently, several researchers have proposed the use of facial thermography for identity recognition [2-5]. The fundamentals behind it are that IR images are independent of external illumination [2] and humans are homoiotherm who are capable of maintaining a constant temperature which is different from that of the surroundings [6]. Especially, it is indicated in [6] that thermal exchange is mainly performed by blood convection, and hence the thermal patterns of faces are irreproducible and hence unique [2, 3]. However, the term “homoiotherm” only indicates approximate constant temperature in deep body, whereas the skin temperature distribution may change from person to person, and from time to time [6]. Further, if considering the facial thermograms, greater variations are resulted from various internal (physiological psychological etc) and external (environmental, imaging etc) conditions. Such variations will severely affect the performance of IR recognition system, as shown in

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 320–328, 2005.
© Springer-Verlag Berlin Heidelberg 2005

[4, 5]. It is obvious that the influence of ambient temperature in IR images is similar to that of illumination in visual images.

Then, how to improve the performance of IR face recognition system under different environments? Essentially, the solution is to use only the physiological features which are likely invariant to changes of ambient conditions. In this paper, the blood perfusion rates are derived from appearance temperature distribution, and these physiological data are used for face recognition.

In the following Section, we will describe our model developed based on thermodynamics and thermal physiology. Section 3 demonstrates the experimental results, and comparisons with the results based on temperature data, followed by conclusions and future work in Section 4.

2 Blood Perfusion Model

The main target of this paper is to derive the blood perfusion model to eliminate or alleviate the effect of ambient temperature. Thus the following assumptions are made in the analysis:

- The only venue of heat exchange is at the skin surface;
- The testing subjects are in a steady state, and hence the deep body temperature is regarded as constant, and no thermal regulation is considered;
- The ambient temperature is lower than body temperature;
- Pathological (like fever, headache, inflammation etc) and psychological (like nervous, blush etc) conditions, are not considered;
- The camera is well calibrated, and accurate temperature can be measured;

In view of the heat transfer and thermal physiology under these assumptions, the skin surface can be described by the following heat equilibrium equation. [6]:

$$Q_r + Q_e + Q_f = Q_c + Q_m + Q_b \tag{1}$$

where Q represents the heat flux per unit area. The subscripts r, e and f stand for radiation, evaporation and convection respectively. These three terms on left hand are the outflows which point from the skin surface to the environment. The subscripts c, m and b stand for body conduction, metabolism and blood flow convection. These are the influx terms in the direction from the body to the skin surface.

2.1 Heat Transfer by Radiation

In most models of internal biological tissues, the contribution from intrinsic radiative heat transfer process is negligible. Considering a body with apparent temperature T_s, the radiation heat flux Q_r to environment with temperature T_e can be described by the Stefan-Boltzmann Law as follows:

$$Q_r = \varepsilon\sigma(T_s^4 - T_e^4) \tag{2}$$

where σ is the Stefan-Boltzmann constant, ε is the emissivity. As we concern the accurate measurement of body temperature without effect of ambient temperature, ε is selected as the body emissivity, i.e. $\varepsilon = 0.98$.

2.2 Heat Transfer by Air Convection

The heat exchange by air convection could be modeled according to Newton's law of cooling:

$$Q_f = h_f(T_s - T_e) \tag{3}$$

where h_f is the coefficient of air convection and is calculated as [7]:

$$h_f = k_f Nu/d \tag{4}$$

with k_f being the thermal conductivity, d is the characteristic length of the object, and *Nu* is the Nusselt number.

Further according to [7], *Nu* is determined by Prandtl number *Pr* and Grashof number *Gr* as follows:

$$Nu = A(Pr \cdot Gr)^M \tag{5}$$

where *A* and *M* are constants to be determined experimentally.

It is indicated in [6] that the *Pr* in air is close to unity, whereas *Gr* could be calculated by [8]:

$$Gr = g\beta(T_s - T_e)d^3/\nu^2 \tag{6}$$

In which g is the local gravitational acceleration; ν is the kinematic viscosity of air; β is the thermal expansion coefficient of air.

In [9], it is proposed to calculate β as follows:

$$\beta = -(\frac{1}{\rho_a})(\frac{\partial \rho_a}{\partial T_e})_P \tag{7}$$

Here, $(\frac{\partial \rho}{\partial T})_P$ indicates a derivative at constant pressure, ρ_a is the air density and can be calculated as follows [9]:

$$\rho_a = 1.2929 \cdot (\frac{273.13}{T_e}) \cdot (\frac{760 - 0.3783 \cdot eVp}{760}) \tag{8}$$

with eVp being the vapor pressure, and is computed as follows:

$$eVp = eVpSat \cdot RH \tag{9}$$

where $eVpSat$, the saturation vapor pressure, equals to $23.76mmHg$ at $T_e = 25^oC$ [10] and RH , the Relative Humidity, is an experimental value. Normally, we select $RH = 50\%$ for calculation and obtain $eVp = 11.88mmHg$. Substitute ρ_a back in Eq.(7), we have $\beta = 3.354 * 10^{-3} K^{-1}$.

Finally, overall equation for the convectional heat flux is presented as follows:

$$Q_f = AK_f d^{3M-1}(\mathrm{Pr}\, g\beta/\nu^2)^M (T_s - T_e)^{M+1} \tag{10}$$

2.3 Heat Transfer by Evaporation

Based on the second assumption, when no sweat secretion is considered, a continuous evaporative heat loss still takes place via the skin surface due to a diffusion of water

vapor from the skin [6]. However, the evaporative rate is normally low under equilibrium condition and is only a second- or higher-order effect [11]. Thus, the Q_e term is ignored for simplicity.

2.4 Heat Transfer by Deep Body Conduction

As the core temperature in the body is normally higher than the skin temperature, a net heat flux Q_c takes place from the body core to the skin surface, which can be described by the Fourier law:

$$Q_c = k_s(T_a - T_s)/D \tag{11}$$

in which, k_s is the coefficient of heat conductivity in the tissues, T_a is the core temperatures, and D is the distance from the body core to the skin surface.

2.5 Heat Transfer by Local Tissue Metabolism

The local tissue metabolic heat refers to heat production component by metabolic mechanism of the surface skin volume of interest. Generally, this term is neglected in most models or treated as a constant. Here we follow Pennes' model [12], and set $Q_m = 4.186 W \cdot m^{-2}$.

2.6 Heat Transfer by Blood Perfusion Convection

Many researches from biology and bioengineering have demonstrated that while conduction from body core provides continuous heat to the skin, the main mechanism of heat exchange to the skin is by circulation of blood [6]. When homeostasis requires that the body conserves heat, blood perfusion in the skin decreases by vasoconstriction, otherwise, blood perfusion increases by vasodilatation. Normally, the convection by blood perfusion is modeled as follows [6]:

$$Q_b = \alpha \rho_b c_b \omega_b (T_a - T_s) \tag{12}$$

where α is the ratio of countercurrent exchange, ρ_b and c_b are the density of the blood and the specific heat of the blood respectively, and ω_b is the blood perfusion rate. Therefore, according to above equations, we have the blood perfusion rate as follows:

$$\omega_b = \frac{\varepsilon\sigma(T_s^4 - T_e^4) + AK_f d^{3M-1}(\Pr g\beta/\nu^2)^M (T_s - T_e)^{M+1} - k_s(T_a - T_s)/D - 4.186}{\alpha\rho_b c_b (T_a - T_s)} \tag{13}$$

Using Eq.(13), a thermal image is converted into blood perfusion data. It reveals from Eq.(13) that an area of the skin with relatively high blood perfusion rate has a surface temperature higher than that of an area with lower perfusion.

3 Experimental Results

To evaluate the feasibility of blood perfusion data for face recognition, a variety of real facial thermograms under different scales, poses, periods, and ambient tempera-

tures have been collected. A face is automatically detected and recognized by our developed system RIFARS, a real-time IR face recognition system as described in [13]. The schematic diagram of the RIFARS is shown in Fig.1. The procedures for preprocessing, face detection and normalization are demonstrated in [13].The features are extracted by the PCA and FLD methods and the classifier uses the RBF neural network as shown in [14] for details. The performance is evaluated in terms of top recognition score, and the results are then compared with those by using temperature information.

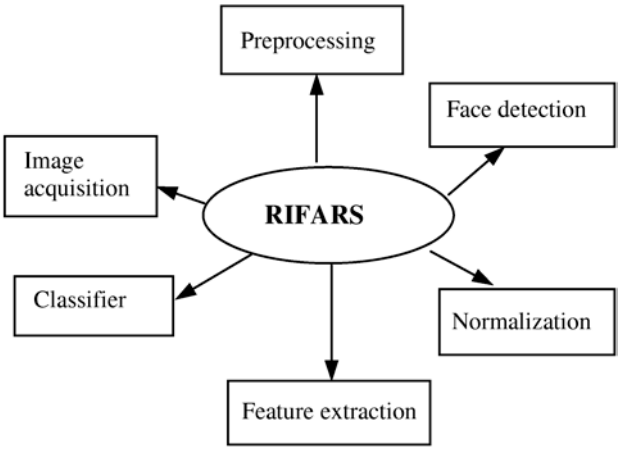


Fig. 1. Schematic diagram of the RIFARS

3.1 Database Collection

The IR data were captured by an IR camera ThermoVision A40 made by FLIR Systems Inc. This camera, which uses an uncooled microbolometer sensor with resolution of 320×240 pixels and the spectral response is 7.5 ~ 13 microns, is specially designed for accurate temperature measurement. The sensitivity is as high as $0.08\,^{o}C$. One of its prominent features is the function of automatic self-calibration to cope with the temperature drift. Further, we have a blackbody modeled MIKRON M340 to check and compensate the accuracy of measurement.

The database comprises 410 data of 41 individuals which were carefully collected at the same condition: i.e., same environment under air-conditioned control with temperature around 25.6 ~ $26.3\,^{o}C$, and each person stood at a distance of about 1 meter in front of the camera. Each person has 10 templates: 2 in frontal-view, 2 in up-view, 2 in down-view, 2 in left-view, and 2 in right-view. All the 10 images of each subject were acquired within 1 minute. As glass is opaque to IR, people are required to remove their eyeglasses.

3.2 Same-Session Data and Recognition Results

After captured the templates, each person was asked for testing immediately. The testing is similar to the watchlist scenario described in [1]: the testing person was

instructed to walk back and forth in front of the camera at a distance between 0.7m and 1.4m. He/She may wear glasses or not, with different poses and expressions. The threshold is set high to avoid being recognized so that the testing could be carried on. 20 testing data for each person were captured, but part of images (such as the images which are blurred, or contain part of a face), may be discarded in preprocessing. The recognition results by using temperature data or blood perfusion features for the same testing data are demonstrated in Table 1.

Table 1. Recognition rate for same-session data

Ambient condition	No of testing data	Data type	Recognition rate
Air-conditioned	707	Temperature	(543)[a] 76.8 %
25.3~26.2 ^{o}C		Blood perfusion	(637) 90.1 %

[a] The number of testing images to be recognized.

For same-session data, it is believed that the environmental condition, physiological and psychological features are invariant even thermal drift occurs in the camera. As the hairstyles are the same as those of templates, the segmentation error is small. During testing, each person was allowed to wear glasses. Our results demonstrated that the glasses are little effect on recognition rate for same-session data. The variations come from two aspects: one is facial shape due to changes of scale, expression, and pose, and another is image quality resulted from out-of-focus effect, which will slightly affect facial thermogram under the aforementioned constrain. The results shown in Table 1 illustrate that the blood perfusion data are less sensitive to shape variations than the thermal data.

3.3 Time-Lapse Data in Air-Conditioned Room and Recognition Results

In this experiment, the testing data were collected from July to November, 2004, all of which were captured in air-conditioned room from morning (24.5 ~24.8 ^{o}C), afternoon (25.7~26.3 ^{o}C) to night (25.4~25.8 ^{o}C, without lighting). The experimental results are shown in Table 2.

Table 2. Recognition rate for time-lapse data

Ambient condition	No of testing data	Data type	Recognition rate
Air-conditioned	1003	Temperature	(245) 24.43 %
24.5~26.3 ^{o}C		Blood perfusion	(867) 86.44 %

As the testing data are captured in air-conditioned room, it is considered that the testing individuals are in steady state without body temperature regulation. However, these time-lapse data comprise a variety of variations: ambient temperature, face shape (especially hair styles, poses), and psychology (for example, relaxed in morning, and tired in afternoon and at night). The effect of hair styles will lead to inconsistence in face normalization, and accordingly results in decrease of recognition rate. However, we found that one crucial factor came from psychology. It was shown that even the ambient temperature was almost the same, the face images collected when the person was overtired cannot be recognized at all, and the effect of psychology on

recognition rate varies from person to person. It is the key reason to affect the performance identified via blood perfusion data. For temperature data, it is shown that the performance decreases significantly for time-lapse data.

3.4 Time-Lapse Data with Eyeglasses in Air-Conditioned Room and Recognition Results

Also collected in air-conditioned room as shown in Section 3.3, but more challenges were put by wearing glasses in this experiment. The recognition rates are tabulated in Table 3.

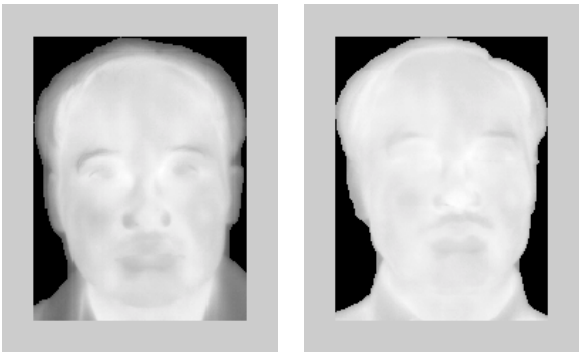
Table 3. Recognition rate for time-lapse data with eyeglasses

Ambient condition	No of testing data	Data type	Recognition rate
Air-conditioned	519	Temperature	(102) 19.65 %
24.6~26.2 ^{o}C		Blood perfusion	(317) 61.08 %

It is indicated in Section 3.2 that whether wearing glasses or not is not important for same-session data, because the face shape are invariant and segmentation of facial images is identical to the templates, and accordingly small variations of facial part have little effect on recognition. However, if the testing images contain a variety of variations, the eyeglasses will result in significantly decrease of performance.

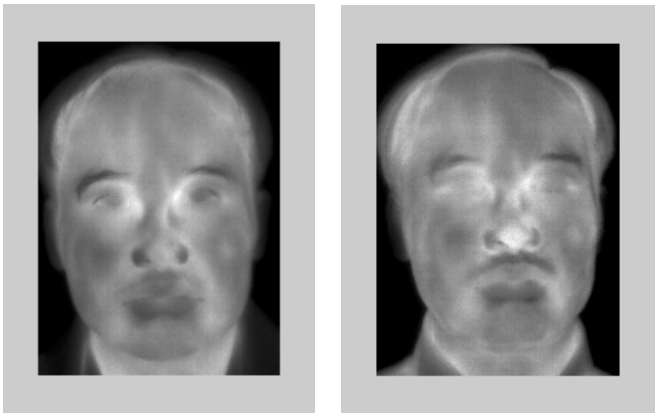
3.5 Time-Lapse Data with Different Environment and Recognition Results

In this experiment, two-session data were acquired under the environment without air-conditioned and wind (indoor). The ambient temperature of the first session was 29.3~29.5 ^{o}C, whereas the ambient temperature of the second session was 32.8 ~ 33.1 ^{o}C, which is much higher than that under which the database were captured. It is noted that these data were collected on the condition that the subjects had no sweat. Even in this case, the original temperature images of the identical subject have great variations as shown in Fig.2. The corresponding blood perfusion data are illustrated in Fig. 3.



(a) (b)

Fig. 2. Temperature images (a) Temperature image in database (T_e = 25.3~26.2 ^{o}C) (b) Testing temperature image (T_e = 32.8~33.1 ^{o}C)



(a) (b)

Fig. 3. Blood perfusion images (a) Corresponding to Fig.2(a), (b) Corresponding to Fig.2(b)

The recognition rates based on temperature information and blood perfusion feature respectively are illustrated in Table 4.

Table 4. Recognition rate under different ambient temperatures

Ambient condition	No of testing data	Data type	Recognition rate
29.4 ~29.5 ^{o}C , no wind	159	Temperature	(21) 13.21 %
		Blood perfusion	(142) 89.31 %
32.8~33.1 ^{o}C , no wind	226	Temperature	(37) 16.37%
		Blood perfusion	(33) 14.60%

It is revealed from Table 4 that the performance by means of blood perfusion data are comparable to that in air-conditioned environment, and greatly outperforms that via temperature data when the ambient temperature is 29.4 ~29.5 ^{o}C . However, the recognition rate drops significantly when the ambient temperature is 32.8~33.1 ^{o}C , even though we carefully segmented the face images. Our results demonstrated that the recognition is seriously dependent on testing subject. For example, the second-session data were collected from 5 persons. For blood perfusion data, only the first person was identified, whereas the second and the third persons were recognized by temperature In such environment, the human bodies are generally not in steady state, and temperature regulation, like sweating, has to take place to adjust body temperature.

4 Conclusions and Future Work

Up to now, developing a reliable face recognition system is still an open problem. As IR images are independent on external illumination, more researchers have been using IR images for face recognition. However, the variations in temperature for IR images are similar to the changes of illumination for visual images, and the IR data are severely determined by environmental temperature. In this paper, a mathematical

model of human skin based on thermodynamics and thermal physiology is derived to convert the facial temperature data into persistent physiological data. Our experiments illustrate that the performance by means of blood perfusion data significantly outperforms that by temperature data. Further analysis of the effect of ambient temperature on recognition, and extraction of unique features from blood perfusion data will be our future work.

References

1. Bone, M., and Blackburn D.: Face Recognition at a Chokepoint – Scenario Evaluation Results. Tech Report, DoD Counterdrug Technology Development Program Office (2002)
2. Socolinsky, D. A., Wolff, L. B., Neuheisel, J. D. and Eveland, C. K.: Illumination Invariant Face Recognition Using Thermal Infrared Imagery. Proc. IEEE Conf. Computer Vision & Pattern Recognition, Vol. 1, (2001) 527-534
3. Prokoski, F. J., Riedel, B. and Coffin, J.S.: Identification of Individuals by Means of Facial Thermography. Proc. IEEE Int. Conf. Security Technology, Crime Countermeasures (1992) 120-125
4. Yoshitomi, Y., Miyaura, T., Tomita, S. and Kimura, S.: Face Identification Using Thermal Image Processing. Proc. IEEE Int. Workshop Robot & Human Communication (1997) 374-379
5. Chen, X., Flynn, P. J. and Bowyer, K.W.: PCA-Based Face Recognition in Infrared Imagery: Baseline and Comparative Studies. Proc. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures, Nice, France (2003) 127-134
6. Houdas, Y. and Ring, E. F. J. Human Body Temperature: Its Measurement and Regulation. Plenum Press. New York (1982)
7. Earle, R. L. Unit Operations in Food Processing. Pergamon Press□1983□55-57
8. Available at: http://www.fact-index.com/g/gr/grashof_number_2.html
9. Weast, R. C., Astle, M. J. and Beyer, W. H. (eds.) CRC Handbook of Chemistry & Physics. 61st Ed, Florida, CRC Press□1980-1981□F-4-5
10. Saturated Vapor Pressure. Available at: http://hyperphysics.phy-astr. gsu.edu/ hbase/ kinetic/watvap.html
11. Love, T. J. Thermography as an Indicator of Blood Perfusion. In: Jain, R. K. and Gullino, P. M. Thermal Characteristics of Tumors: Applications in Detection and Treatment. Annals of New York Academy of Sciences, Vol. 335, (1980) 429-437
12. Pennes, H.H. Analysis of Tissue and Arterial Blood Temperatures in the Resting Human Forearm. J. Appl. Physiol. Vol. 1, (1948) 93-122
13. Wu, S-Q, Jiang L-J, Cheng, L., Wu, D., Wu, S., Xie, S-L, and Yeo, Allen C. B.: RIFARS: A Real-Time Infrared Face Recognition System. Asian Biometrics Workshop. Singapore (2003) 1-6
14. Er, M. J., Wu, S-Q., Lu, J., and Toh, H. L.: Face Recognition Using Radial Basis Function (RBF) Neural Networks. IEEE Trans. Neural Networks 13 (2002) 697-710

On Finding Differences Between Faces

Manuele Bicego[1], Enrico Grosso[1], and Massimo Tistarelli[2]

[1] DEIR - University of Sassari, via Sardegna 58 - 07100 Sassari - Italy
[2] DAP - University of Sassari, piazza Duomo 6 - 07041 Alghero (SS) - Italy

Abstract. This paper presents a novel approach for extracting characteristic parts of a face. Rather than finding a priori specified features such as nose, eyes, mouth or others, the proposed approach is aimed at extracting from a face the most distinguishing or dissimilar parts with respect to another given face, *i.e.* at "finding differences" between faces. This is accomplished by feeding a binary classifier by a set of image patches, randomly sampled from the two face images, and scoring the patches (or features) by their mutual distances. In order to deal with the multi-scale nature of natural facial features, a local space-variant sampling has been adopted.

1 Introduction

Automatic face analysis is an active research area, whose interest has grown in the last years, for both scientific and practical reasons: on one side, the problem is still open, and surely represents a challenge for Pattern Recognition and Computer Vision scientists; on the other, the stringent security requirements derived from terroristic attacks have driven the research to the study and development of working systems, able to increase the total security level in industrial and social environments.

One of the most challenging and interesting issue in automatic facial analysis is the detection of the "facial features", intended as characteristic parts of the face. As suggested by psychological studies, many face recognition systems are based on the analysis of facial features, often added to an holistic image analysis. The facial features can be either extracted from the image and explicitly used to form a face representation, or implicitly recovered and used such as in the PCA/LDA decomposition or by applying a specific classifier.

Several approaches have been proposed for the extraction of the facial features ([1–5], to cite a few). In general terms, all feature extraction methods are devoted to the detection of a priori specified features or gray level patterns such as the nose, eyes, mouth, eyebrows or other, non anatomically referenced, fiducial points. Nevertheless, for face recognition and authentication, it is necessary to also consider additional features, in particular those features that really characterize a given face. In other words, in order to distinguish the face of subject "A" from the face of subject "B", it is necessary to extract from the face image of subject "A" all features that are significantly different or even not present in face "B", rather than extract standard patterns.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 329–338, 2005.
© Springer-Verlag Berlin Heidelberg 2005

This paper presents a novel approach towards this direction, aiming at "finding differences" between faces. This is accomplished by extracting from one face image the most distinguishing or dissimilar areas with respect to another face image, or to a population of faces.

2 Finding Distinguishing Patterns

The amount of distinctive information in a subject's face is not uniformly distributed within its face image. Consider, as an example, the amount of information conveyed by the image of an eye or a chin (both sampled at the same resolution). For this reason, the performance of any classifier is greatly influenced by the uniqueness or degree of similarity of the features used, within the given population of samples. On one side, by selecting non-distinctive image areas increases the required processing resources, on the other side, non-distinctive features may drift or bias the classifier's response.

This assert is also in accordance with the mechanisms found in the human visual system. Neurophysiological studies from impaired people demonstrated that the face recognition process is heavily supported by a series of ocular saccades, performed to locate and process the most distinctive areas within a face [6–10].

In principle, this feature selection process can be performed by extracting the areas, within a given subject's face image, which are most dissimilar from the same areas in a "general" face. In practice, it is very difficult to define the appearance of a "general face". This is an abstract concept, definitely present in the human visual system, but very difficult to replicate in a computer system. A more viable and practical solution is to determine the face image areas which mostly differ from any other face image. This can be performed by feeding a binary classifier with a set of image patches, randomly sampled from two face images, and scoring the patches (or features) by their mutual distances, computed by the classifier. The resulting most distant features, in the "face space", have the highest probability of being the most distinctive face areas for the given subjects.

In more detail, the proposed algorithm extracts, from two face images, a set of sub-images centered at random points within the face image. The sampling process is driven to cover most of the face area[1]. The extracted image patches constitute two data sets of location-independent features, each one characterizing one of the two faces. A binary Support Vector Machine (SVM) [16, 17] is trained to distinguish between patches of the two faces: the computed support vectors define a hyperplane separating the patches belonging to the two faces. Based on the distribution of the image patches projected on the classifier's space, it is possible to draw several conclusions. If the patch projection "lies" very close to the computed hyperplane (or on the opposite side of the hyperplane), it means

[1] A similar image sampling model has been already used in other applications such as image classification (the so called patch-based classification [11–14]) or image characterization (the epitomic analysis proposed by Joijc and Frey in [15])

that the classifier is not able to use the feature for classification purposes (or it may lead to a misclassification). On the other hand, if the patch projection is well located on the subject's side of the hyperplane and is very far from the separating hyperplane, then the patch clearly belongs to the given set (*i.e.* to that face) and it is quite different from the patches extracted from the second face.

According to this intuition, the degree of distinctiveness of each face patch can be weighted according to the distance from the trained hyperplane. Since the classifier has been trained to separate patches of the first face from patches of the second face, it is straightforward to observe that the most important differences between the two faces are encoded in the patches far apart from the separating hyperplane (i.e. the patches with the highest weights).

In this framework the scale of the analysis is obviously driven by the size of the extracted image patches. By extracting large patches only macro differences are determined, loosing details, while by reducing the size of the patches only very local features are extracted, loosing contextual information. Both kinds of information are important for face recognition. A possible solution is to perform a multi scale analysis, by repeating the classification procedure with patches at different sizes, and then fusing the determined differences. The drawback is that each analysis is blind, because no information derived from other scales could be used. Moreover, repeating this process for several scales is computationally very expensive.

A possible, and more economic, alternative to a multi-scale classification, is to extract "multi-scale" patches, *i.e.* image patches which encode information at different resolution levels. This solution can be implemented by sampling the image patches with a log-polar mapping [18]. This mapping resembles the distribution of the ganglion cells in the human retina, where the sampling resolution is higner in the center (fovea) and decreases toward the periphery. By this resampling of the face image, each patch contains both low scale (high resolution) and contextual (low resolution) information.

The proposed approach for the selection of facial features consists of three steps:

1. two distinct and geometrically disjoint sets of patches are extracted, at random positions, from the two face images;
2. a SVM classifier is trained to define an hyperplane separating the two sets of patches;
3. for each of the two faces, the face patches are ranked according to the distances from the computed hyperplane.

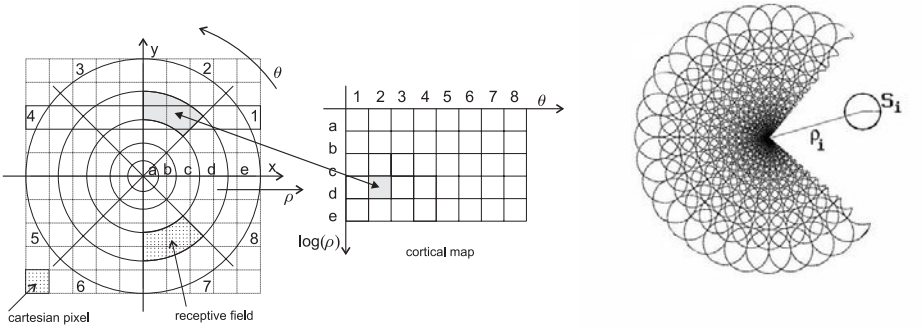
The processes involved by each step are detailed in the remainder of the paper.

2.1 Multi-scale Face Sampling

A number of patches are sampled from the original face image, centered at random points. The randomness in the selection of the patch center assures that

the entire face is analyzed, without any preferred side or direction. Moreover, a random sampling enforces a blind analysis without the need for a priori alignment between the faces.

The face image is re-sampled at each selected random point following a log-polar law [18]. The resulting patches represent a local space-variant remapping of the face image, centered at the selected point. The analytical formulation of the log-polar mapping describes the mapping that occurs between the retina (retinal plane (x, y)) and the visual cortex (log-polar or cortical plane ($\log(\rho)$, θ)). The derived logarithmic-polar law, taking into account the linear increment in size of the receptive fields, from the central region (fovea) towards the periphery, is described by the diagram in figure 1(a).



(a) (b)

Fig. 1. (a) Retino-cortical log-polar transformation. (b) Arrangement of the receptive fields in the retinal model

The log-polar transformation applied is the same described in [18] which differs from the models proposed in [19, 20]. The parameters required to define the log-polar sampling are: the number of receptive fields per eccentricity (N_a) and the radial and angular overlap of neighboring receptive fields (O_r and O_a).

For each receptive field, located at eccentricity ρ_i and with radius S_i, the angular overlap factor is defined by $K_0 = \frac{S_i}{\rho_i}$. The amount of overlapping is strictly related to the number of receptive fields per eccentricity N_a. In particular if $K_0 = \frac{\pi}{N_a}$ all receptive fields are disjoint. The radial overlap is determined by:

$$K_1 = \frac{S_i}{S_{i-1}} = \frac{\rho_i}{\rho_{i-1}}.$$

The two overlap parameters K_0 and K_1 are not independent, in fact:

$$K_1 = \frac{\rho_i}{\rho_{i-1}} = \frac{1 + K_0}{1 - K_0}.$$

As for the angular overlap, the radial overlap is not null only if:

$$K_1 < \frac{1 + K_0}{1 - K_0}.$$

Given the log-polar parameters N_a, O_r, O_a, K_0 and K_1 are computed as:

$$K_0 = \pi \frac{O_a}{N_a}, \qquad K_1 = \frac{O_r + K_0}{O_r - K_0}.$$

The image resolution determines the physical limit in the size of the smallest receptive fields in the fovea. This, in turn, determines the smallest eccentricity:

$$\rho_0 = \frac{S_0}{K_0}$$

Defining $\rho_0 \in [0.5 - 5]$, the original image resolution is preserved.

2.2 The SVM Classifier

In the literature Support Vector Machines have been extensively employed as binary classifiers in face recognition and authentication [21, 22], object classification [23], textile defects classification [24] and other applications as well.

The SVM classifier holds several interesting properties: quick training process [25], accurate classification, and, at the same time, a high generalization power [17]. Moreover, only two parameters need to be set: the regularization constant C and the size of the kernel for the regularization function.

In the proposed approach the *Radial Basis Function (RBF)* regularization kernel has been adopted, because it allows the best compromise between classification accuracy and generalization power. In order to obtain an acceptable generalization from the input data, the value of sigma has been carefully determined.

The set of log-polar image patches, sampled from each face image, are firstly vectorized and subsequently fed to a Support Vector Machine [16, 17]. As the SVM is a binary classifier, the data from the two subjects are used to build a set of support vectors able to distinguish them. Therefore, according to the procedure adopted to build a classifier for authentication purposes, the patches from one subject are used to represent the "client" class, while the patches from the second subject represent the "impostor" class.

2.3 Determining Face Differences

The SVM classifier, obtained from the input patches, defines an hyperplane separating the features belonging to the two subjects. The differences between the two subjects could be determined, for each correctly classified patch, from the absolute distance from the hyperplane: higher distances identify more characteristic facial features.

More formally, let $\mathcal{C}(\mathbf{x})$ be the class assigned by the trained SVM to an unknown patch $\mathbf{x}$, then:

$$\mathcal{C}(\mathbf{x}) = sign(f(\mathbf{x})) \tag{1}$$

where $f(\mathbf{x})$ represents the distance between the point $\mathbf{x}$ and the hyperplane represented by the SVM. When using a kernel $K(\mathbf{x}_i, \mathbf{x}_j)$, the distance $f(\mathbf{x})$ is computed as

$$f(\mathbf{x}) = b + \sum_{i=1}^{D} \alpha_i \mathcal{C}(\mathbf{x}_i) K(\mathbf{x}, \mathbf{x}_i) \tag{2}$$

where b and α_i are parameters determined in the training phase, and $\mathbf{x}_i$ are the points of the training set.

Given the trained SVM, the weight ω of the patch P_i belonging to the face k is computed as follows:

$$\omega(P_i) = \begin{cases} |f(P_i)| & \text{if } \mathcal{C}(P_i) = k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This analysis is repeated for both faces. It is important to note that the patches which are in the uncorrect side of the hyperplane are discarded (weight equal to 0), since the classifier could not provide any useful information about them (it is not able to correctly classify those patches).

3 Experimental Results

In order to verify the real applicability of the proposed method, two experiments were performed. In the first experiment a synthetic artifact (a black dot) is added to a face image and this is compared against the original image (see Fig. 2). In the second experiment two face images from two different subjects are compared (see Fig. 3). In both experiments gray level images were used, with a resolution of 310x200 pixels. The images have been re-sampled, at random positions, with 1000 log-polar patches. Each log-polar patch has a resolution of 23 eccentricities and 35 receptive fields for each eccentricity, with an overlap equal to 10% along the two directions. The *Radial Basis Function (RBF)* regularization kernel has been adopted for the SVM, with parameters $\sigma = 400$ and $C = 10$.

The results of the synthetic experiment is displayed in Fig. 2. To facilitate the understanding of the computed image differences, only the first ten patches with higher weights (distances from the computed hyperplane) are displayed. From the sequence of patches resulting in figure 2 the black dot is clearly identified.

In the experiment performed on two real face images, the 52 patches with higher distances for each face have been considered. The computed results are shown in Fig. 4 and 5.

In order to facilitate the visualization, similar patches have been grouped together, using the K-means method [26]. For the first face, six semantically different regions have been found, whereas in the second face nine different regions were considered. For each patch retained in the figure, the number of similar patches in the group is displayed. From these pictures some relevant differences between the two faces are detected. In the first face, for example, the forehead (both right and left part), the nose and the eyes are clearly identified. It is worth noting that also the fold of the skin on the right cheek is detected. As for the

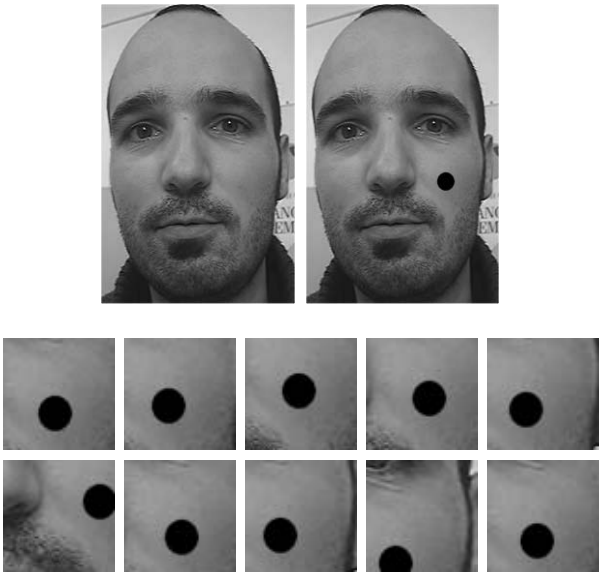


Fig. 2. Synthetic experiment. (top) The two images used in the experiment. (bottom) The 10 most weighted patches extracted when comparing the two faces. Only the patches related to the modified face are displayed



Fig. 3. (left) Original images used in the comparison experiment. (right) Random image points used for sampling the space-variant patches

second person (Fig. 5) the eyeglass are clearly identified as distinctive features (both right, left, upper and central parts). In fact, 27 out of the first 52 most weighted patches are located on them. Another distinctive pattern is the shape of the mouth, together with the chin, and the shape of the forehead.

As it can be noted, the extracted patterns seem to have some complementarities for the two faces. In fact, some distinctive areas are still present in both faces (regions around the eyes and the nose) while other distinctive and subtle details are preserved.

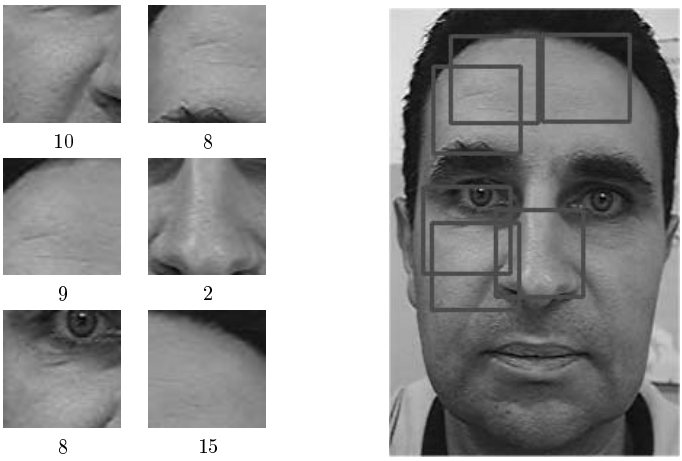


Fig. 4. Results of the detection of the most distinguishing features for the first face. Similar patches have been grouped together. (a) The representative patches (the number of components of each group is displayed below the patch) and (b) the location of the patches on the face

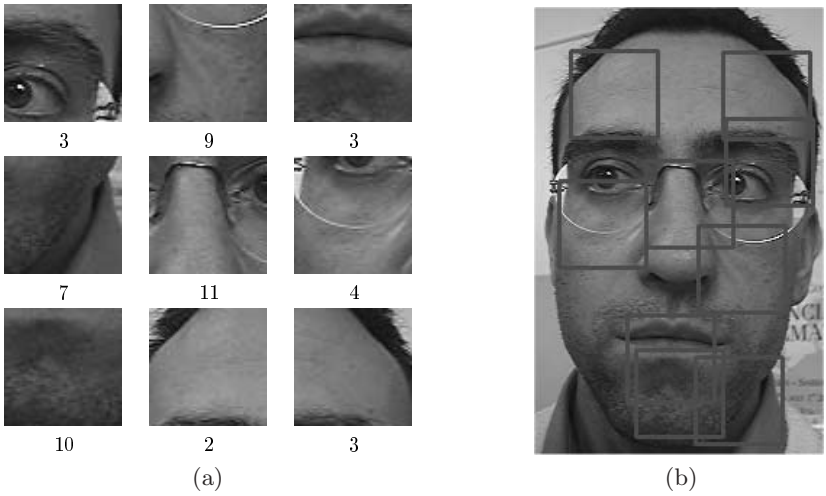


Fig. 5. Results of the detection of the most distinguishing features for the second face. Similar patches have been grouped together. (a) The representative patches (the number of components of each group is displayed below the patch) and (b) the location of the patches on the face

4 Conclusions

In this paper a new approach for finding differences between faces has been proposed. A Support Vector Machines classifier is trained to distinguish between

two sets of space-variant patches, randomly extracted from two different face images. The "distinctiveness" of each patch is computed as the distance from the separating hyperplane computed from the support vectors.

Even though the experiments performed are very preliminary, already demonstrate the potential of the algorithm in determining the most distinctive patterns in the analyzed faces. The proposed approach can be very effective to tailor the face representation according to the most distinctive features of a subject's face, for recognition purposes.

A future development of this research includes the combination of the extracted features, which could be performed by "back propagating" the patches weights to the face, to build a true "difference map".

Another interesting issue is the comparison of more than two faces, i.e. finding the differences between a given face and the rest of the world. In this way it may be possible to extract the general characteristic features of any given face. This can be achieved by choosing the negative examples in the SVM training as formed by all patches randomly sampled from several different faces. A further issue could be the investigation of different sampling techniques, *i.e.* methods that could reduce the number of samples needed to significantly cover the whole face.

Acknowledgments

This work has been supported by funds from the 6th Framework Programme European Network of Excellence "BioSecure".

References

1. Craw, I., Tock, D., Bennett, A.: Finding face features. In: Proc. of European Conf. Computer Vision. (1992) 92–96
2. Graf, H., Chen, T., Petajan, E., Cosatto, E.: Locating faces and facial parts. In: Proc. Int. Workshop Automatic Face and Gesture Recognition. (1995) 41–46
3. Campadelli, P., Lanzarotti, R.: Fiducial point localization in color images of face foregrounds. Image and Vision Computing **22** (2004) 863–872
4. Ming-Hsuan, Y., Kriegman, D., Ahuja, N.: Detecting faces in images: a survey. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 34–58
5. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys **35** (2003) 399 – 458
6. Goren, C., Sarty, M., Wu, P.: Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics **56** (1975) 544–549
7. Yarbus, A.: Eye movements and vision. Plenum Press, New York (1967)
8. Nahm, F., Perret, A., Amaral, D., Albright, T.: How do monkeys look at faces? Journal of Cognitive Neuroscience **9** (1997) 611–623
9. Haith, M., Bergman, T., Moore, M.: Eye contact and face scanning in early infancy. Science **198** (1979) 853–854
10. Klin, A.: Eye-tracking of social stimuli in adults with autism (2001) Paper presented at the meeting of the NICHD Collaborative Program of Excellence in Autism. Yale University, New Haven, CT.

11. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: Proc. European Conf. on Computer Vision. Volume 4. (2002) 113–130
12. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition. Volume 2. (2003) 264
13. Dorko, G., Schmid, C.: Selection of scale-invariant parts for object class recognition. In: Proc. Int. Conf. on Computer Vision. Volume 1. (2003) 634–640
14. Csurka, G., Dance, C., Bray, C., Fan, L., Willamowski, J.: Visual categorization with bags of keypoints. In: Proc. Workshop Pattern Recognition and Machine Learning in Computer Vision. (2004)
15. Jojic, N., Frey, B., A.Kannan: Epitomic analysis of appearance and shape. In: Proc. Int. Conf. on Computer Vision. Volume 1. (2003) 34–41
16. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
17. Burges, C.: A tutorial on support vector machine for pattern recognition. Data Mining and Knowledge Discovery **2** (1998) 121–167
18. Grosso, E., Tistarelli, M.: Log-polar stereo for anthropomorphic robots. In: Proc. European Conference on Computer Vision. Volume 1., Springer-Verlag (2000) 299–313
19. Braccini, C., Gambardella, G., Sandini, G., Tagliasco, V.: A model of the early stages of the human visual system: Functional and topological transformation performed in the peripheral visual field. Biol. Cybern. **44** (1982) 47–58
20. Tistarelli, M., Sandini, G.: On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. IEEE Trans. on Pattern Analysis and Machine Intellingence **15** (1993) 401–410
21. Jonsson, K., Kittler, J., Li, Y.P., Matas, J.: Support vector machines for face authentication. In: Proc. of Brit. Machine Vision Conf., Nottingham, UK. (1999) 543–553
22. Bicego, M., Iacono, G., Murino, V.: Face recognition with Multilevel B-Splines and Support Vector Machines. In: Proc. of ACM SIGMM Multimedia Biometrics Methods and Applications Workshop. (2003) 17–24
23. Pontil, M., Verri, A.: Support vector machines for 3-D object recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 637–646
24. Murino, V., Bicego, M., Rossi, I.: Statistical classification of raw textile defects. In: Proc. of IEEE Int. Conf. on Pattern Recognition. Volume 4. (2004) 311–314
25. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods - Support Vector Learning (1998)
26. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (1999)

Face Detection Using Look-Up Table Based Gentle AdaBoost

Cem Demirkır and Bülent Sankur

Boğaziçi University, Electrical-Electronic Engineering Department,
80815 Bebek, İstanbul
{cemd,sankur}@boun.edu.tr
http://busim.ee.boun.edu.tr

Abstract. In this work, we propose a face detection method based on the Gentle AdaBoost algorithm which is used for construction of binary tree structured strong classifiers. Gentle AdaBoost algorithm update values are constructed by using the difference of the conditional class probabilities for the given value of Haar features proposed by [1]. By using this approach, a classifier which can model image classes that have high degree of in-class variations can be constructed and the number of required Haar features can be reduced.

1 Introduction

Classical object detection schemes necessitate complex and computationally heavy classifiers for face detection in gray level images. Since the face detector is applied at each location and scale it requires significant computational power. However most of the locations in the searched scene do not contain any face and in fact the odds of finding a face at any one location is very small. Thus most of the non-face blocks can be eliminated with very simple classifiers. To find an efficient alternative to this exhaustive search approach, a rejection based classification approach is used to eliminate rapidly non-face regions in an image. Based on such a rejection strategy, Viola and Jones [1] used cascaded non-face rejection stages with low face elimination rate. Their algorithm consisting of simple classifiers are based on easily computable Haar features yield good detection with low false alarm rate. The required number of Haar features using their approach depends on the target false alarm rate. Liehart [6] made performance comparison of boosting algorithms using haar feature based binary-output weak classifiers.

Various extensions of this detector structure have been proposed in [2], [3] and [4]. For example Wu [4] has proposed a multi-view face detector using Real AdaBoost confidence-rated Look-Up-Table (LUT) classifiers to detect faces under rotation and pose variations. Our work is in the same line as the Viola-Jones scheme. The contribution consists in the use of a simple real valued Gentle AdaBoost (GAB) algorithm procedure to construct cascaded classifier structure. The GAB approach helps to reduce the required number of Haar features vis-a-vis the boosting approach that instead uses binary output weak classifiers. Using

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 339–345, 2005.
© Springer-Verlag Berlin Heidelberg 2005

LUT based confidence values for each Haar feature the information wasted in the binary weak classifier DAB approach can be utilized in the cascaded classifier training. We also used classifier output propagation as proposed in [4] to further reduce the number of features.

The rest of the paper is organized as follows: in Section 2 we define the GAB procedure. In Section 3, we define the GAB based Haar feature selection and strong classifier construction. In Section 4, we show we can the use of previous classifier output in the following classifier construction. In Section 5 we give experiments and results, and finally in Section 6 the conclusions.

2 GAB Algorithm

Boosting is a classification methodology which applies sequentially reweighted versions of the input data to a classifier algorithm, and taking a weighted majority vote of sequence classifiers thereby produced. At each application of the classification algorithm to the reweighted input data, classification algorithm finds an additional classifier $f_m(x)$ at stage m. GAB algorithm is a modified version of the Real AdaBoost (RAB) algorithm and it is defined in Figure 1. The main difference between GAB and RAB is the way it uses the estimates of the weighted class probabilities to update the weak classifier functions, $f_m(x)$.In GAB the update is given by $f_m(x) = P_w(y = 1|x) - P_w(y = -1|x)$, while in the RAB algorithm is given by half the log-ratio $f_m(x) = \frac{1}{2} log \frac{P_w(y=1|x)}{P_w(y=-1|x)}$ [5]. Log-ratios can be numerically unstable, leading to very large update values, while the update in GAB lies in the range [-1,1]. This more conservative algorithm has classification performance similar to RAB algorithm, and outperforms it both especially when stability problems arise.

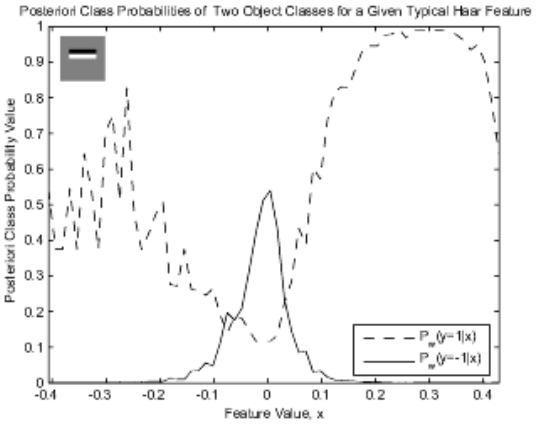
GAB Algorithm

1. Start with weights $w_i = 1/N, i = 1, ..., N, F(x) = 0$.
2. Repeat for $m = 1, 2, \ldots, M$.
   (a) Estimate $f_m(x)$ by weighted least squares fitting of y to x.
   (b) Update $F(x) \leftarrow F(x) + f_m(x)$
   (c) Set $w_i \leftarrow w_i \exp[-y_i . f_m(x_i)], i = 1, 2, \ldots, N$, and renormalize so that $\sum_i w_i = 1$.
3. Output the classifier sign $[F(x)] = \text{sign}\left[\sum_{m=1}^{M} f_m(x)\right]$.

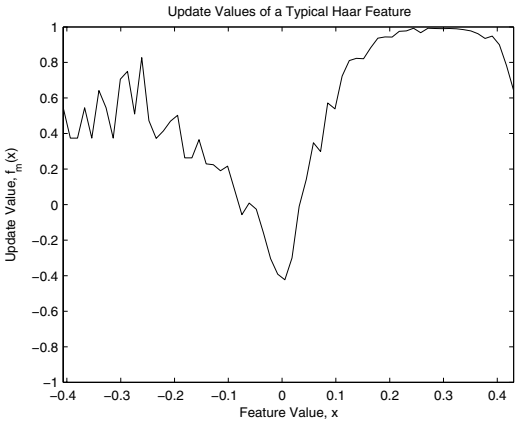
Fig. 1. The GAB algorithm allows for the estimator $f_m(x)$ to range over real numbers

3 GAB Based Haar Feature Selection

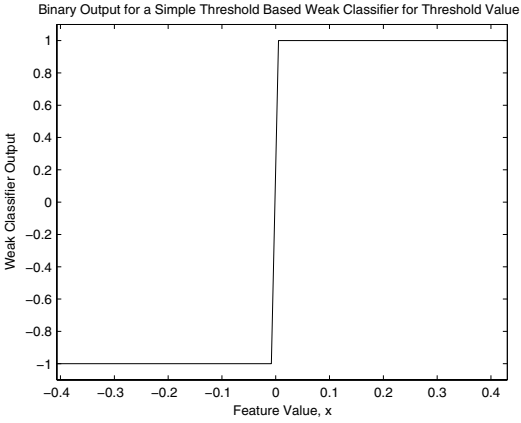
In the original algorithm the simple weak classifiers are built by simply comparing the Haar features to a threshold and thereby producing binary outputs. The feature and its threshold are selected to yield the minimum error at every stage.



(a) Likelihood values of a typical chosen Haar feature by GAB algorithm for two object classes. The corresponding Haar feature is shown in the upper left of the figure



(b) GAB update values $f_m(x) = P_w(y = 1|x) - P_w(y = -1|x)$



(c) Binary output of the classifier in the case a simple threshold based weak classifier is used

Fig. 2. Comparison of GAB update scheme (Fig. a,b) to the simple binary output scheme (Fig. c)

In contrast, in GAB based feature selection mechanism the weak classifiers are not forced to yield the binary outputs, instead they give the values of the update functions $f_m(x)$ at finite number of samples. The comparison of the update values for the binary case and GAB-based case is shown for a typical chosen Haar feature in Figure 2. We use the GAB training algorithm described in Figure 1 to construct a strong stage classifier using confidence values $f_m(x)$ for each feature. Under equal prior probabilities for the object classes, the update $f_m(x)$ is given by

$$\begin{aligned} f_m(x) &= P_w(y=1|x) - P_w(y=-1|x) \\ &= \frac{P_w(x|y=1)P(y=1) - P_w(x|y=-1)P(y=-1)}{P_w(x|y=1)P(y=1) + P_w(x|y=-1)P(y=-1)} \\ &= \frac{P_w(x|y=1) - P_w(x|y=-1)}{P_w(x|y=1) + P_w(x|y=-1)} \end{aligned} \tag{1}$$

where $P_w(x|y=\pm 1)$ are the likelihood values computed by using histograms of feature values x for two different object hypothese. Histogram bins are updated by summing the sample weights of the training set. The subscript w denotes the likelihood values with respect to updated sample weights at each boosting round.

A stage of Haar feature classifier construction using GAB

1. Start with weights $w_i = 1/2p$ and $1/2l$ where p and l are the number of positive and negatives class samples.
2. Repeat for $m = 1, 2, \ldots, M$.
   (a) For each Haar feature j, $f_m(x) = P_w(y=1|x) - P_w(y=-1|x)$ using only the feature j values.
   (b) Choose the best feature confidence set of values $f_m(x)$ giving the minimum weighted error $e_m = E_w\left[1_{(y_i \neq sign[f_m(x_i)])}\right]$ for all feature j.
   (c) Update $F(x) \leftarrow F(x) + f_m(x)$
   (d) Set $w_i \leftarrow w_i \exp\left[-y_i.f_m(x_i)\right], i = 1, 2, \ldots, N$, and renormalize so that $\sum_i w_i = 1$.
3. Output the classifier sign $[F(x)] = \text{sign}\left[\sum_{m=1}^{M} f_m(x)\right]$.

Fig. 3. At each iteration AdaBoost finds a set of update values $f_m(x)$ which use the values of the feature corresponding to the minimum error

GAB algorithm chooses the best Haar feature resulting with the minimum weighted error $e_m = E_w\left[1_{(y_i \neq sign[f_m(x_i)])}\right]$ from among all available features. The chosen update values are accumulated in the classifier output $F(x)$ value. The output of the classifier, $F_m(x)$, is thresholded by T_m such that the desired target false alarm rate, f, and detection rate, d , of the classifier is achieved. We also used the approach in [4], where the previous stage classifier output,$F_m(x)$, is inputted to the next stage classifier. According to this, in the training process,

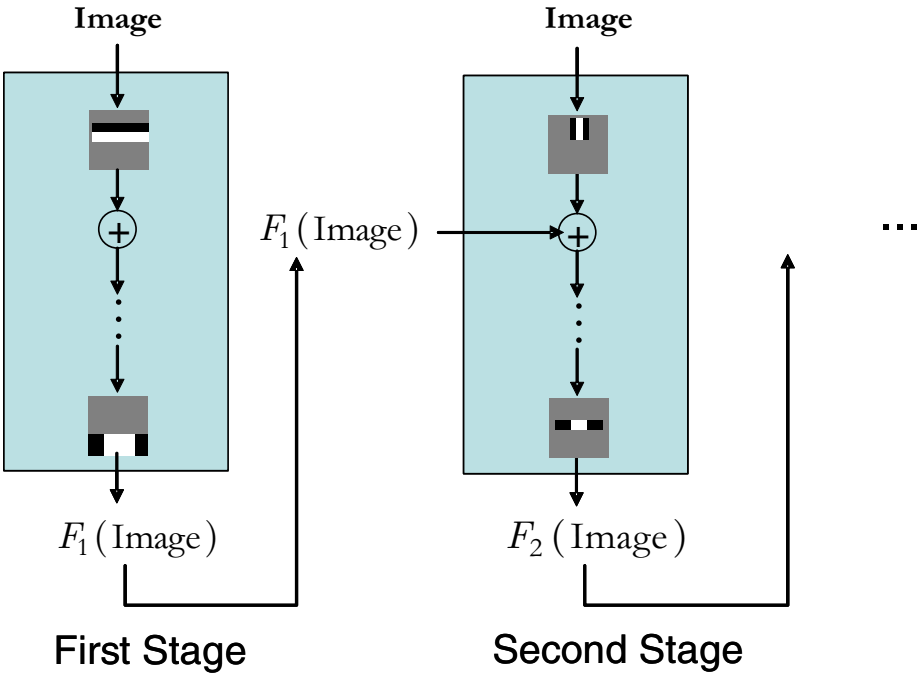


Fig. 4. Propagating the classifier output to the next stage classifier, the previous stage classifier outputs are used to compute the posterior class probabilities of the object classes, $P_w(x|y = 1)$ and $P_w(x|y = -1)$, the update values computed from them is used as an update corresponding to the first feature of the classifier

classifier output values for the training samples of the next stage is used to prepare the posteriori class probabilities of the object classes. Thereby the first update value $f_1(x)$ value of the succeeding classifier is computed by using the histogram values resulting from the classifier outputs of the previous stage. In this sense each previous stage conveys its accumulated information to the next stage classifier. A cascaded classifier is produced by this kind of GAB procedure as illustrated in the Fig 4.

4 Experiments

Figure 5 illustrates a chosen Haar feature, its corresponding posterior class probabilities as a function of feature values. The rightmost figure shows the classifier output as more and more classifiers are added, abscissa indicates the training sample index. Adding more features, samples of two classes can be separated from each other by simply thresholding the final classifier output $F(x)$.

We implemented GAB algorithm for near frontal face detection in cluttered scenes. For training we used about 10000 aligned face and 6000 non-face images with the size of 24x24 at each stage training. Face images contain near frontal

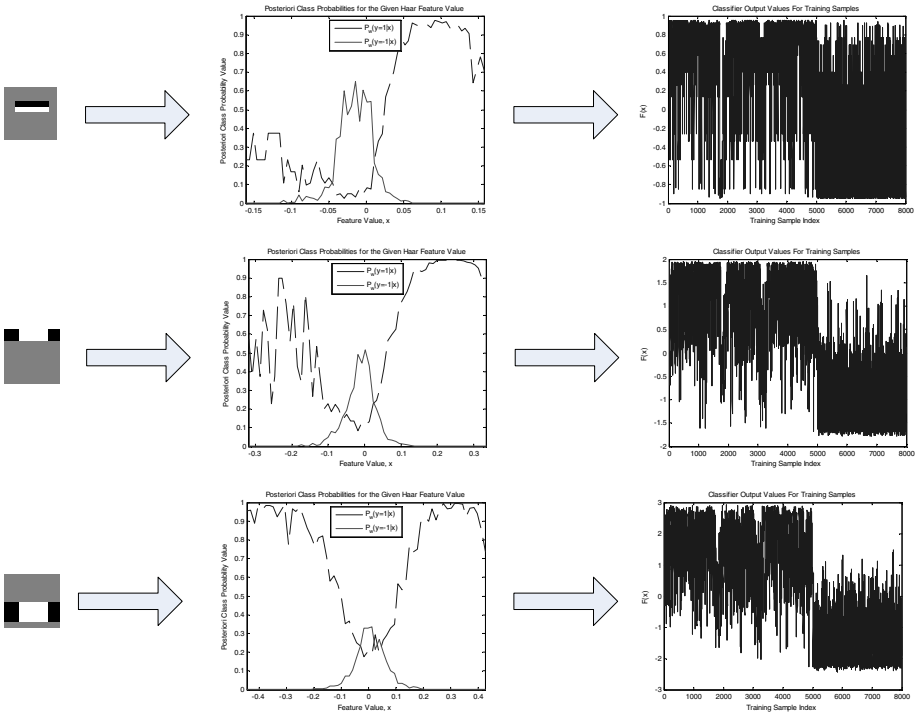


Fig. 5. First column shows the Haar features selected by GAB, second column shows the posterior class probabilities of the object classes, $P_w(x|y=1)$ and $P_w(x|y=-1)$, for the corresponding Haar feature, the last column shows classifier output values $F(x)$ values for the training samples after being updated with each $f_m(x_i)$

faces which are subject to $\pm 15°$ inplane rotations and in the range of $\pm 30°$ out-of-plane rotations. The number of points to sample the GAB update functions, $f_m(x_i)$, used in our work is 64 and the false alarm probability was chosen as 10^{-6}. These values are stored for each selected feature. In test process, these stored Look-Up Table (LUT) values of the update functions are used to compute the confidence values of each computed feature values. On the same training set, we trained the cascaded structure by using the proposed LUT-based GAB procedure and threshold-based Discrete AdaBoost (DAB) approach . The number of total features produced by GAB based procedure is about 20 percent of the DAB based training case for the same overall target detection and false alarm rate. We tested two methods on the CMU test set containing 503 faces in the 130 images. The number of features and False Alarm (FA)/Detection Rates (DR) are given in Table 1 for the two methods.

As seen from Table 1, not only the LUT-based GAB method performance is higher than the DAB method and also requires much fewer feature as compare to the DAB method, but about one fifth of the number of features the detection performance is even better.

Table 1. Performance and number of feature comparison for the methods. FA : False Alarm, DR : Detection Ratio

Method	Number of FAs/DR	Number of total features
LUT-based GAB	15/85.2%	531
DAB	15/80.1%	2652

5 Conclusions

In this work, we developed a simple and efficient LUT-based GAB training procedure using Haar like features for the near frontal face detection. We tested and compared two methods on the public common test set, CMU test set. This procedure necessitates significantly fewer features with respect to the DAB-based training case for the near frontal face detection problem. We plan to use this approach to implement a rotation invariant multi-view face detection system.

References

1. Viola, P., Jones, M.: Rapid Object Detection Using A Boosted Cascade of Simple Features. IEEE Conference on Computer Vision and Pattern Recognition, (2001).
2. Li, S. Z.: Statistical Learning of Multi View Face Detection. ECCV 2002, Copenhagen, Denmark.
3. Viola, P., Jones, M.: Fast Multi View Face Detection. Technical Report TR2003-96, July 2003, Mitsubishi Electric Research Laboratories.
4. Wu, B., Ai, H., Huang, C.: Fast Rotation Invariant Multi-View Face Detection based on Real Adaboost. AFG'04, May 17-19, 2004, Seoul, Korea.
5. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic regression: a Statitical View of Boosting. Technical Report, Standford University, August 17, 1998.
6. R. Liehart, A. Kuranov, V. Pisarevsky, Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection, MRL Tech report, 2002.

Post-processing on LDA's Discriminant Vectors for Facial Feature Extraction

Kuanquan Wang[1], Wangmeng Zuo[1], and David Zhang[2]

[1] Department of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China
cswmzuo@163.com
[2] Biometrics Research Centre, Department of Computing,
The Hong Kong Polytechnic University, Kowloon, Hong Kong
csdzhang@comp.polyu.edu.hk

Abstract. Linear discriminant analysis (LDA) based methods have been very successful in face recognition. Recently, pre-processing approaches have been used to further improve recognition performance but few investigations have been made into the use of post-processing techniques. This paper intends to explore the feasibility and efficiency of the post-processing technique on LDA's discriminant vectors. In this paper we propose a Gaussian filtering approach to post-process the discriminant vectors. The results of our experiments demonstrate that, post-processing technique can be used to improve recognition performance.

1 Introduction

As an important issue in face recognition system, facial feature extraction can be classified in two categories, geometric or structural methods and holistic methods [1]. So far, holistic methods, which use the whole face region as the input, have been a major facial feature extraction approaches and among various holistic methods, the state-of-art approaches are principal component analysis (PCA) and linear discriminant analysis (LDA) based methods [2, 3].

Recently, pre-processing approaches have been introduced to further improve the recognition performance of PCA and LDA-based methods. 2D-Gabor filters [4], edge detection [5] and wavelet techniques [6] have been used for facial image pre-processing before the application of PCA or LDA. Most recently, Wu et al. proposed to apply the projection-combined version of the original image for PCA [7].

Unlike the pre-processing techniques, few works have dealt with the use of post-processing to improve recognition performance. Precious work has shown that, LDA's discriminant vectors are very noisy and wiggly [8]. One general approach to address this problem is to add a penalty matrix to the within-class covariance matrix [9]. Since the discriminant vector can be mapped into image, in this paper we believe that appropriate image post-processing techniques can also be used to address this problem. To validate this view, we propose to use a Gaussian filtering approach to post-process discriminant vectors and carry out a series of experiments to test the effectiveness of the post-processing.

The remainder of this paper is organized as follows. In Section 2, we briefly review two representative LDA-based approaches, Fisherfaces and D-LDA. Section 3

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 346–354, 2005.
© Springer-Verlag Berlin Heidelberg 2005

presents the proposed Gaussian filtering method that is used to post-process discriminant vectors. In Section 4, the ORL and FERET database is used to evaluate the proposed Gaussian filtering method. Section 5 offers our conclusion.

2 The LDA-Based Facial Feature Extraction Methods

Linear discriminant analysis is an effective feature extraction approach used in pattern recognition [10]. It finds the set of the optimal vectors that map features of the original data into a low-dimensional feature space in such a way that the ratio of the between-class scatter to the within-class scatter is maximized. When LDA is applied to facial feature extraction, the recognition performance would be degraded due to the singularity of the within-class scatter matrix S_w. To date, a considerable amount of research has been carried out on this problem.

Although the aim of this paper is to study the effectiveness of the post-processing, it is not possible to test the effect of the post-processing on all the LDA-based methods. Consequently, we reviewed only two representative approaches, Fisherfaces [3] and D-LDA [11].

2.1 Fisherfaces

The Fisherfaces method is essentially linear discriminant analysis in a PCA subspace. When using Fisherfaces, each image is mapped into a high dimensional vector by concatenating together the rows of the original image. Let $\boldsymbol{X}$ denotes a training set $X = \{x_1^{(1)}, x_2^{(1)}, \cdots, x_{N_1}^{(1)}, x_1^{(2)}, x_2^{(2)}, \cdots, x_{N_2}^{(2)}, \cdots, x_j^{(i)}, \cdots, x_{N_C}^{(C)}\}$, where C is the number of classes, N_i is the number of training samples of class i, and $x_j^{(i)}$ is the jth vector of class i. Thus the total scatter matrix S_t is defined as

$$S_t = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_j^{(i)} - \bar{x})(x_j^{(i)} - \bar{x})^T , \tag{1}$$

the within-class scatter matrix is

$$S_w = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_j^{(i)} - \overline{x^{(i)}})(x_j^{(i)} - \overline{x^{(i)}})^T , \tag{2}$$

and the between-class scatter matrix is

$$S_b = \sum_{i=1}^{C} N_i (\overline{x^{(i)}} - \bar{x})(\overline{x^{(i)}} - \bar{x})^T , \tag{3}$$

where $\overline{x^{(i)}}$ is the mean vectors of class i, and $\bar{x}$ is the mean vectors of all training samples. The PCA projection $T_{pca} = [\nu_1, \nu_2, \cdots, \nu_{N-C}]$ can be obtained by calculating the eigenvalues and vectors of the total scatter matrix S_t, where $N = \sum_{i=1}^{C} N_i$ is the total number of training samples, and ν_k is the eigenvector corresponding to the kth largest eigenvalue of S_t. Then LDA projection T_{lda} is obtained by maximizing the Fisher's criteria

$$T_{lda} = \arg\max_{W} \frac{\left| W T_{pca}^T S_b T_{pca} W \right|}{\left| W T_{pca}^T S_w T_{pca} W \right|}. \tag{4}$$

Thus the final projection T_f used for feature extraction is

$$T_f^T = T_{lda}^T * T_{pca}^T. \tag{5}$$

2.2 D-LDA

D-LDA is another representative LDA-based method that has been widely investigated in facial feature extraction [11, 12]. The key idea of the D-LDA method is to find a projection that simultaneously diagonalizes both the between-class scatter matrix S_b and the within-class scatter matrix S_w. To diagonalize S_b, the D-LDA method first finds the matrix V with constraint

$$V^T S_b V = \Lambda, \tag{6}$$

where $V^T V = I$ and Λ is a diagonal matrix sorted in decreasing order. Let Y denote the first m columns of V, and calculate $D_b = Y^T S_b Y$. Then we calculate $Z = Y D_b^{-1/2}$, and diagonalize $Z^T S_w Z$ by calculate matrix U and diagonal matrix D_w with the constraint

$$U^T (Z^T S_w Z) U = D_w. \tag{7}$$

Finally the D-LDA projection T_{dlda} is defined as

$$T_{dlda} = D_w^{-1/2} U^T Z^T \tag{8}$$

3 Post-processing on Discriminant Vectors

3.1 Why Post-processing

Using the ORL database, we give an intuitional illustration of Fisherfaces and D-LDA's discriminant vectors. The ORL database contains 400 facial images with 10 images per individual. Ten images of one person are shown in Fig.1. The images in the ORL database vary in sampling time, light conditions, facial expressions, facial details (glasses/no glasses), scale and tilt. Moreover, all the images are taken against a dark homogeneous background, with the person in an upright frontal position. The tolerance for some tilting and rotation is up to about 20°. These gray images are 112×92 [13]. In this experiment we choose the first five images of each individual for training and thus obtained a training set consisting of 200 facial images. Then Fisherfaces and D-LDA were used to calculate the discriminant vectors.

Fig. 2(a) shows a set of Fisherfaces' discriminant vectors obtained from the training set and Fig. 2(b) shows five discriminant vectors obtained using D-LDA. It is observed that the discriminant vectors in Fig. 2(a)(b) were not smooth. Since a facial image is a 2D smooth surface, it is reasonable to consider that better recognition performance would be obtained by further improving the smoothness of the discriminant vectors using the post-processing techniques.



Fig. 1. Ten images of one person from the ORL database

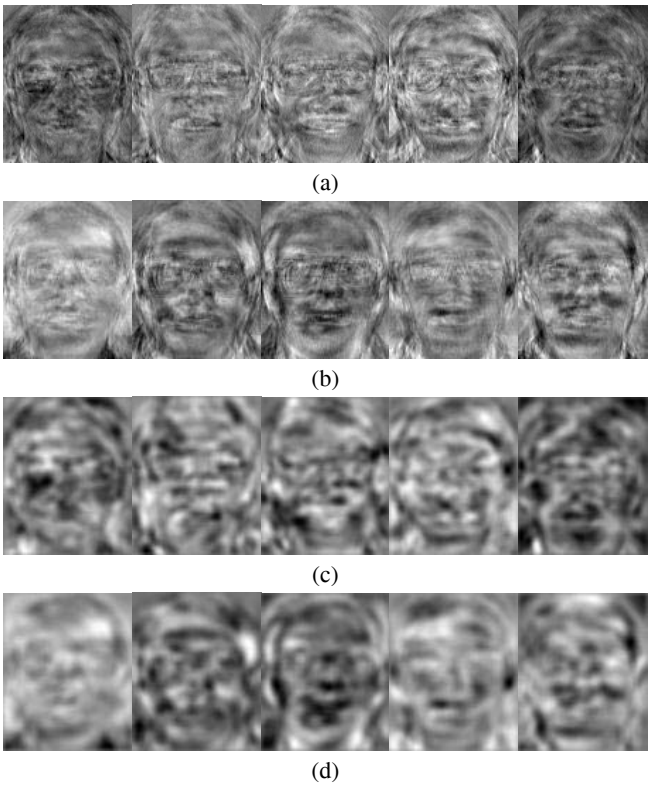


Fig. 2. An illustration of different LDA-based methods' discriminant vectors: (a) Fisherfaces, (b) D-LDA, (c) post-processed Fisherfaces, (d) post-processed D-LDA

3.2 Post-processing Algorithm

When the discriminant vectors obtained using Fisherfaces or D-LDA were reshaped into images, we observed that the discriminant vectors were not smooth. We expect that, this problem can be solved by introducing a post-processing step on discriminant vectors.

We propose to post-process the discriminant vectors using a 2D-Gaussian filter. A Gaussian filter is an ideal filter in the sense that it reduces the magnitude of high spatial frequency in an image and has been widely applied in image smoothing [14]. The 2D-Gaussian filter could be used to blur the discriminant images and remove noise. A two-dimensional Gaussian function is defined as

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2}, \tag{9}$$

where $\sigma>0$, which is the standard deviation. First we define a 2D-Gaussian model M according to the standard deviation σ. Once the standard deviation is determined, the window size $[w, w]$ can be determined as w=4~6×σ, and the Gaussian model M is defined as the w×w truncation from the Gaussian kernel $G(x,y)$. Then we record the 2-norm of each discriminant vector $\|v_i\|=\sqrt{v_i^T v_i}$, and map it into its corresponding image I_i by de-concatenating it into rows of I_i. The Gaussian filter M is used to smooth discriminant image I_i,

$$I_i'(x,y)=I(x,y)\otimes M(x,y). \tag{10}$$

$I_i'(x,y)$ is transformed into a high dimensional vector v_i'. Finally we obtain the post-processed discriminant vector $v_i''=v_i'*\|v_i\|/\|v_i'\|$ and the corresponding LDA projector $T_{pLDA}=[v_1'', v_2'', \cdots, v_m'']$, where m is the number of discriminant vectors.

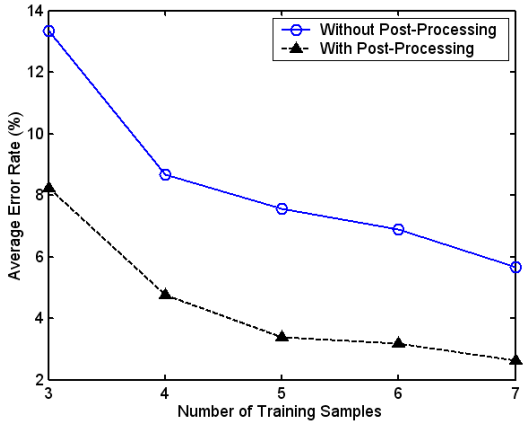
Other image smoothing techniques, such as wavelet and nonlinear diffusion filtering, can also be used to post-process the discriminant vectors. Since the aim of this paper is to investigate the feasibility of post-processing in improving recognition performance, we adopt the simple Gaussian filtering method.

4 Experimental Results and Discussions

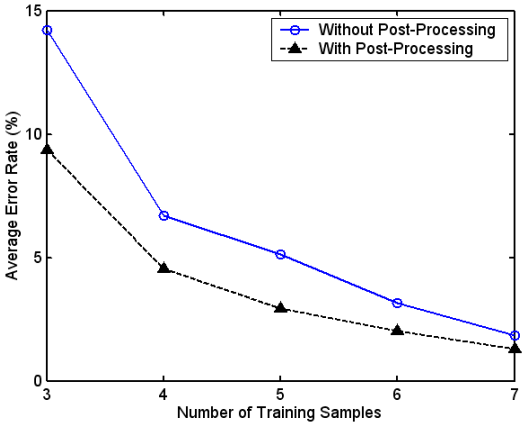
In this section, we use two popular face databases, the ORL, and the FERET database to evaluate the effectiveness of the proposed post-processing approach. Since the aim is to evaluate the feature extraction approaches, a simple nearest neighbor classifier is adopted.

Using the ORL database, we give an intuitional demonstration of the ability of Gaussian filter in smoothing the discriminant vectors. Fig. 2(c) and Fig. 3(d) show the first five post-processed discriminant vectors of Fisherfaces and D-LDA. Compared with Fig. 2(a) and Fig. 2(b), we can observe that the proposed method improved the smoothness of discriminant images.

For the ORL database, we randomly select n samples per person for training, resulting in a training set of 40×n images and a testing set of 40×(10-n) images with no overlapping between the two sets. To reduce the variation of recognition results, the averaged error rate (AER) is adopted by calculating the mean error rate over 20 runs. We set the window of the Gaussian filter as $[w, w]$ = [11, 11] and the variance σ = 2.5. Fig. 3(a) shows the AER obtained using Fisherfaces and post-processed Fisherfaces with different number of training samples n. Fig. 3(b) compares the AER obtained using D-LDA and post-processed D-LDA. It is simple to see that 2D-Gaussian filter can improve the recognition performance.



(a)



(b)

Fig. 3. Comparison of the averaged error rates with and without post-processing for different LDA-based methods: (a) Fisherfaces, (b) D-LDA

Table 1 compares the AER obtained with and without post-processed approaches when the number of training samples is 5. The AER of post-processed Fisherfaces is 3.38, much less than that obtained by classical Fisherfaces. The AER of post-processed D-LDA is 2.93, while the AER obtained by D-LDA is 5.12. We also compared the proposed post-processed LDA with some recently reported results using the ORL database, as listed in Table 2 [9, 12, 15-17]. What is to be noted is that all the error rates are obtained with the number of training samples n=5. It can be observed that post-processed LDA-based method is very effective and competitive in facial feature extraction.

The FERET face image database is the second database used to test the post-processing method [18]. We choose a subset of the FERET database consisting of 1400 images corresponding to 200 individuals (each individual has seven images, including a front image and its variations in facial expression, illumination, ±15° and ±25° pose). The facial portion of each original image was cropped to the size of

80×80 and pre-processed by histogram equalization. In our experiments, we random selected three images of each subject for training and thus resulting in a training set of 600 images and a testing set of 800 images. Fig. 4 illustrates some cropped images of one person.

Table 1. Error rates obtained using the ORL database with and without post-processing

Methods	Fisherfaces	D-LDA
Without Post-Processing	7.55	5.12
With Post-Processing	3.38	2.93

Table 2. Other results recently reported on the ORL database

Methods	Error Rate (%)	Year
DF-LDA [12]	≈4.2	2003
RDA [9]	4.75	2003
NKFDA [15]	4.9	2004
ELDA [16]	4.15	2004
GSLDA [17]	4.02	2004



Fig. 4. Some cropped images of one person in the FERET database

Previous work on the FERET database indicates that the dimensionality of PCA subspace has an important effect on Fisherfaces' recognition [19]. Thus we investigate the recognition performance of Fisherfaces and post-processed Fisherfaces with different number of principal components (PCs). The number of discriminant vectors is set as 20 according to [20]. The averaged recognition rate is used by calculate the mean across 10 tests. The window of the Gaussian filter is set as $[h, w] = [9, 9]$ and the variance is set as $\sigma=1.5$. Fig. 5 shows the averaged recognition rates obtained using Fisherfaces and post-processed Fisherfaces with different numbers of PCs. The highest averaged recognition rate of post-processed Fisherfaces is 87.12%, and that of Fisherfaces is 84.87%. It is observed that post-processing has little improvement on Fisherfaces' recognition rate when the number of PCs is less than 80. When the number of PCs is greater than 100, post-processed Fisherfaces is distinctly superior to Fisherfaces in recognition performance. Besides, the dimensionality of PCA subspace has a much less effect on the performance of post-processed Fisherfaces, whereas the averaged recognition rate of Fisherfaces varied greatly with the number of PCs.

5 Conclusion

Other than the using of pre-processing techniques, this paper shows that post-processing can also be used to improve the performance of the LDA-based methods. In this paper we proposed a 2D-Gaussian filter to post-process discriminant vectors. Experimental results indicate that the post-processing technique can be used to improve LDA's recognition performance. While using the ORL with 5 training samples

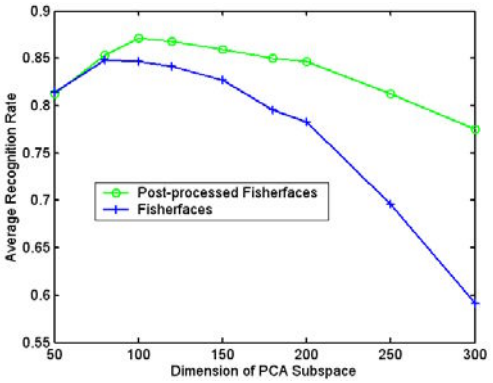


Fig. 5. Comparison of the averaged recognition rates obtained by Fisherfaces and post-processed Fisherfaces with different number of principal components

per individual, the averaged error rate obtained by post-processed Fisherfaces is 3.38, and the averaged error rate of post-processed D-LDA is 2.93. A large set of faces, a subset of FERET database consisting of 1400 images of 200 individuals, is also used to test the post-processing approach, and post-processed Fisherfaces can achieve a recognition rate of 87.12% on this subset.

Some problems worthy of further study still remain. Further work should include the automatic determination of the window and variance of the Gaussian filter, the investigation of other possible post-processing techniques such as wavelets, exploration of the effect of post-processing on other LDA-based methods, and the application of post-processing in other biometrics such as palmprint and gait biometrics.

Acknowledgements

The work is partially supported by the National Science Foundation of China under Grant No. 90209020.

References

1. Zhao, W., Chellappa, R., Phillips, P.J., and Rosenfeld, A.: Face recognition: a literature survey. ACM Computing Surveys, 35 (2003): 399-458.
2. Turk, M., and Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience, 3 (1991): 71-86.
3. Belhumeour, P.N., Hespanha, J.P., and Kriegman, D.J.: Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Analysis and Machine Intelligence, 19 (1997): 711-720.
4. Liu, C. and Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Processing, 11 (2002): 467-476.
5. Yilmaz, A., and Gokmen, M.: Eigenhill vs. eigenface and eigen edge. Pattern recognition, 34 (2001): 181-184.
6. Chien, J.T., and Wu, C.C.: Discriminant waveletfaces and nearest feature classifiers for face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002): 1644-1649.

7. Wu, J., and Zhou, Z.-H.: Face recognition with one training image per person. Pattern Recognition Lett., 23 (2002): 1711-1719.
8. Zhao W, Chellappa R, Phillips P.J.: Subspace Linear Discriminant Analysis for Face Recognition. Tech Report CAR-TR-914, Center for Automation Research, University of Maryland (1999).
9. Dai, D., Yuen, P.C.: Regularized discriminant analysis and its application to face recognition. Pattern Recognition 36 (2003): 845-847.
10. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990).
11. Yu, H. and Yang, J.: A direct LDA algorithm for high-dimensional data with application to face recognition. Pattern Recognition 34 (2001): 2067-2070.
12. Lu, J., Plataniotis, K.N., and Venetsanopoulos, A.N.: Face recognition using LDA-based algorithms. IEEE Trans. Neural Networks 14 (2003) 195-200.
13. The ORL face database. AT&T (Olivetti) Research Laboratories, Cambridge, U.K. Online: Available at http://www.uk.research.att.com/facedatabase.html.
14. Pratt, W. K., Digital Image Processing, 2nd edn. Wiley, New York (1991).
15. Liu, W., Wang, Y., Li, S.Z., and Tan, T.: Null space approach of Fisher discriminant analysis for face recognition.In: Maltoni, D. and Jain, A.K. (eds.): BioAW 2004, Lecture Notes in Computer Science, Vol. 1281. (2004) 32-44.
16. Zheng, W., Zhao, L., and Zou, C.: An efficient algorithm to solve the small sample size problem for LDA. Pattern Recognition 37 (2004): 1077-1079.
17. Zheng, W., Zou, C., Zhao, L.: Real-time face recognition using Gram-Schmidt orthogonalization for LDA. the 17th International Conference on Pattern Recognition (ICPR'04) (2004) 403-406.
18. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1090-1104.
19. Liu, C., and Wechsler, H.: Enhanced Fisher linear discriminant models for face recognition. the 14th International Conference on Pattern Recognition (ICPR'98) (1998) 1368-1372.
20. Yang, J., Yang, J.-Y., Frangi, A.F.: Combined Fisherfaces framework. Image and Vision Computing 21 (2003) 1037-1044.

An Integrated Prediction Model for Biometrics

Rong Wang, Bir Bhanu, and Hui Chen

Center for Research in Intelligent Systems University of California,
Riverside Riverside, California 92521, USA
{rwang,bhanu,hchen}@vislab.ucr.edu

Abstract. This paper addresses the problem of predicting recognition performance on a large population from a small gallery. Unlike the current approaches based on a binomial model that use match and non-match scores, this paper presents a generalized two-dimensional model that integrates a hypergeometric probability distribution model explicitly with a binomial model. The distortion caused by sensor noise, feature uncertainty, feature occlusion and feature clutter in the gallery data is modeled. The prediction model provides performance measures as a function of rank, population size and the number of distorted images. Results are shown on NIST-4 fingerprint database and 3D ear database for various sizes of gallery and the population.

1 Introduction

The goal of pattern recognition is to classify patterns into a number of classes. Patterns can be images, signals or any other type of measurements that need to be classified [1]. Currently, in order to ensure the high confidence in security, biometrics (e.g. fingerprint, palm, face, gait, signature and speech) are used. Depending on application there are two kinds of biometric recognition systems: verification systems and identification systems [2]. A verification system stores users' biometrics in a database. Then it will compare a person's biometrics with the stored features to verify if this person is who she/he claims to be. This is a one-to-one matching problem. The system can accept or reject this person according to the verification result. An identification system is more complex, where for a query the system searches the entire database to find out if there are any biometric features saved in the database that can match the query. It conducts one-to-many matching [2].

Usually a biometric recognition system consists of three stages: image acquisition, feature extraction and matching. Distortion often occurs in these stages which is caused by the sensor noise, feature uncertainty, feature occlusion and feature clutter. In a biometric recognition system before we can widely use the recognition algorithm we need to evaluate its performance on a large population. Since we have very limited data, we can build a statistical model which is based on a small gallery to estimate its performance on large population. Considering the distortion problem that may occur in large population we present an integrated model which considers the distortion to predict the large population performance from a small gallery. Unlike the previous approaches based on a binomial model that use match and non-match score distributions, we present a generalized two-dimensional model that integrates a hypergeometric model explicitly with a binomial model.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 355–364, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Our paper is organized as follows. In section 2 we present the related work. In section 3 we describe the distortion model which includes uncertainty, occlusion and clutter. The detail of the integrated model are given here. Results are shown in section 4. The integrated model is tested on NIST-4 fingerprint database and 3D ear database for various sizes of small gallery and the large population. Conclusions are given in section 5.

2 Related Work

Until now the prediction models are mostly based on the feature space and similarity scores [3]. Tan et al. [4] present a two-point model and a three-point model to estimate the error rate for the point based fingerprint recognition. Their approach not only measures minutiae's position and orientation, but also the relations between different minutiae to find the probability of correspondence between fingerprints. They assume that the uncertainty area of any two minutiae may overlap. Hong et al. [5] present a method to predict the upper and lower bound for object recognition performance. They consider the data distortion problem in the prediction. In their method performance is predicted in two steps: compute the similarity between each pair of model; use the similarity information along with the statistical model to determine an upper and lower bound for recognition performance. Johnson et al. [6] build a *cumulative match characteristic* (CMC) model that is based on the feature space to predict the gait identification performance. Mahalanobis distance and L_2 norm are used to compute similarity within the feature space. They make an assumption about the density that the population variation is much greater than the individual variation. When this assumption is invalid this approach cannot be used.

Wayman [7] and Daugman [8] develop a binomial model that uses the non-match score distribution. This model underestimates recognition performance for large galleries. Phillips et al. [9] create a moment model, which uses both the match score and the non-match score distributions. Since all the similarity scores are sampled independently, their results underestimate the identification performance. Johnson et al. [10] improve the moment model by using a multiple non-match scores set. They average match scores on the whole gallery. For each match score they count the number of non-match scores that is larger than this match score, which leads to an error. In reality the distribution of match score is not always uniform. Grother et al. [11] introduce the joint density function of the match and non-match scores to solve the underestimation problem.

In this paper we present a generalized two-dimensional model that integrates a hypergeometric model with a binomial model to predict the large population performance from a small gallery. It considers the data distortion problem in the large population. The number of distorted images follows hypergeometric distribution. Like Hong et al. [5] our distortion model includes feature uncertainty, occlusion and clutter. The distortion model needs users to define some parameters, such as feature uncertainty *probability density function* (PDF), occlusion amount, clutter amount, clutter region, and clutter PDF etc. Then according to the different numbers of distorted images we get the distributions of match score and non-match score. After this we use the CMC curve [6] to rank all these scores. A CMC curve can show different probabilities of recognizing

biometrics depending upon how similar the features for this query biometrics are in comparison to the other biometrics in the gallery. Finally we use a binomial distribution to compute the probability that the match score is within rank r. In this paper we consider the performance when the rank $r = 1$.

3 Technical Approach

We are given two sets of data: gallery and probe. Gallery is a set of biometric templates saved in the database. For each individual there is one template saved in the gallery. Probe is a set of query biometrics. Large population is the unknown data set whose recognition performance needed to be estimated based on the given gallery and probe set.

3.1 Distortion Model

Our distortion model includes feature uncertainty, occlusion and clutter. Assume $F = \{f_1, f_2, \cdots, f_k\}$ is feature set of the biometrics, where $f_i = (x, y, t)$, x and y represent feature's location, t represents feature's other attributes except for location. Then the distortion algorithm [5] does the following:

a) Uncertainty: Assume the uncertainty PDF follows uniform distribution. It represents how likely each feature is to be perturbed. We replace each feature $f_i = (x, y, t)$ with $f_i^{'}$ which is chosen uniformly at random from the set

$$\{(x', y', t'), (x', y') \in 4NEIGHBOR(x, y), (1 - \alpha)t \leq t' \leq (1 + \alpha)t\}$$

where α is a coefficient, usually $0 \leq \alpha \leq 1$. $4NEIGHBOR(x, y)$ means 4 points around (x,y), they are $\{(x - 1, y), (x + 1, y), (x, y - 1), (x, y + 1)\}$. The unit is pixel.

b) Occlusion: Assume the number of features to be occluded is O. Uniformly choose O features out of the k features, remove these features.

c) Clutter: Add C additional features, where each feature is generated by picking a feature according to the clutter PDF from the *clutter region* (CR). The clutter PDF determines the distribution of clutter over the clutter region. Clutter region is used to determine where clutter features should be added. The clutter region typically depends upon the given model to be distorted. We usually use a bounding box to define the clutter region

$$CR = \{(x, y, t), x_{min} \leq x \leq x_{max}, y_{min} \leq y \leq y_{max}, t_{min} \leq t \leq t_{max}\}$$

where x_{min} and x_{max} represent the minimum and maximum value of x, the same definition for y_{min}, y_{max}, t_{min} and t_{max}.

We define the distortion region of feature f, denoted by $DR(f)$, as the union of all features that could be generated as uncertain version of f. In order to simplify, we use uniform PDF for uncertainty and clutter. In fact other PDFs are also possible and can be implemented.

3.2 Prediction Model

Our two-dimensional prediction model considers the distortion problem which is much more conform with the reality than our previous work [3]. Assume we have two kinds

of different quality biometric images, group #1 and group #2. Group #1 is a set of good quality biometric images without distortion. Group #2 is a set of poor quality biometric images with feature uncertainty, occlusion and clutter. In general, the size of these two groups are N_1 and N_2. We randomly pick p images from group #1 and group #2. Then the number of distorted images y which are chosen from group #2 should follow hypergeometric distribution.

$$f(y) = \frac{\mathrm{C}_{p-y}^{N_1}\mathrm{C}_{y}^{N_2}}{\mathrm{C}_{p}^{N_1+N_2}} \tag{1}$$

where

$$\mathrm{C}_{p-y}^{N_1} = \frac{N_1!}{(p-y)!(N_1-p+y)!}$$
$$\mathrm{C}_{y}^{N_2} = \frac{N_2!}{y!(N_2-y)!}$$
$$\mathrm{C}_{p}^{N_1+N_2} = \frac{(N_1+N_2)!}{p!(N_1+N_2-p)!}$$

where $N_1 + N_2$ is the total number of images in these two groups, $p - y$ is the number of images chosen from group #1.

In order to simplify the description we assume sizes of gallery and probe set are all n. For each image in the probe set we compute the similarity scores with every image in the gallery. Then we have one match score and $n - 1$ non-match scores for this image. Here we assume that the match score and the non-match score are independent. Thus for a given number of distorted images we get a set of match scores $M_i = [m_{i,1}, m_{i,2}, \cdots, m_{i,n}]$ and a set of corresponding non-match scores

$$NM_i = \begin{bmatrix} n_{i,1,1} & \cdots & n_{i,n,1} \\ \vdots & \ddots & \vdots \\ n_{i,1,(n-1)} & \cdots & n_{i,n,(n-1)} \end{bmatrix}$$

where i represents the number of images selected from group #2, $i = 1, 2, \cdots, n$. Now for a given number of distorted images i, jth image has a set of similarity scores which include one match score and $n - 1$ non-match scores

$$S_{ij} = \begin{bmatrix} m_{i,j} & n_{i,j,1} & \cdots & n_{i,j,(n-1)} \end{bmatrix}$$

where $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, n$.

If we have enough match scores and non-match scores then we can estimate their distributions. From above we know that the similarity score distributions depend not only on the similarity scores but also on the number of images with distortion. Here we assume $ms(x|y)$ and $ns(x|y)$ represent the distributions of match scores and non-match scores given the number of distorted images. Assume if the similarity score is higher then the biometrics are more similar. The error occurs when a given match score is smaller than the non-match score. For a given number of distorted images the probability that the non-match score is greater than or equal to the match score x is $NS(x)$ where

$$NS(x) = \int_x^{\infty} ns(t|y)f(y)dt \tag{2}$$

Then the probability that the non-match score is smaller than the match score is $1 - NS(x)$.

Here we assume that the similarity score distributions are similar for small gallery and large population. If the size of large population is N, then for jth image we can have a set of similarity scores, which include one match score and $N - 1$ non-match scores. We rank the similarity scores in decreasing order. Then for a given number of images with distortion the probability that the match score x rank r is given by the binomial probability distribution

$$\mathrm{C}_{r-1}^{N-1}\left(1 - \int_x^{\infty} ns(t|y)f(y)\right)^{N-r}\left(\int_x^{\infty} ns(t|y)f(y)\right)^{r-1} \tag{3}$$

Integrating over all the match scores, for a given number of images with distortion the probability that the match scores rank r can be written as

$$\int_{-\infty}^{\infty} \mathrm{C}_{r-1}^{N-1}\left(1 - \int_x^{\infty} ns(t|y)f(y)\right)^{N-r}\left(\int_x^{\infty} ns(t|y)f(y)\right)^{r-1} ms(x|y)f(y)dx \tag{4}$$

We integrate over all the number of images chosen from group #2, the probability that the match scores rank r can be written as

$$\int_{-\infty}^{\infty} \mathrm{C}_{r-1}^{N-1}\left(1 - \int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{N-r} \left(\int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{r-1} \sum_{y=0}^{n} ms(x|y)f(y)dx \tag{5}$$

In theory the match scores can be any values within $(-\infty, \infty)$. We get the probability that the match scores are within rank r is

$$P(N,r) = \sum_{i=1}^{r} \int_{-\infty}^{\infty} \mathrm{C}_{r-1}^{N-1}\left(1 - \int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{N-r} \left(\int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{r-1} \sum_{y=0}^{n} ms(x|y)f(y)dx \tag{6}$$

Considering the correct match take place above a threshold t, the probability that the match score is within rank r becomes

$$P(N,r,t) = \sum_{i=1}^{r} \int_{t}^{\infty} \mathrm{C}_{r-1}^{N-1}\left(1 - \int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{N-r} \left(\int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{r-1} \sum_{y=0}^{n} ms(x|y)f(y)dx \tag{7}$$

For the problem where rank $r = 1$ then the prediction model with the threshold t becomes

$$P(N,1,t) = \int_t^{\infty} \left(1 - \int_x^{\infty} \sum_{y=0}^{n} ns(t|y)f(y)\right)^{N-1} \sum_{y=0}^{n} ms(x|y)f(y)dx \qquad (8)$$

In this model we make two assumptions: match scores and non-match scores are independent and their distributions are similar for large population. In this model N is the size of large population whose performance needs to be estimated. Small size gallery is used to estimate distributions of $ms(x|y)$ and $ns(x|y)$.

4 Experimental Results

In this section we verify our model on NIST-4 fingerprint database and ear database for different sizes of small gallery and large population. Then we compare the performance of our integrated model with our previous binomial model on the NIST-4 fingerprint database.

4.1 Integrated Prediction Model

Fingerprint Database: All the fingerprints we use in the experiments are from *NIST Special Database 4* (NIST-4). It consists of 2000 pairs of fingerprints, each of them is labeled 'f' or 's' that represent different impressions of a fingerprint followed by an ID number. The images are collected by scanning inked fingerprints from paper. The resolution of the fingerprint image is 500 DPI and the size is 480×512 pixels. Figure 1 is a pair of fingerprints from NIST-4 database.



f0006_09 s0006_09

Fig. 1. Sample images from NIST-4

Usually the minutiae features are used for fingerprint recognition which can be expressed as $f = (x, y, c, d)$, where x and y are the locations of the minutiae, c is the class of minutiae, and d is the direction of minutiae. We define the percentage of minutiae with distortion for one fingerprint as g. In our experiments we choose $g = 5\%$, 7%, and 8% respectively. By applying distortion model to these 2000 pairs of fingerprints according to different distortion percentages, we get 6000 pairs of distorted fingerprints. Assume the number of minutiae is num_j, usually one pair of fingerprints have different number of minutiae so $j = 1, 2, \cdots, 4000$. The distortion model is as following:

(a) Uncertainty: Uniformly choose $U = g \times num_j$ minutiae features out of the num_j replace each $f_i = (x, y, c, d)$ with $f_i^{'}$ chosen uniformly at random from the set

$$\{(x', y', c', d'), (x', y') \in 4NEIGHBOR(x, y), c' = c \pm 1, d' = d \pm 3^{\circ}\}$$

where $i = 1, 2, \cdots, U$.

(b) Occlusion: Uniformly choose $O = g \times num_j$ minutiae features out of the num_j remove these minutiae.

(c) Clutter: Add $C = g \times num_j$ additional minutiae, where each minutiae is generated by picking a feature uniformly at random from the clutter region. Here we choose the clutter region as

$$CR = \{(x, y, c, d), 50 \leq x \leq 450, 60 \leq y \leq 480, c = \{0, 1, 2, 3, 4\}, 10^{\circ} \leq d \leq 350^{\circ}\}$$

In our experiments we use the uniform distribution as the uncertainty PDF and the clutter PDF. The number of features with uncertainty, occlusion and clutter are the same. By adding different percentage of minutiae with distortion g we have four groups of fingerprint images, each group has 2000 pairs of fingerprints. Group #1 is the original fingerprints in NIST-4, group #2 is the fingerprints with $g = 5\%$, group #3 with $g = 7\%$, and group #4 with $g = 8\%$.

Assume our small galley size $n = 50$. We randomly pick up 50 fingerprints pairs from group #1 and group #2. Then the number of fingerprints chosen from group #2 which denoted by y follows hypergeometric distribution,

$$f(y) = \frac{C_y^{50} C_{50-y}^{50}}{C_{50}^{100}} \tag{9}$$

Now we have 50 pairs of images including the original images and the distorted images. The images labeled with 'f' are used as the gallery and the others labeled with 's' are used as the probe set. We use fingerprint verification approach which based on the triplets of minutiae to compute the similarity scores for these fingerprints [12]. Then we get the distributions of the match score and the non-match score. Figure 2 is the distributions of the match score and the non-match score for different number of distorted images. From Figure 2 it's clear that these distributions depend not only on similarity scores also on the number of distorted images. Here we choose the threshold for correct match $t = 12$. For the verification problem we consider the case when rank $r = 1$. This small gallery $n = 50$ applies in the integrated prediction model which can predict the large population performance, here we choose $N = 6000$. Now we get the prediction result for $g = 5\%$. By repeating the above process we get the estimation results for $g = 7\%$ and $g = 8\%$. Average these three prediction values we get the estimation result for large population $N = 6000$. We choose different size of small gallery $n = 70$. By repeating the above process we obtain the estimation results for large population. Figure 3 shows the absolute error between the predicted and experimental verification performance. The absolute error is smaller than 0.08 when the population size is larger than 1000. That means our integrated prediction model can efficiently predict the fingerprint recognition performance for large population.

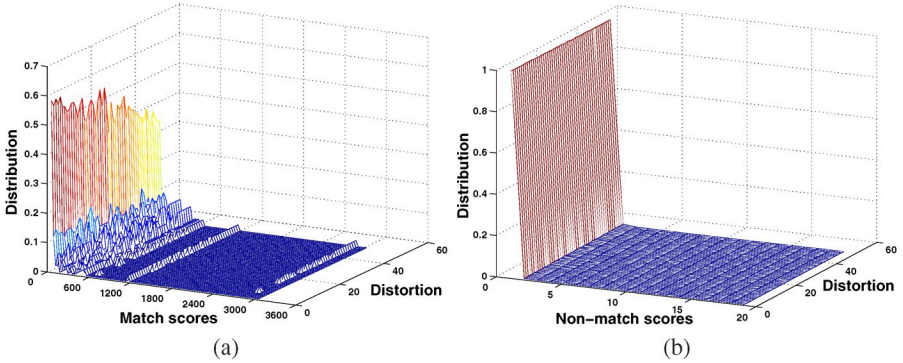


Fig. 2. Similarity scores distributions. (a) Match scores distribution. (b) Non-match scores distribution

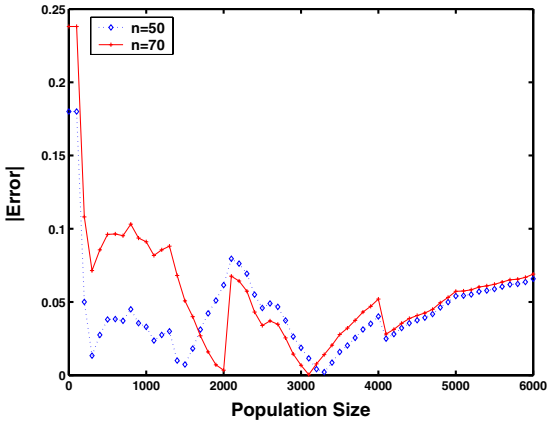


Fig. 3. Absolute error between the integrated prediction model and experimental fingerprint recognition performance

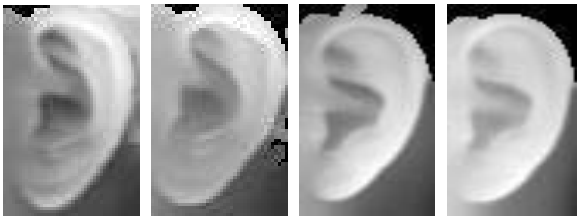


Fig. 4. Sample images from 3D ear database

Ear database: Ear data we use in this experiment are acquired by using Minolta Vivid 300. The image contains 200×200 grid points which has 3D coordinate (x, y, z). This data set has 52 subjects and every subject has two range images which are taken at different viewpoints. Figure 4 shows two pairs of ear. We add Gaussian noise $N(0, \sigma =$

Table 1. Prediction using the integrated model and experimental ear recognition performance

Gallery Size	Experiment	Prediction
50	88.00%	83.28%

Table 2. Predicted ear recognition performance for different sizes of large population by the gallery of 52 objects

Gallery Size	Prediction Results
100	81.67%
150	81.22%
200	81.07%
250	81.01%
300	80.99%

$0.6mm$) to these images. Then we have two image groups: group #1 has 52 images without noise, group #2 has 52 images with Gaussian noise. We randomly choose 52 images from these two image groups as our small gallery we can predict the recognition performance for different large population sizes. Table 1 shows the comparison of the predicted and actual recognition performance with rank $r = 1$. The error between them is 0.0472 which indicate that our integrated prediction model can predict ear recognition performance for large population. Table 2 shows predicted recognition performance for different sizes of large population by the small gallery of 52 objects.

4.2 Comparison with Previous Work

In our previous work [3], we use binomial model to predict the fingerprint recognition performance when rank $r = 1$. In this model the prediction problem is expressed as:

$$P(N, r, t) = \int_t^{\infty} \left(1 - \int_x^{\infty} ns(t)dt\right)^{N-r} ms(x)dx \tag{10}$$

Compared with equation (8) binomial model did not consider the distortion problem in large population. Figure 5 is the prediction error between the integrated model and the binomial model under the same small gallery size for fingerprint database. The prediction error made by integrated model is much smaller than that of the binomial model which indicate that the integrated model is suitable for the distortion problem.

5 Conclusions

We have presented an integrated model which can predict large population performance from a small gallery. This model considers the distortion problem which happens in large population. Results are shown on NIST-4 fingerprint database and 3D ear database for various sizes of small gallery and population. From the above results we can see that compared with previous approaches our model improve the prediction results and can be used to predict the large population performance. In this paper we mainly focused on the biometrics recognition system, in fact this prediction model can be used to predict other kind of object recognition system.

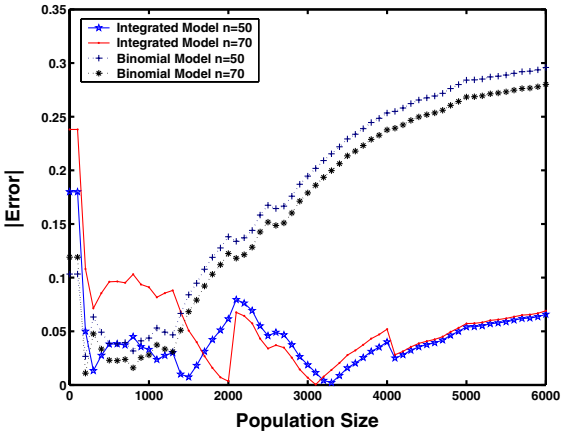


Fig. 5. Prediction error between the integrated model and the binomial model for fingerprint recognition performance

References

1. S. Theodoridis and K. Koutroumbas, "Pattern Recognition," *Academic Press*, Second Edition, 2003.
2. D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, "Handbook of Fingerprint Recognition," *Springer (New York)*, 2003.
3. B. Bhanu, R. Wang, and X. Tan, "Predicting fingerprint recognition performance from a small gallery," *ICPR Workshop on Biometrics: Challenges arising from Theory to Practice*, pp. 47-50, 2004.
4. X. Tan, and B. Bhanu, "On the fundamental performance for fingerprint matching," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol.2, pp. 18-20, 2003.
5. E. S. Hong, B. Bhanu, G. Jones and X. B. Qian, "Performance modeling of vote-based object recognition," *Proceedings Radar Sensor Technology IX*, vol. 5077, pp. 157-166, August 2003.
6. A. Y. Johnson, J. Sun, and A. F. Bobick, "Predicting large population data cumulative match characteristic performance from small population data," *The 4th International Conference on AVBPA*, pp. 821-829, June 2003.
7. J. L. Wayman, "Error-rate equations for the general biometric system," *IEEE Robotics & Automation Magazine*, vol. 6, issue 1, pp. 35-48, 1999.
8. J. Daugman, "The importance of being random: statistical principles of iris recognition," *Pattern Recognition*, 36(2):279-291, February 2003.
9. P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face Recognition Vendor Test 2002," *Evaluation Report*, March 2003.
10. A. Y. Johnson, J. Sun and A. F. Boick, "Using similarity scores from a small gallery to estimate recognition performance for large galleries," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 100-103, 2003.
11. P. Grother, and P. J. Phillips, "Models of large population recognition performance," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2, pp.68-75, 2004.
12. X. Tan, and B. Bhanu, "Robust fingerprint identification," *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 277-280, 2002.

Active Shape Models with Invariant Optimal Features (IOF-ASMs)

Federico Sukno[1,2], Sebastián Ordás[1], Costantine Butakoff[1], Santiago Cruz[2], and Alejandro Frangi[1]

[1] Department of Technology, Pompeu Fabra University, Barcelona, Spain
[2] Aragon Institute of Engineering Research, University of Zaragoza, Spain

Abstract. This paper is framed in the field of statistical face analysis. In particular, the problem of accurate segmentation of prominent features of the face in frontal view images is addressed. Our method constitutes an extension of *Cootes et al.* [6] linear Active Shape Model (ASM) approach, which has already been used in this task [9]. The technique is built upon the development of a non-linear appearance model, incorporating a reduced set of differential invariant features as local image descriptors. These features are invariant to rigid transformations, and a subset of them is chosen by Sequential Feature Selection (SFS) for each landmark and resolution level. The new approach overcomes the unimodality and gaussianity assumptions of classical ASMs regarding the distribution of the intensity values across the training set. Validation of the method is presented against the linear ASM and its predecesor, the Optimal Features ASM (OF-ASM) [14] using the AR and XM2VTS databases as testbed.

1 Introduction

In many automatic systems for face analysis, following the stage of face detection and localization and before face recognition is performed, facial features must be extracted. This process currently occupies a large area within computer vision research.

A human face is part of a smooth 3D object mostly without sharp boundaries. It exhibits an intrinsic variability (due to identity, gender, age, hairstyle and facial expressions) that is difficult if not impossible to characterize analytically. Artifacts such us make-up, jewellery and glasses cause further variation. In addition to all these factors, the observer's viewpoint (i.e. in-plane or in-depth rotation of the face), the imaging system, the illumination sources and other objects present in the scene, affect the overall appearance. All these intrinsic and extrinsic variations make the segmentation task difficult and discourage a search for fixed patterns in facial images. To overcome these limitations, statistical learning from examples is becoming popular in order to characterize, model and segment prominent features of the face.

An Active Shape Model (ASM) is a flexible methodology that has been used for the segmentation of a wide range of objects, including facial features [9]. In

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 365–375, 2005.
© Springer-Verlag Berlin Heidelberg 2005

the seminal approach by Cootes et al. [6] shape statistics are computed from a Point Distribution Model (PDM) and a set of local grey-level profiles (normalized first order derivatives) is used to capture the local intensity variations at each landmark point. In [5] Cootes et al. introduced another powerful approach to deformable template models, namely the Active Appearance Model (AAM). In AAMs a combined PCA of the landmarks and pixel values inside the object is performed. The AAM handles a full model of appearance, which represents both shape and texture variation.

The speed of ASM-based segmentation is mostly based on the simplicity of its texture model. It is constructed with just a few pixels around each landmark whose distribution is assumed to be gaussian and unimodal. This simplicity, however, becomes a weakness when complex textures must be analyzed. In practice, its local grey-levels around the landmarks can vary widely and pixel profiles around an object boundary are not very different from those in other parts of the image. To provide a more complete intensity model, van Ginneken et al. [14] proposed an Optimal Features ASM (OF-ASM), which is non-linear and allows for multi-modal intensities distribution, since it is based on a k-nearest neighbors (kNN) classification of the local textures. The main contribution of that approach is an increased accuracy in the segmentation task that has shown to be particularly useful in segmenting objects with textured boundaries in medical images. However, its application to facial images is not straightforward. Facial images have a more complicated geometry of embedded shapes and present large texture variations when analyzing the same *region* for different individuals. In this work we will discuss those problems and develop modifications to the model in order to make it deal with face complexities, as well as the replacement of the OF-ASM derivatives so that the intensity model is invariant to rigid transformations. The new method, coined Invariant Optimal Features ASM (IOF-ASM) will also attack the segmentation speed problem, mentioned as a drawback in [14]. The performance of our method is compared against both the original ASM and the OF-ASM approaches, using the AR [11] and XM2VTS [12] databases as test bed. Experiments were split into segmentation accuracy and identity verification rates, based on the Lausanne protocol [12].

The paper is organized as follows. In Section 2 we briefly describe the ASM and OF-ASM, while in Section 3 the proposed IOF-ASM is presented. In Section 4 we describe the materials and methods for the evaluation and show the results of our experiments and Section 5 concludes the paper.

2 Background Theory

2.1 Linear ASM

In its original form [6], ASMs are built from the covariance matrices of a Point Distribution Model (PDM) and a local image appearance model around each of those points.

The PDM consists of a set of landmarks placed along the edges or contours of the regions to segment. It is constructed by applying PCA to the aligned set

of shapes, each represented by a set of landmarks [6]. Therefore, the original shapes $\mathbf{u}_i$ and their model representation $\mathbf{b_i}$ are related by the mean shape $\overline{\mathbf{u}}$ and the eigenvectors matrix $\mathbf{\Phi}$:

$$\mathbf{b}_i = \mathbf{\Phi}^T(\mathbf{u}_i - \overline{\mathbf{u}}), \qquad \mathbf{u}_i = \overline{\mathbf{u}} + \mathbf{\Phi}\mathbf{b}_i \tag{1}$$

It is possible to use only the first M eigenvectors with the largest eigenvalues. In that case (1) becomes an approximation, with an error depending on the magnitude of the excluded eigenvalues. Furthermore, under the assumption of gaussianity, each component of the $\mathbf{b_i}$ vectors is constrained to ensure that only *valid shapes* are represented:

$$b_i^m \leq \beta\sqrt{\lambda_m} \qquad 1 \leq i \leq N, \quad 1 \leq m \leq M \tag{2}$$

where β is a constant, usually set between 1 and 3, according to the degree of flexibility desired in the shape model and λ_m are the eigenvalues of the covariance matrix.

The intensity model is constructed by computing second order statistics for the normalized image gradients, sampled on each side of the landmarks, perpendicularly to the shape's contour. The matching procedure is an iterative alternation of landmark displacements based on image information and PDM fitting, performed in a multi-resolution framework. The landmark displacements are provided using the intensity model, by minimizing the Mahalanobis distance between the candidate gradient and the model's mean.

2.2 Optimal Features ASM

As an alternative to the construction of normalized gradients and to the use of the *Mahalanobis* distance as a cost function, van Ginneken et al. [14] proposed to use a non-linear gray-level appearance model and a set of features as local image descriptors. Again, the landmark points are displaced to fit edge locations during optimization, along a profile perpendicular to the object's contour at every landmark. However, the best displacement here will be the one for which everything on one side of the profile is classified as being outside the object, and everything on the other side, as inside of it. Optimal Features ASMs (OF-ASMs) use local features based on image derivatives to determine this. The idea behind that is the fact that a function can be locally approximated by its Taylor series expansion provided that the derivatives at the point of expansion can be computed up to a sufficient order. The set of features is made optimal by sequential feature selection [8] and interpreted by a kNN classifier with weighted voting [3], to hold for non-linearity.

3 Invariant Optimal Features ASM

3.1 Multi-valued Neural Network

In our approach we used a non linear classifier in order to label image points near a boundary or contour. In principle, any classifier can be used, as long as it can

cope with the non-linearity. Between the many available options, our selection was the Multivalued Neural Network (MVNN) [2], mainly based on the need to improve segmentation speed. This is a very fast classifier, since its decision is based only on a vector multiplication in the complex domain. Furthermore, a single neuron is enough to deal with non-linear problems [1], which avoids the need for carefully tuning the number of layers (and neurons in each of them) that characterizes multi-layer perceptron networks.

The MVNN will have as many inputs as the number of features selected for each landmark (say N_F), all of them being integer numbers, and a single integer output. The classification is performed by a single neuron, which for every input x_k finds a corresponding complex number on the unit circle:

$$q_k = \exp(j2\pi x_k) \quad 1 \leq k \leq N_F \tag{3}$$

x_k being the k-th input variable value (discrete), q_k its corresponding complex-plane representation, N_F the number of inputs (features) and j the imaginary unit $\sqrt{-1}$. Then, the neuron maps the complex inputs to the output plane by means of the N_F-variable function f_{N_F}:

$$f_{N_F}(z) = \exp(j2\pi \frac{k}{N_O}) \quad when \quad 2\pi \frac{k}{N_O} \leq \arg(z) < 2\pi \frac{k+1}{N_O} \tag{4}$$

$$z = w_0 + w_1 q_1 + w_2 q_2 + ... + w_{N_F} q_{N_F} \tag{5}$$

where w_k are the network weights learnt during the training phase. The f_{N_F}'s image is a complex plane, which has been divided into N_O angular sectors, like a quantization of $\arg(z)$. In other words, the neuron's output is defined as the number of the sector in which the weighted sum z has fallen. Notice that the number of sectors of the input and output domains does not need to be the same.

3.2 Irreducible Cartesian Differential Invariants

A limitation of using the derivatives in a cartesian frame as features in the OF-ASM approach is the lack of invariance with respect to translation and rotation (rigid transformations). Consequently, these operators can only cope with textured boundaries with the same orientations as those seen in the training set. To overcome this issue we introduce a multi-scale feature vector that is invariant under 2D rigid transformations.

It is known [7][15] that Cartesian differential invariants describe the differential structure of an image independently of the chosen cartesian coordinate system. The term irreducible is used to indicate that any other algebraic invariants can be reduced to a combination of elements in this minimal set. Table 1 shows the Cartesian invariants up to second order.

To make our approach invariant to rigid transformations we use these invariants at three different scales, $\sigma = 1$, 2 and 4. The zero order invariants were not used since the differential images are expected to provide more accurate and

Table 1. Tensor and Cartesian formulation of invariants

Tensor Formulation	**2D Cartesian Formulation**
L	L
L_{ii}	$L_{xx} + L_{yy}$
$L_i L_i$	$L_x^2 + L_y^2$
$L_i L_{ij} L_j$	$L_x^2 L_{xx}^2 + 2L_{xy} L_x L_y + L_y^2 L_{yy}^2$
$L_{ij} L_{ji}$	$L_{xx}^2 + L_{xy}^2 + L_{yy}^2$

stable information about facial contours (edges). For each landmark and resolution level, a sequential feature selection algorithm [8] was used to reduce the size of the feature vector. In this way, only a subset of the invariants drove the segmentation process.

3.3 Texture Classifiers for IOF-ASM

IOF-ASM is basically an improved OF-ASM. The first two improvements are the new and much faster classifier and the use of features invariant to rigid transformations of the input image. Only one improvement is left that we will be stated below.

Let us look back for a moment at OF-ASM. Its training is based on a landmarked set of images for which all of the derivative images are computed and described by local histograms statistics. The idea behind this method is that, once trained, texture classifiers will be able to label a point as inside or outside the region of interest based on the texture descriptors (the features) or, ideally, on a smaller (*optimal*) subset of them. Therefore, labelling inside pixels with 1 and outside pixels with 0 and plotting the labels corresponding to the profile pixels, the classical step function is obtained, and the transition will correspond to the landmark position.

Nevertheless, there are a couple of reasons why this will not happen. The first one is that certain regions of the object are thinner than the size of the grid, and then the correct labelling of the points will look more like a bar than like a step function. An indicative example arises when the square grid is placed on the eyes or eyebrows contours, especially if using a multiresolution framework, as ASM does (Fig. 1). Another problem is that the classifiers will not make a perfect decision, so the labelling will look much noisier than the ideal step or bar. Moreover, Fig. 1 illustrates how, for certain landmarks where there is a high contour curvature (i.e mouth/eyes/eyebrows corners), most of the grid points lie outside the contour, promoting quite an unbalanced training of the classifiers.

To tackle these problems, in our IOF-ASM we introduced new definitions of input and output of the classifiers. For each landmark, instead of the Gaussian weighted histograms used in OF-ASM, we place a square grid, with a subgrid at each cell of the main grid. In other words, in previous approaches fixed positions along the normal were used to sample pixels. We extended this approach and defined a grid with its center on the landmark. Then for each cell of this grid we use a classifier, whose inputs are taken from a subgrid centered at each of the cells of the main grid.

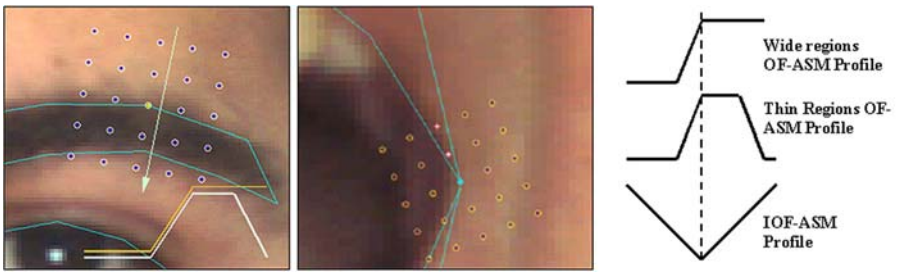


Fig. 1. A typical eyebrow image and a 5x5 grid with the arrow indicating the normal to the contour (Left); The same grid in the mouth corner, where only 3 points lie inside the lip (Center); and the typical graphs of the profiles for OF- and IOF-ASM (Right)

Regarding the outputs, the bi-valued labelling (inside-outside) is replaced with the distance of the pixels to the landmarked contour. Then, for each cell of the main grid the classifiers are trained to return the distance to reach the landmark. Since those centers are placed along normals to the contour, the typical plot of the labels will take a shape of letter "V", with its minimum (the vertex) located at the landmark position, irrespective of which region is sampled or its width relative to the grid size.

At matching time, this labelling strategy allows for introducing a confidence metric. The best position for the landmark is now the one which minimizes the profile distance to the ideal "V" profile, excluding the *outliers*. An outlier here is a point on the profile whose distance to the ideal one is greater than one. It can be easily understood by noticing that such a point is suggesting a different position to place the landmark (i.e. its distance would be smaller if the V is adequately displaced). If the number of outliers exceeds 1/3 of the profile size, then the image model is not trustworthy and the distance for that position is set to infinity. Otherwise, preference is given to the profiles with fewer outliers. The function to minimize is:

$$f(k) = N_{OL} + \frac{1}{N_P - N_{OL}} \sum_{i=1}^{N_P - N_{OL}} |p_i - v_i| \tag{6}$$

where k are the candidate positions for the landmark, N_{OL} is the number of outliers, N_P the profile size, and $\mathbf{p}$ and $\mathbf{v}$ are the input and ideal (V) profiles, respectively.

4 Experiments

The performance of the proposed method was compared with the ASM and OF-ASM schemes. Two datasets were used. The first one is a subset of 532 images from the AR database [11], showing four different facial expressions of 133 individuals. This database was manually landmarked with a 98-point PDM template that outlines the eyes, eyebrows, nose, mouth and face silhouette. The second dataset is the XM2VTS [12] database, composed of 2360 images (8 for each of 295 individuals).

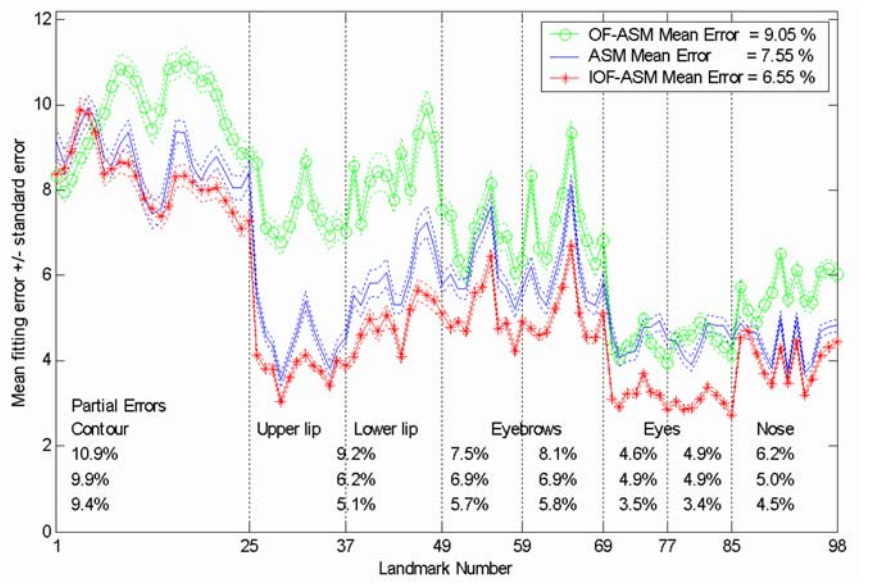


Fig. 2. Mean fitting error performance in 532 frontal images of the AR database

Both segmentation accuracy and identity verification scores have been tested. In order to make verification scores comparable between the two datasets, the AR images were divided into five groups, preserving the proportions of the Lausanne Protocol [12] (configuration 2) for the XM2VTS database. In this way, we came out with 90 users (two images/each for training, one for evaluation and one for testing), 11 evaluation impostors (44 images) and 32 test impostors (128 images). It must be pointed out that the individuals on each group were randomly chosen, making sure that there is the same proportion of men/women in all of them. Moreover, they are also balanced in the amount of images per facial expression.

4.1 Segmentation Accuracy

The segmentation accuracy was tested on the AR dataset only, since this task needs the input images to be annotated. For the same reason, all models constructed in our experiments were based on the *Training Users* group of the AR dataset. Table 2 summarizes the parameters used by the 3 compared models. Additionally, we use 150 PCA modes for the PDM, $\beta = 1.5$ (see (2)) and a search range of ± 3 pixels along the profiles at each resolution. The segmentation results are displayed in Fig. 2. The displayed curves show the mean Euclidean distance ($\pm$ the standard error) between the segmentation using the corresponding model and the manual annotations for each landmark, normalized as a percentage of the distance between the true centers of the eyes. The mean eyes-distance in the AR dataset is slightly greater than 110 pixels, so the curves give the fitting error approximately in pixel units.

Table 2. Parameters used to build the statistical models

Parameter	ASM	OF-ASM	IOF-ASM
Profile length	8	n/a	7
Grid size	n/a	5×5	7×1
Each grid point	n/a	$\alpha = 2\sigma$	7×5 patch
Resolutions	4	5	5
Selected features	n/a	6 of 36	70 of 420

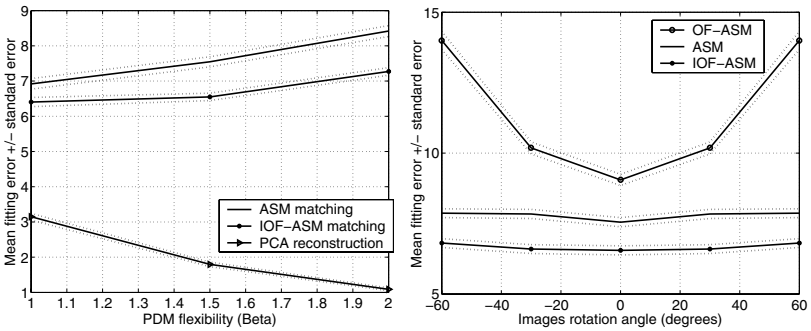


Fig. 3. Segmentation errors vs. PDM flexibility (Left) and vs. rotation angles of the input images (Right)

It is clear from Fig. 2 that OF-ASM produces a segmentation error significantly larger than the other methods, due to the problems that were previously discussed, mainly regarding shape complexity. On the other hand, IOF-ASM outperforms ASM in all regions, and the difference is statistically significant in several landmarks. The average improvement of IOF-ASM with respect to the ASM segmentation is of 13.2%, with a maximum and minimum of 28.5% and 5.2% in the eyes and silhouette contour respectively.

Fig. 3 (Left) shows further comparison of ASM and IOF-ASM accuracy when varying the PDM flexibility parameter β (see (2)). It can be seen that as β increases, the difference between the error of both models tends to grow. At the same time, the PCA reconstruction error introduced by the PDM decreases, so the segmentation relies more on the image model precision. This behavior enforces the hypothesis of performance improvement in favor of IOF-ASM.

The three models are always initialized with their *mean shape* centered at the true face centroid (according to the annotations) and scaled to 80% of the mean size of the faces in the database. Notice in Table 2 that the image model search range for all models is ± 3 pixels per resolution level, giving a total of $\pm 3 \times 2^{N_R - 1}$ pixels, for N_R resolutions. Considering such initialization, the initial distance between the model landmarks and the annotated positions will be up to 68 pixels in the lower lip and the chin, and up to 40 pixels in the rest of the face, so N_R should be fixed at least to 5. However, in our experiments the best performance for ASM was obtained with 4 resolutions, and therefore we used this value.

Table 3. Identity Verification Scores

Database	Set	Parameter	ASM	IOF-ASM
AR	Evaluation	EER	3.3%	3.3%
	Test	FAR	3.6%	3.9%
		FRR	3.3%	<1%
XM2VTS	Evaluation	EER	11.0%	6.8%
	Test	FAR	10.9%	6.9%
		FRR	12.8%	7.3%

4.2 Rotation Invariance

It was emphasized in Section 3.2 that the IOF-ASM features extracted from the images are invariant to rigid transformations. ASM exhibits the same invariance, but OF-ASM does not. To verify this fact we repeated the experiments of the previous section but using rotated versions of the images, ranging from -60 to +60 degrees. The PDM was constructed from the rotated images, such that the starting shape (based on the *mean shape*) was also rotated. But the image models were not changed (i.e. they were based on the non-rotated images) so that their invariance is the only thing to test.

The results of the experiment are presented in Fig. 3 (on the right). As expected, there is a clear increment of the segmentation error in the OF-ASM as the rotation angle departs from zero. On the other hand, the ASM and IOF-ASM performance is only marginally affected.

4.3 Identity Verification

Once demonstrated that IOF-ASM is more accurate in segmenting facial images, there is the question of weather or not it will improve recognition tasks as well. The development of a state-of-the-art classifier is beyond the scope of this paper. Our approach consisted of a whitened angle classifier, known to be a good choice for PCA-based metrics [13]. In order to obtain the inputs for the classifier the final shape matched by the model is used to warp image pixels into some *mean shape*. Then, texture parameters are computed from it using PCA. In our experiments, the warping was done by means of a Delaunay triangulation.

The error rates obtained with this strategy are presented in Table 3. The evaluation sets Equal Error Rate (EER) [4] were used to fix the working point of the classifier and get the False Acceptance (FAR) and False Rejection (FRR) rates from the test sets. Despite the fact that AR database error rates of ASM and IOF-ASM are comparable, the DET curves [10] in Fig. 4 show that there is some advantage for IOF-ASM, especially in the test set. These curves are not needed for the XM2VTS database, since the error rates differ significantly.

5 Summary and Conclusions

In this paper a new segmentation method has been presented to solve some limitations of its predecessor, the OF-ASM approach. The main contributions

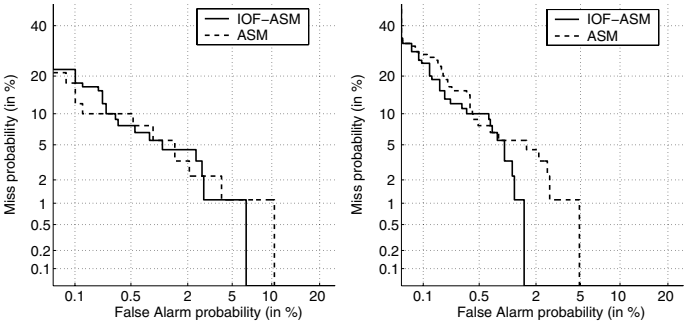


Fig. 4. DET curves for the AR Evaluation (Left) and Test (Right) Sets

introduced here are the rigid transformations invariance, the ability to deal with shape complexities (such as multiple embedding) and the speed up of the segmentation process (up to 5 times with the AR database training set), by means of faster texture classifiers.

Experiments were presented showing that the non-linear intensity model outperformed the linear one, and yielded smaller segmentation error, especially when matching the eyes, eyebrows and some points of the lips, where the pixel value distributions are expected to be clearly non-unimodal. The invariance under 2D rotations was also successfully tested on a wide angles range.

The influence of the accuracy improvement respect to the ASM was reported on identity verification. The IOF-ASM demonstrated superior performance both in the AR database, partially used to construct the model, and the XM2VTS database, whose images were not involved in the model construction.

Acknowledgments

This work was partially funded by grants TIC2002- 04495-C02 and FIT-390000-2004-30 from the Spanish Ministry of Science and Technology. FS is supported by a BSCH grant; SO is supported by an FPU grant from the Spanish Ministry of Education. AF holds a Ramón y Cajal Research Fellowship.

References

1. I. Aizenberg, C. Butakoff, V. Karnaukhov, N. Merzlyakov, and O. Milukova. Blurred image restoration using the type of blur and blur parameters identification on the neural network. In *SPIE Proceedings on Image Processing: Algorithms and Systems*, volume 4667, pages 460–471, California, USA, 2002.
2. I.N. Aizenberg, N.N. Aizenberg, and J. Vandewalle. *Multi-valued and universal binary neurons: theory, learning, applications.* Kluwer Academic Publishers, Boston/Dordrecht/London, 2000.
3. S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, and A.Y. Wu. An optimal algorithm for approximate nearest neighbor serching in fixed dimensions. *International Jorunal of Computer Vision*, 45(6):891–923, 1998.

4. R.M. Bolle, N.K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems-the subsets bootstrap. *Computer Vision and Image Understanding*, 93:1–33, 2004.
5. T.F. Cootes, G. Edwards, and C.J. Taylor. Active appearance models. In *Proceedings of European Conference on Computer Vision*, volume 2, pages 484–498, Springer, 1998.
6. T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.
7. L. Florack. *The Syntactical Structure of Scalar Images*. PhD thesis, Utrecht University, Utrecht, The Nedherlands, 2001.
8. M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classiers. *Pattern Recognition*, 33:25–41, 2000.
9. A. Lanitis, C.J. Taylor, and T.F. Cootes. Automatic interporetation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and machine intelligence*, 19(7):743–756, 1997.
10. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *Proceedings of of Eurospeech (EUROSPEECH'97)*, pages 1895–1898, 1997.
11. A. Martínez and R. Benavente. The AR face database. technical report. Computer Vision Center, Barcelona, Spain, 1998.
12. K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Proceedings of International Conference on Audio- and Video-Based Person Authentication*, pages 72–77, 1999.
13. V. Perlibakas. Distance measures for PCA-based face recognition. *Pattern Recognition Letters*, 25(6):711–724, 2004.
14. B. van Ginneken, A.F. Frangi, J.J. Staal amd B.M. ter Har Romeny, and M.A. Viergever. Active shape model segmentation with optimal features. *IEEE Transactions on Medical Imaging*, 21(8):924–933, 2002.
15. K.N. Walker, T.F. Cootes, and C. J. Taylor. Correspondence using distinct points based on image invariants. In *British Machine Vision Conference*, volume 1, pages 540–549, 1997.

Palmprint Authentication Using Time Series

Jian-Sheng Chen, Yiu-Sang Moon, and Hoi-Wo Yeung

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{jschen,ysmoon,hwyeung}@cse.cuhk.edu.hk

Abstract. Automated personal authentication using biometric features is getting more and more popular for solving the security problems. A new branch of biometric technology, palmprint authentication, has attracted increasing amount of attention because palmprints are abundant of line features and thus low resolution images can be used. In this paper, we propose a new approach for palmprint feature extraction, template representation and matching. Using of time series technologies such as SAX representation and MINDIST calculation is the key to make this new approach simple, flexible and reliable. Experiment shows that this approach can achieve an accuracy of 98.7% when performing one to one verification on a 600 palmprints database. This new approach, which is very computationally efficient, also facilitates the biometric feature fusion as well as palmprint identification using incomplete templates.

1 Introduction

Automated personal authentication using biometric features has been widely studied during the last two decades. Previous research efforts have made it possible to apply biometric systems to practical applications for commercial or security purposes. Among all the existing biometric technologies, fingerprint is the most successful one. However it is still quite difficulty to detect minutiae from dry/wet fingers or fingers from the elders. The high cost of fingerprint sensors has also hampered the social acceptance of this technology. Recently, a novel biometric feature, palmprint [1], has attracted an increasing amount of attention because it has several advantages: palmprints are abundant of line features; low-resolution imaging can be employed; faking a palmprint is quite difficult because the texture is very complicated and one seldom leaves his/her whole palmprint somewhere unintentionally.

Many approaches have been proposed for palmprint authentication [1–7]. These approaches have mainly focused on selecting appropriate features for describing the palmprints. The representation of the features has never been carefully studied. However, in a practical palmprint authentication system, the database might contain templates from millions of people. Choosing a suitable feature representation to make the database small in size and fast responsive to query while keeping high verification accuracy is vital. Also, as multi-biometrics has been proved to be an effective way for improving the reliability of the biometric systems, the convenience and flexibility of using palmprint features as a fusion partner with other different biometric features also needs to be considered.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 376–385, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we will propose a new approach for palmprint feature extraction and representation. This new approach makes use of "time series", a concept which has been widely studied in data mining. The rest of this paper is organized as the follows. Section 2 is a review of the previous palmprint authentication research and a brief introduction to the time series technologies to be used in the paper. Section 3 is the detail description of our approach. Experiments and results are elaborated in section 4. The last section is a conclusion of our work and a discussion on the further work.

2 Related Work

In previous studies, the palmprint features used can be classified into three categories: structural features, statistical features and algebraic features. Typical structural features include principal lines, minutia points and delta points [1, 6]. Statistical features such as texture energy were used in [3, 4, 5], algebraic features such as eigenpalms and fisherpalms were proposed in [2, 7]. Our new approach actually allows adoption of different kinds of features. However, simple statistical features such as gray-level average and variance will be presented in this paper as the key contribution of our work is to use "time series" for the representation of the features.

Time series has been widely studied in data mining [8, 9], bioinformatics [10, 11] and branches of pattern recognition such as biometrics [12, 13]. Technologies such as classification, indexing and motif detection for time series are quite mature. Basically, a time series is a collection of observations made sequentially in time. Nevertheless, as long as the data of interest can be represented sequentially, time series technologies can be applied. In this paper the palmprint features are extracted as data sequences. Time series technologies are applied for the template representation and matching.

One important development in time series research is the introduction of Symbolic Aggregate approXimation (SAX) [14]. SAX is a simple and effective tool for solving most time series problems. The SAX representation is obtained by first transforming the time series into Piecewise Aggregate Approximation (PAA) representation; then predetermined breakpoints are used to map the PAA coefficients into SAX symbols. Basically, SAX converts a real valued data sequence into a string of symbols. MINDIST (minimum distance) is employed to measure the similarity of two SAX symbol strings [14]. In our work, SAX is used for the palmprint template representation and MINDIST is adopted as the matching score of two templates.

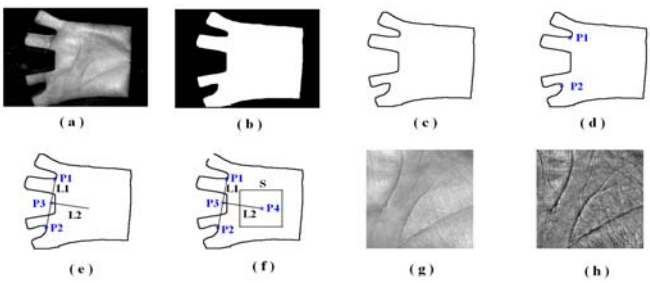


Fig. 1. The palmprint preprocessing

3 Palmprint Representation and Verification

3.1 Palmprint Image Preprocessing

The public palmprint database from the Biometric Research Center, The Hong Kong Polytechnic University, is used in this paper. The palmprints (Fig. 1a) in the database are preprocessed before the feature extraction:

1) A threshold is applied to convert the original images into binary images. Isolated pixels are removed. A binarized palmprint is shown in Fig. 1b;
2) The palm border (Fig. 1c) is obtained using a border tracking algorithm [21];
3) The curvature maxima $P_1$and P_2 between the fingers are located using a curvature maxima finding algorithm [22] (Fig. 1d);
4) P_1 and P_2 are linked to get line L_1 Use another line L_2 to pass through the middle point P_3 of L_1 perpendicularly (Fig. 1e);
5) A point P_4 is found in L_2 so that the length of P_3 P_4 equals to a predefined value. A square S (Fig. 1f) of fixed size (135x135 pixels) is extracted with P_4 as its center (Fig. 1f). S is the region of interest (ROI), from which the features will be extracted;
6) The intensity of S is smoothed so that the illumination becomes uniform (Fig. 1h).

Several ROI samples are shown in Fig. 2.

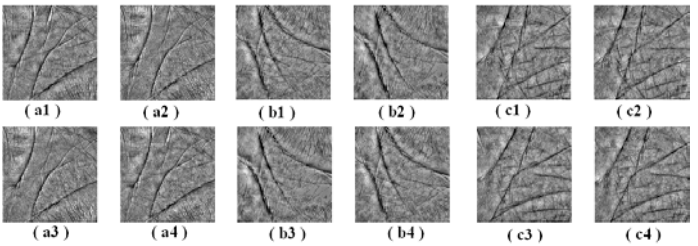


Fig. 2. Extracted ROI squares, (a1) ~ (a4); (b1) ~ (b4); (c1) ~ (c4) are from the same palm

3.2 Feature Extraction and Representation

For time series representation purpose, ROI squares need to be decomposed into sequential data. There are many possible ways of decomposing a 2D image into sequential data. In the MPEG technology, the DCT output of a frame is transformed into sequential data by applying zigzag scan [20]. However, this method is not suitable here. As shown in Fig. 1(f), the ROI square is not rotated according to the direction of L_1 and L_2. This is because of the low resolution (135x135) of the ROI squares. Distortion and blur caused by the rotation become significant no matter what kind of interpolation method is used. Since the line L_2 has different orientations in different palmprint images, the ROI squares extracted from different palmprint images of the same palm are usually different in direction as shown in Fig. 3.

To solve this problem, we adopt a spiral as the track for the decomposition. The polar equation of a spiral is $r = a\theta^n$. In our implementation, we simply use the Archimedes' spiral with the polar equation $r = a\theta$. a is set to 0.8 empirically. To counteract the effect of direction variation of L_2, assumed to be θ_0, we simply include θ_0 in

the polar equation of the spiral as $r = a(\theta + \theta_0)$. Thus, we "rotate" the spiral to adapt the direction of L_2. It is easy to observe that no matter what θ_0 is, the spiral is invariant in direction with respect to the palm.

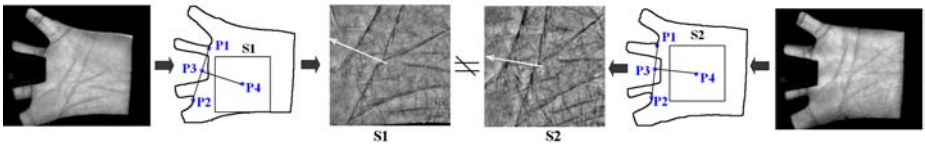


Fig. 3. ROI square direction problem

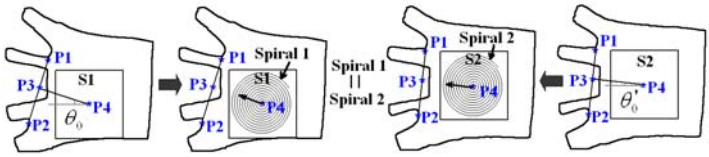
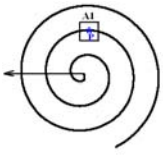


Fig. 4. Use spiral as the feature sequence extraction track



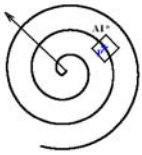
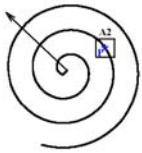
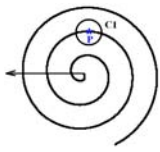


Fig. 5. The problem of using rectangular local area





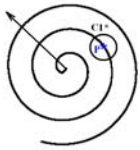


Fig. 6. Use circular local area

Features are extracted along the spiral. For each point P in the spiral, certain local information at P is used as the feature. Many kinds of local textural features can be used here, such as average intensity, variance, and cross correlation. The shape of the local area has to be carefully selected as the implicit rotation of the spiral must be taken into consideration. Choosing a small rectangle at each tracking point as the local area does not work because rectangle is not rotationally invariant. In Fig. 5, after rotation, A1 changes to A1* instead of A2. To address this problem, we use a circle C as the local area with P as its center so that after rotation C remains invariant (Fig. 6.). The verification accuracy differs when different local texture features are adopted. In this paper, we use average intensity and variance.

For each point on the spiral, the local textural feature is extracted to form the data sequence Q which usually has a quite high dimension (around 2000). Dimension deduction is achieved by converting Q into its SAX representation Q_{sax} [14]. The length and level number of Q_{sax} can be used to control the degree of the dimension deduction [14]. Fig. 7 shows one example of the extracted data sequence and the corresponding SAX representations with different lengths and levels.

The SAX representation Q_{sax} is used as the template of the palmprint. The matching score is obtained by calculating the MINDIST of the two templates. Fig. 8 shows the matching of two templates extracted from two different palmprint images taken from the same palm. The smaller the gray areas are, the smaller the MINDIST is, and the more similar the two palmprints are.

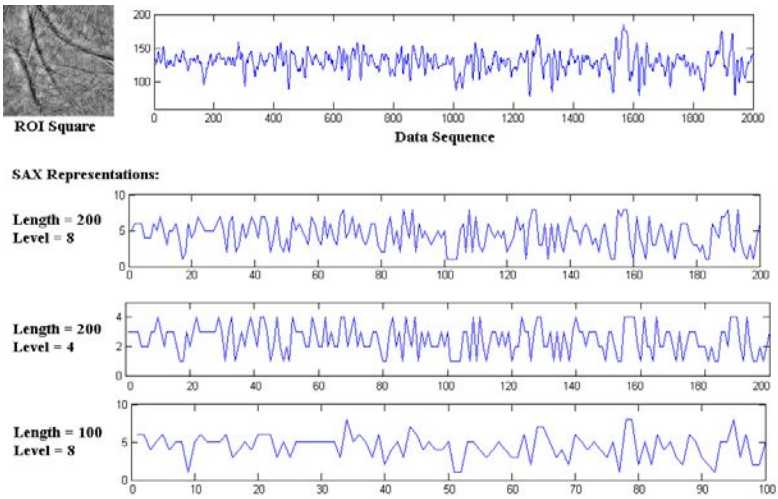


Fig. 7. Data sequence and the corresponding SAX representations with different parameters

Suppose the spiral length is 2000, the radius of C is 5 pixels and intensity average is used as the local feature. The feature extraction process needs around 240,000 additions and 2000 divisions. The computations needed for the SAX conversion and the template matching (MINDIST calculation) are trivial [14]. Compared to the existing methods, the computational complexity of our new approach is considerably low. In an ordinary desktop PC, the time needed for SAX extraction and matching for one palmprint is far less than one second. Thus, our approach is also suitable to be implemented in the relatively slow mobile embedded systems such as PDA.

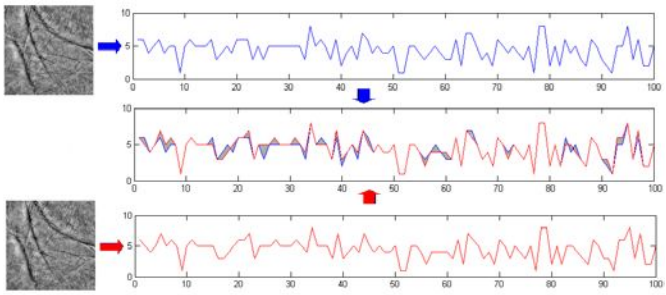


Fig. 8. Palmprint matching using the MINDIST of the SAX representation

4 Experiments and Results

The experiments described in this section are all performed on the PolyU Palmprint Palmprint Database [15]. The database contains 600 grayscale palmprint images (384x284 pixels; 96dpi) from 100 different palms. Six images from each palm were collected in two sessions, 3 images for each session. The average interval between the two sessions was two months. Three kinds of experiments are performed:

1) Accuracy test. The templates for all the palmprints are matched with each other. The verification accuracy is estimated using different features (average intensity & variance), different SAX lengths and different SAX level numbers.
2) Fusion test. The templates extracted using different parameters are fused together to investigate the effectiveness of biometric feature fusion.
3) Incomplete template identification. Only part of a template (substring of the SAX representation) is used to identify the palmprint from the database.

4.1 Accuracy Test

The templates of all 600 palmprints in the database are extracted and matched (1500 genuine matches, 178200 imposter matches) with each other. Different SAX lengths and level numbers are used and local average intensity and local variance are adopted respectively. The verification accuracy results are listed in Table 1 and Table 2. The ROC curves of several typical test cases are shown in Fig. 9.

Table 1. The verification accuracy using local average intensity

SAX Length	SAX Level	EER(%)	SAX Length	SAX Level	EER(%)
200	10	1.33	200	5	1.51
200	9	1.34	200	4	1.60
200	8	1.51	200	3	1.79
200	7	1.56	100	8	2.01
200	6	1.59	50	8	3.63

Table 2. The verification accuracy using local variance

SAX Length	SAX Level	EER(%)	SAX Length	SAX Level	EER(%)
200	10	3.55	200	5	3.88
200	9	3.56	200	4	3.88
200	8	3.43	200	3	3.30
200	7	3.48	100	8	4.75
200	6	3.68	50	8	6.40

From Table 1 and Table 2 we can see that our new approach has achieved a very high verification accuracy of 98.7%. Also we can see that when the number of the SAX level changes, the verification accuracy does not change abruptly, but the change of SAX length does affect the verification accuracy dramatically. To build a practical biometric system, there must be a tradeoff between the template size and the system accuracy. The experimental results above have shown that longer SAX length is more preferable than bigger SAX level number for a practical biometric system.

SAX representations are small in terms of storage requirement. When SAX length equals to 200 and the level number is 4, each template occupies only 400 bits or 50 bytes. This is one of the advantages of our approach.

4.2 Fusion Test

Multi-biometrics has been proved to be effective for improving the reliability of biometric systems [16, 17]. By using the SAX representation (actually symbol string),

feature fusion (actually string concatenation) can be implemented with ease. The weight of each feature can be easily adjusted by choosing different SAX lengths. In our experiment, we concatenate the SAX representations using different feature extraction schemes (average intensity & variance) to implement a very simple feature fusion. The experiment results are listed in Table 3 in which we can see that the verification accuracy after the fusion is better than using a single feature only. Also, we can see that although test cases II and III have the same SAX length (150) after fusion, the verification accuracy differs a lot. This is because the weight of feature 1 (average intensity) in test case III is higher than that in test case II, and feature 1 (average intensity) is generally more reliable than feature 2 (variance) regarding to the verification accuracy (Table1 and Table2).

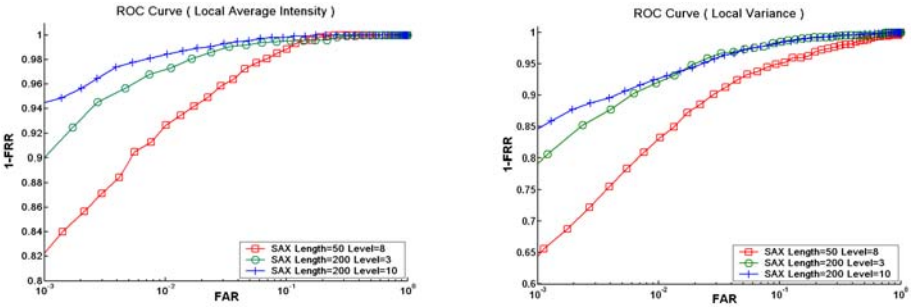


Fig. 9. ROC curves of typical test cases

Table 3. The verification accuracy using feature fusion

Feature 1: Average Intensity			Feature 2: Variance			Feature-Fusion EER (%)	
SAX Length	SAX Level	EER (%)	SAX Length	SAX Level	EER (%)		
100	8	2.01	100	8	4.75	1.99	I
50	8	3.63	100	8	4.75	3.46	II
100	8	2.01	50	8	6.40	1.85	III
50	8	3.63	50	8	6.40	3.08	IV

4.3 Incomplete Template Identification

The purpose of this experiment is to test whether it is possible to identify the input palmprint by using only part of its SAX representation. As the palmprint images are always taken using low resolution input devices such as webcam, and the distortion of the palm is often more obvious than the fingerprints, so it is quite difficulty to control the overall quality of the input palmprint images. However, if we can discard the areas of low quality and use the incomplete template to do the authentication, the palmprint authentication technology will become more practical.

We perform the following experiment: Among the 100 palms, we choose 50 to form a database D. For each palm P_x in D, the master template T_x is obtained by calculating the average value of the SAX representations of 2 (randomly chosen) out of the 6 palmprints of P_x. Then remaining 500 palmprints (palmprints used for master templates generation are excluded) are identified according to D. Thus, there will be

200 genuine identification cases (the palms to be identified are really in D) and 300 imposter cases (the palms to be identified are NOT in D).

For each live template T (SAX symbol string), a substring T_S of T is used for the identification. The start position of T_S in T is randomly chosen. The master template T_m of palm P_m which has a substring bearing the smallest MINDIST (d_{min}) to T_S is found out through sequential search. If d_{min} is smaller than a preset upper-bound threshold d_t, then T is identified as belonging to P_m. d_t is adjusted to find out the smallest possible number of the incorrect identification cases (false acceptance and false rejection). In our experiment, we chose the number of the SAX level to be 10 and the length of the master templates to be 200. The results are shown in Table 4.

Table 4. The accuracy of incomplete template identification

Length(T_s)/Length(T)	Minimum Possible Number of Fail Cases	
100%	1	0.2%
50%	7	1.4%
30%	27	5.4%
20%	51	10.2%
10%	120	24.0%

From the above table we can see that even by using only 30% the live templates, we can still achieve pretty high identification accuracy of 94.6%. Fig. 10 shows the matching of an incomplete template with the corresponding master template. As the length of the master template is short (200) and MINDIST calculation process is very fast, the time needed for the identification process is acceptable even if sequential search is adopted.

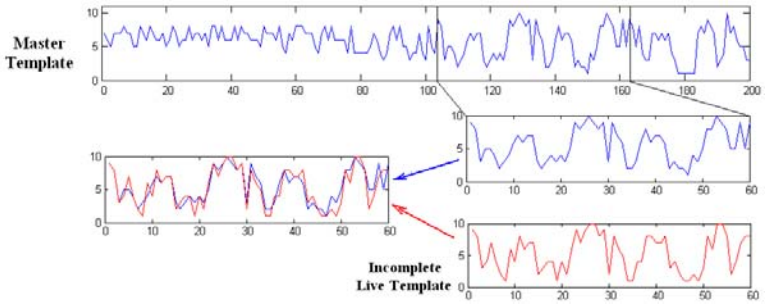


Fig. 10. Incomplete template matching

5 Conclusion and Future Work

We have proposed a novel approach for the palmprint authentication using time series technologies. Using of SAX for the template representation is the key of this new approach. It has the following advantages: first, it is simple to implement and the overall computational complexity is very low compared to previous works; second, it is very flexible as both the local features and the SAX parameters can be adjusted according to different system requirements; third, the SAX representation (essentially

symbol string) makes it very convenient for the implementation of multi-biometrics using feature fusion; fourth, feasibility of the incomplete template matching (substring of SAX) makes this approach more applicable.

However, there are still several problems to be solved. First, the size of the palmprint database used in the experiment is relatively small. Bigger database should be involved to further test the effectiveness of this approach; second, the verification accuracy (98.7%) can still be improved, considering the highest reported results (99.4% in [3], 99.2% in [2]). Using more appropriate local features may improve the accuracy; third, after a careful study of the palmprints for which our approach fails, we find out that most of these palmprints are too bright in certain sub-areas so that the local features are seriously blurred. How to automatically discard these low-quality areas needs to be investigated; fourth, using Dynamic Time Warping (DTW) might help to address the problem of palm distortion and further increase the accuracy [19]; finally, fast string indexing algorithms [18] can be used to improve the efficiency of the incomplete template matching. Also, using non-contiguous SAX slices can make the incomplete template matching more applicable.

Acknowledgement

This work was partially supported by the Hong Kong Research Grants Council Project 2300011, "Towards Multi-Modal Human-Computer Dialog Interactions with Minimally Intrusive Biometric Security Functions.

References

1. W. Shu, D. Zhang, Automated Personal Identification by Palmprint, Opt. Eng. 37 (8), pp. 2359 – 2362, 1998
2. X.Q. Wu, D. Zhang, K. Q. Wang, Fisherpalms Based Palmprint Recognition, Pattern Recognition Letters, vol. 24, pp. 2829-2838, 2003
3. W.K. Kong, D. Zhang, W. Li, Palmprint Feature Extraction using 2-D Gabor Filters", Pattern Recognition, vol. 36, pp. 2339-2347, 2003
4. X.Q. Wu, K.Q. Wang, D. Zhang, Wavelet Based Palmprint Recognition, Proceedings of the First International Conference on Machine Learning and Cybernetics, vol. 3, pp. 1253-1257, 2002
5. W.Li, D. Zhang, Z. Xu, Palmprint Identification by Fourier Transform, Int. J. Pattern Recognition Artificial Intell. 16 (4), pp. 417 – 432, 2002
6. D. Zhang, W. Shu, Two Novel Characteristics in Palmprint Verification: Datum Point Invariance and Line Feature Matching, Pattern Recognition, 32, pp. 691-702, 1999
7. G.M. Liu, D. Zhang, K.Q. Wang, Palmprint Recognition Using Eigenpalms Features, Pattern Recognition Letters 24 (2003) pp. 1463 – 1467
8. M. Vlachos, G. Kollios, D. Gunopulos, Discovering Similar Multidimensional Trajectories, Proceedings of the 18th International Conference on Data Engineering, 2002
9. J. Lin, E. Keogh, S. Lonardi, P. Patel, Finding Motifs in Time Series, Proceedings of the 2nd Workshop on Temporal Data Mining, 8th ACM SIGKDD, pp. 53 – 68, July, 2002
10. A. Apostolico, M.E. Bock, S. Lonardi, Monotony of Surprise and Large-Scale Quest for Unusual Words. In proceedings of the 6th Int'l Conference on Research in Computational Molecular Biology, pp 22-31, April, 2002
11. Gionis, H. Mannila, Finding Recurrent Sources in Sequences, Proceedings of the 7th International Conference on Research in Computational Molecular Biology, April, 2003

12. T.M. Rath, R. Manmatha, "Word Image Matching Using Dynamic Time Warping", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pp. 521-527, June, 2003
13. S.A.C. Schuckers, S.T.V. Parthasaradhi, R. Derakshani, L. A. Hornak, "Comparison of Classification Methods for Time-Series Detection of Perspiration as a Liveness Test in Fingerprint Devices", Proceedings of ICBA 2004, Lecture Notes in Computer Science, LNCS 3072, pp. 256 – 263, July, 2004
14. J. Lin, E. Keogh, S. Lonardi, B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms", Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, CA. pp. 2 -11, June, 2003
15. PolyU Palmprint Palmprint Database, http://www.comp.polyu.edu.hk/~biometrics
16. A.K. Jain, Arun Ross, Multibiometric Systems, Communications of the ACM, Vol. 47, No. 1, pp. 34 – 40, Jan. 2004
17. Kurmar, D. Wong, H.C. Shen, A.K. Jain, Personal Verification Using Palmprint and Hand Geometry Biometric, Proceedings of AVBPA 2003, pp. 668 – 678, June, 2003
18. D. Gusfield, Algorithms on strings, trees, and sequences: computer science and computational biology, New York, Cambridge University Press, 1997
19. S.W. Kim, S.H. Park, W. Chu, Efficient processing of similarity search under time warping in sequence databases: an index-based approach, Information Systems, Vol. 29, Issue 5, pp. 405-420, 2004
20. K.R. Rao, P.C. Yip, The Transform and Data Compression Handbook, CRC Press, Boca Raton, Fla, 2001
21. A. Rosenfeld and A.C. Kak, Digital Picture Processing, Academic Press, San Diego, 1982
22. M.H. Han, D. Jang, The use of maximum curvature points for the recognition of partially occluded objects, Pattern Recognition, vol. 23, pp. 21-23, 1990

Ear Biometrics by Force Field Convergence

David J. Hurley, Mark S. Nixon, and John N. Carter

School of Electronics and Computer Science,
University of Southampton, SO17 1BJ, UK
djh@analyticalengines.co.uk, {msn,jnc}@ecs.soton.ac.uk

Abstract. We have previously described how force field feature extraction can be used to exploit the directional properties of a force field generated from an ear image to automatically locate potential wells and channels which then form the basis of characteristic ear features. We now show how an analysis of the mechanism of this algorithmic field line approach leads to an additional closed analytical description based on the divergence of force direction revealing even more information in the form of anti-wells and anti-channels. In addition to furnishing specific implementation details for much faster FFT based computation and demonstrating brightness insensitivity, the technique is validated by achieving a recognition rate of 99.2% on a set of 252 ear images taken from the XM2VTS face database. These results demonstrate the inherent automatic extraction advantage of the new technique, especially when compared with more traditional PCA where we show that the ear set has to be more accurately extracted and registered in order to achieve comparable results. We show that it performs even more favourably against PCA under variable brightness conditions, and we also demonstrate its excellent noise performance by showing that noise has little effect on recognition results. Thus we have introduced a powerful new extension to complement our existing technique and we have validated it by achieving good ear recognition results, and in the process we have contributed to the mounting evidence that the human ear has considerable biometric value.

1 Introduction

In the past there has been little interest in ears as a biometric but this is now changing with recent contributions from computer vision researchers [1,2,3,4]. Ear prints have been used in forensics [5] but their reliability has recently been challenged in the Courts [6,7]. We have proposed the Force Field Transform and Force Field Feature Extraction in the context of ear biometrics [8,9,10,11] which effectively filters the ear image by convolving it with a huge inverse square kernel more than four times the size of the image, the force then being the gradient of the resulting massively smoothed image. The force field paradigm allows us to draw upon a wealth of proven techniques from vector field calculus; indeed the present paper uses the divergence operator on the force direction leading to a nonlinear operator which we call convergence of force direction. Other researchers [12,13,14] have successfully applied traditional physics based mathematical modeling techniques to image processing with good results. The extreme kernel size results in the smoothed image having a general dome shape which gives rise to brightness sensitivity issues, but we argue by showing that the field line features are hardly distorted that this will have little overall effect and this conclusion is borne out by including brightness variation in our recognition

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 386–394, 2005.
© Springer-Verlag Berlin Heidelberg 2005

tests. On the other hand, the dome shape leads to an automatic extraction advantage and this is demonstrated by using poorly registered and poorly extracted images for our recognition tests and then comparing the results with those for PCA under the same conditions, where we see that the ear images have to be accurately extracted and registered for PCA to achieve comparable results. Here we extend the description of the implementation and investigate further performance attributes in noise. As such, we highlight a powerful new direction and show by experimental results that ears have substantial promise as a biometric.

2 Extracting and Recognizing Ears

2.1 Force Field Feature Extraction

Here we review the force field transform and algorithmic field line feature extraction before introducing convergence feature extraction. The mathematical concepts we use can be found in basic works on electromagnetics [16] and a more detailed description of the transform can be found in [9,11]. We consider faster computation using convolution and the FFT and also consider the question of brightness sensitivity both theoretically and by demonstration.

The image is first transformed to a force field by treating the pixels as an array of mutually attracting particles that attract each other according to the product of their intensities and inversely to the square of the distances between them. Each pixel is assumed to generate a spherically symmetrical force field so that the total force $\mathbf{F}(\mathbf{r}_j)$ exerted on a pixel of unit intensity at the pixel location with position vector $\mathbf{r}_j$ by a remote pixels with position vector $\mathbf{r}_i$ and pixel intensities $P(\mathbf{r}_i)$ is given by the vector summation,

$$\mathbf{F}(\mathbf{r}_j) = \sum_i \left\{ \begin{array}{c} P(\mathbf{r}_i) \dfrac{\mathbf{r}_i - \mathbf{r}_j}{\left|\mathbf{r}_i - \mathbf{r}_j\right|^3} \forall i \neq j \\ 0 \forall i = j \end{array} \right\} \tag{1}$$

The underlying energy field $E(\mathbf{r}_j)$ is similarly described by,

$$E(\mathbf{r}_j) = \sum_i \left\{ \begin{array}{c} \dfrac{P(\mathbf{r}_i)}{\left|\mathbf{r}_i - \mathbf{r}_j\right|} \forall i \neq j \\ 0 \forall i = j \end{array} \right\} \tag{2}$$

To calculate the force and energy fields for the entire image these calculations should be performed for every pixel but this requires the number of applications of equations 1 and 2 to be proportional to the square of the number of pixels, so for faster calculation the process is treated as a convolution of the image with the force field corresponding to a unit value test pixel, and then invoking the Convolution Theorem to perform the calculation as a frequency domain multiplication, the result of which is then transformed back into the spatial domain. The force field equation for an $M \times N$ pixel image becomes,

$$forcefield = \sqrt{M \times N}\, \Im^{-1}[\Im(unit\ forcefield) \times \Im(image)] \tag{3}$$

$$
\begin{aligned}
ff(pic) := \;& sr \leftarrow 2\cdot(rows(pic)-1),\ sc \leftarrow 2\cdot(cols(pic)-1) \\
& r \leftarrow rows(pic)-1,\ c \leftarrow cols(pic)-1 \\
& \text{for } rr \in 0..sr \\
& \quad \text{for } cc \in 0..sc \\
& \quad\quad usr_{rr,cc} \leftarrow \frac{(r+c\cdot j)-(rr+cc\cdot j)}{(|r+c\cdot j-(rr+cc\cdot j)|)^3} \\
& usr_{3\cdot rows(pic)-3,\,3\cdot cols(pic)-3} \leftarrow 0 \\
& pic_{3\cdot rows(pic)-3,\,3\cdot cols(pic)-3} \leftarrow 0 \\
& oup \leftarrow \sqrt{rows(pic)\cdot cols(pic)}\cdot icfft\left(\overrightarrow{(cfft(usr)\cdot cfft(pic))}\right) \\
& ff \leftarrow submatrix(oup, r, 2\cdot r, c, 2\cdot c)
\end{aligned}
$$

Fig. 1. Force field by convolution in Mathcad

where $\Im$ stands for the Fourier Transform and $\Im^{-1}$ for its inverse. Figure 1 shows how to implement this in *Mathcad* in which 1**j** denotes the complex operator and **cfft** and **icfft** denote the Fourier and inverse Fourier transforms, respectively. Also, because the technique is based on a natural force field there is the prospect of a hardware implementation in silicon by mapping the image pixels to electric charges, which would lead to very fast real time force field calculation.

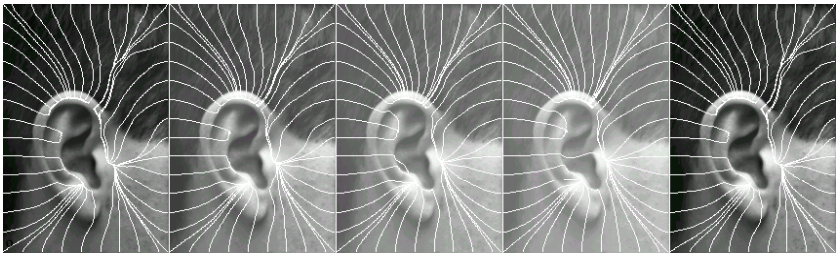
Figure 5(a) demonstrates field line feature extraction for an ear image where a set of 44 test pixels is arranged around the perimeter of the image and allowed to follow the field direction so that their trajectories form field lines which capture the general flow of the force field. The test pixel positions are advanced in increments of one pixel width, and the test pixel locations are maintained as real numbers, producing a smoother trajectory than if they were constrained to occupy exact pixel grid locations. Notice the two obvious potential wells in the lower part of the field.

The effect of brightness change will first be analysed by considering its effect on the energy field and then confirmed by visual experiment. Should the individual pixel intensity be scaled by a factor a and also have and an additive intensity component b, we would have,

$$E(\mathbf{r}_j) = \sum_i \left\{ \begin{matrix} \dfrac{aP(\mathbf{r}_i)+b}{|\mathbf{r}_i-\mathbf{r}_j|} \forall i \neq j \\ 0 \forall i = j \end{matrix} \right\} = a \sum_i \left\{ \begin{matrix} \dfrac{P(\mathbf{r}_i)}{|\mathbf{r}_i-\mathbf{r}_j|} \forall i \neq j \\ 0 \forall i = j \end{matrix} \right\} + \sum_i \left\{ \begin{matrix} \dfrac{b}{|\mathbf{r}_i-\mathbf{r}_j|} \forall i \neq j \\ 0 \forall i = j \end{matrix} \right\} \tag{4}$$

We see that scaling the pixel intensity by the factor a merely scales the energy intensity by the same factor a, whereas adding an offset b is more troublesome, effectively adding a pure energy component corresponding to an image with constant pixel intensity b. The effect of the offset and scaling is shown in Figure 2 with the channels superimposed. We see that scaling by a factor of 10 in (e) has no effect as expected. The original image in (a) has a mean value of 77 and a standard deviation of 47. Images (b) to (d) show the effect of progressively adding offsets of one standard deviation. At one standard deviation the effect is hardly noticeable and even at 3 standard deviations the change is by no means catastrophic as the channel structure alters little.

We therefore conclude that operational lighting variation in a controlled biometrics environment will have little effect. These conclusions are borne out by the results of the corresponding recognition experiments in Table 1.



(a) original (b) 1 std. dev. (c) 2 std. devs. (d) 3 std. devs. (e) scaled ×10

Fig. 2. Effect of additive and multiplicative brightness changes

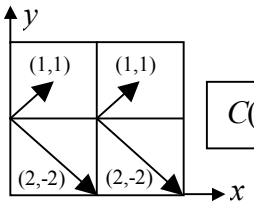
2.2 Convergence Feature Extraction

The analytical method came about as a result of analyzing in detail the mechanism of field line feature extraction. As shown in Figure 5(d), when the arrows usually used to depict a force field are replaced with unit magnitude arrows, thus modeling the directional behavior of exploratory test pixels, it becomes apparent that channels and wells arise as a result of patterns of arrows converging towards each other, at the interfaces between regions of almost uniform force direction. As this brings to mind the divergence operator of vector calculus, it was natural to investigate the nature of any relationship that might exist between channels and wells and this operator. This resulted not only in the discovery of a close correspondence between the two, but also revealed extra information corresponding to the interfaces between diverging arrows, leading to a more general description of channels and wells in the form of a mathematical function in which wells and channels are revealed to be peaks and ridges respectively in the function value. The new function maps the force field to a scalar field, taking the force as input and returning the additive inverse of the divergence of the force direction. The function will be referred to as the force direction convergence field $C(\mathbf{r})$ or just convergence for brevity. A more formal definition is given by

$$C(\mathbf{r}) = -\text{div}\,\mathbf{f}(\mathbf{r}) = -\lim_{\Delta A \to 0} \frac{\oint \mathbf{f}(\mathbf{r}) \cdot d\mathbf{l}}{\Delta A} = -\nabla \cdot \mathbf{f}(\mathbf{r}) = -\left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} \right) \tag{5}$$

where $\mathbf{f}(\mathbf{r}) = \dfrac{\mathbf{F}(\mathbf{r})}{|\mathbf{F}(\mathbf{r})|}$, ΔA is incremental area, and $d\mathbf{l}$ is its boundary outward normal. This function is real valued and takes negative values as well as positive ones where negative values correspond to force direction divergence. This calculation is illustrated for the top right hand element of a simple 4-element artificially constructed force field shown in Figure 3. The negative sign indicates negative convergence or net outward flux in this particular case.

Figure 4 shows a particular implementation of convergence in Mathcad where the matrix addressing scheme would require our simple example to be rotated by 90º clockwise to give the correct answer. FF represents the force field and DF is the direction field.



$$C(1,1) = -1/\sqrt{2} - 1/\sqrt{2} \;+\; 1/\sqrt{2} - (-1/\sqrt{2})\} \;=\; -\sqrt{2}$$

Fig. 3. Convergence calculation

$$\begin{aligned}
&C(FF) := \left|\begin{array}{l}
DF \leftarrow \dfrac{\overrightarrow{FF}}{|FF|} \\
\text{for } r \in 1..\text{rows}(DF) - 1 \\
\quad \text{for } c \in 1..\text{cols}(DF) - 1 \\
\quad\quad \left|\begin{array}{l}
dr \leftarrow \text{Re}\left(DF_{r,c}\right) - \text{Re}\left(DF_{r-1,c}\right) \\
dc \leftarrow \text{Im}\left(DF_{r,c}\right) - \text{Im}\left(DF_{r,c-1}\right) \\
C_{r,c} \leftarrow dr + dc
\end{array}\right. \\
-C
\end{array}\right.
\end{aligned}$$

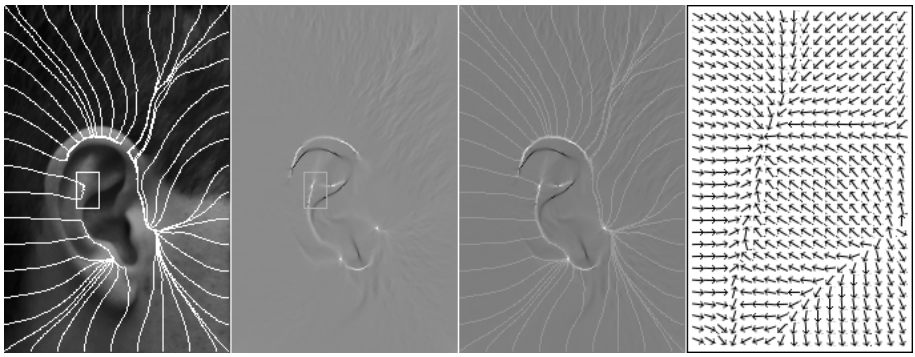
Fig. 4. Convergence implemented in Mathcad

We must also stress that convergence is non-linear because it is based on force direction rather than force. This nonlinearity means that we are obliged to perform the operations in the order shown; we cannot take the divergence of the force and then divide by the force magnitude. Div(grad/|grad|) ≠ (div grad)/|grad|. This is quite easily illustrated by a simple example using the scalar field e^x in Equation 6,

$$\left\{\begin{array}{c}
\text{div}\,(\text{grad}/|\,\text{grad}\,|) \\
\nabla \cdot \left(\dfrac{\nabla e^x}{|\nabla e^x|}\right) = \nabla \cdot \dfrac{e^x \mathbf{i}}{e^x} = \nabla \cdot \mathbf{i} = 0
\end{array}\right\} \neq \left\{\begin{array}{c}
(\text{div grad})/|\,\text{grad}\,| \\
\dfrac{\nabla \cdot \nabla e^x}{|\nabla e^x|} = \dfrac{e^x}{e^x} = 1
\end{array}\right\} \tag{6}$$

where **i** is a unit vector in the *x* direction. The convergence is zero because we have a field of parallel unit magnitude vectors, whereas in the second case the vectors are parallel but the magnitude changes, resulting in a net outflow of flux at any point. This illustrates that even though convergence looks very much like a Laplacian operator, it definitely is not.

Figure 5 shows the relationship between field lines (a) and convergence (b) by merging the two fields in (c). A small rectangular section of the force direction field indicated by a small rectangular insert in (a) and (b) is shown magnified in (d). We can see clearly how channels coincide with white convergence ridges and also that wells coincide with convergence peaks which appear as bright spots. Notice the extra information in the center of the convergence map that is not in the field line map. Negative convergence values representing antichannels appear as dark bands, and positive values corresponding to channels appear as white bands. We see that the



(a) field lines (b) convergence field (c) superimposition (d) force direction

Fig. 5. Convergence field

antichannels are dominated by the channels, and that the antichannels tend to lie within the confines of the channels. Notice also the correspondence between converging arrows and white ridges, and between diverging arrows and black ridges. The features detected tend to form in the center of the field due to its overall dome shape, with channels and wells tending to follow intensity ridges and peaks whereas antichannels and antiwells tend to follow intensity troughs and hollows.

3 Ear Recognition

The technique was validated on a set of 252 ear images taken from 63 subjects selected from the XM2VTS face database [15] by multiplicative template matching of ternary thresholded convergence maps where levels less than minus one standard deviation are mapped to -1, whilst those greater than one standard deviation map to +1, and those remaining map to 0. A threshold level of one standard deviation was chosen experimentally resulting in the template channel thickness shown in Figure 6(c). This figure also shows a rectangular exclusion zone centered on the convergence magnitude centroid; the centroid of the convergence tends to be stable with respect to the ear features and this approach has the added advantage of removing unwanted outliers such as bright spots caused by spectacles. The size of the rectangle was chosen as 71×51 pixels by adjusting its proportions to give a good fit for the majority of the convergence maps. Notice how for image 000-2 which is slightly lower than the other three, that the centroid-centered rectangle has correctly tracked the template downwards.

The inherent automatic extraction advantage was demonstrated by deliberately not accurately extracting or registering the ears in the sense that the database consists of 141×101 pixel images where the ears have only an average size of 111×73 and are only roughly located by eye in the center of these images. This can be seen clearly in Figure 6(a) where we see a marked variation both in vertical and horizontal ear-location, and also that there is a generous margin surrounding the ears. The force field technique gives a correct classification rate of 99.2% on this set, whereas running PCA [18] on the same set gives a result of only 62.4% but when the ears are accurately extracted by cropping to the average ear size of 111×73, running PCA then

gives a result of 98.4%, thus demonstrating the inherent extraction advantage. The first image of the four samples from each of the 63 subjects was used in forming the PCA covariance matrix. Figure 7 shows the first 4 eigenvectors for the 111x73-pixel images. The effect of brightness change by addition was also tested where we see that in the worst case where every odd image is subjected to an addition of 3 standard deviations the force field results only change by 2%. whereas those for PCA under the same conditions fall by 36%, or by 16% for normalized intensity PCA, thus confirming that the technique is robust under variable lighting conditions.

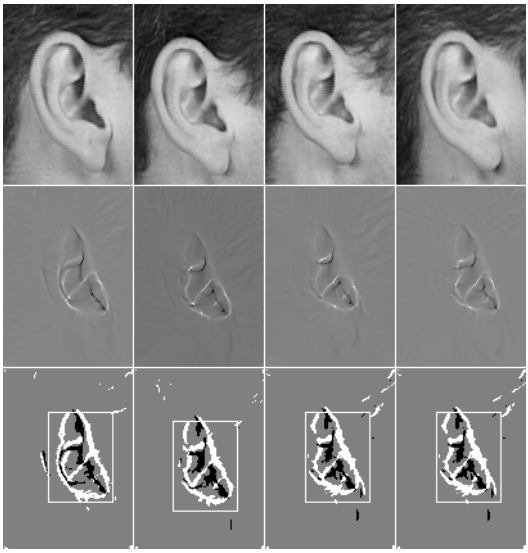


Fig. 6. Feature extraction for subject 000

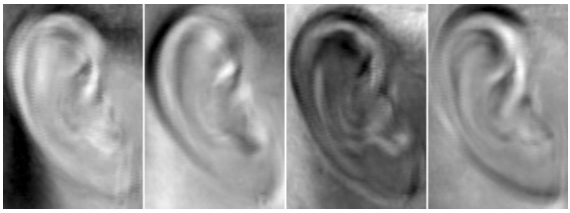


Fig. 7. First 4 eigenvectors for 111×73 pixel images

These results are presented in Table 1 where we also include the decidability index after Daugman [17] which combines the mean and standard deviation of the intra-class and inter-class measurement distributions giving a good single indication of the nature of the results. This index d' measures how well separated the distributions are, since recognition errors are caused by their overlap. The measure aims to give the highest scores to distributions with the widest separation between means, and smallest standard deviations. If the two means are μ_1 and μ_2 and the two standard deviations are σ_1 and σ_2 then d' is defined as

Table 1. Comparison of force field (FFE) and PCA recognition results

Image type	method	passes	Noise $20\log_{10}$S/N	CCR	bright. add. (std devs.)	decidability
141×101 with deliberately poor extraction and registration	FFE	250/252	Nil	99.2%	0	3.432
	FFE	251/252	18dB	99.6%	0	3.488
	FFE	249/252	12dB	98.8%	0	3.089
	FFE	241/252	6dB	95.6%	0	1.886
	FFE	250/252	Nil	99.2%	1	3.384
	FFE	247/252	Nil	98.0%	2	3.137
	FFE	245/252	Nil	97.2%	3	2.846
	PCA	118/189	Nil	62.4%	0	1.945
111×73 with accurate extraction and registration	PCA	186/189	Nil	98.4%	0	3.774
	PCA	186/189	18dB	98.4%	0	3.743
	PCA	186/189	12dB	98.4%	0	3.685
	PCA	177/189	6dB	93.6%	0	3.606
	PCA	130/189	Nil	68.8%	1	1.694
	PCA	120/189	Nil	63.6%	2	0.878
	PCA	118/189	Nil	62.4%	3	0.476
	PCA	181/189	Nil	95.6%	1 normalized	3.171
	PCA	172/189	Nil	91.0%	2 normalized	1.91
	PCA	166/189	Nil	82.5%	3 normalized	1.14

$$d' = \frac{|\mu_1 - \mu_2|}{\sqrt{\left(\sigma_1^{\,2} + \sigma_2^{\,2}\right)/2}} \tag{7}$$

Notice that the best case index for PCA is slightly higher than the value of 3.43 obtained for the 141×101 images but this could be attributed to the reduction in data set size from 252 to 189 also to the fact that the images have been more fully extracted for PCA. We have also included noise performance figures where noise has been modeled as additive noise with a zero mean Gaussian distribution The signal to noise ratios of 6dB, 12dB, and18dB used are calculated as $20\log_{10}$(S/N). We see that the technique enjoys excellent noise tolerance where even for extreme noise of 6dB the performance only falls by about 3.6%. Interestingly at a ratio of 18dB the recognition rate actually improves over the noiseless recognition rate, but this must be put down to the combination of small changes and the random nature of the noise process. For reference we have also included the corresponding noise results for PCA under the same conditions, where we see that PCA also performs well under noisy conditions but not quite as well as FFE at 6dB where the fall is about 4.8%.

4 Conclusions

In conclusion we may say that in the context of ear biometrics we have developed a new linear transform that transforms an ear image, with very powerful smoothing and without loss of information, into a smooth dome shaped surface whose special shape facilitates a new form of feature extraction that extracts the essential ear signature without the need for explicit ear extraction. We have described much faster force field calculation using convolution and the FFT and we have shown that the technique is robust under variable lighting conditions both by analysis and also by ex-

periment. We have introduced convergence feature extraction and shown that it is a valuable extension to our earlier field line feature extraction. We have validated the technique by experiment where we have shown that it compares very well with PCA especially under variable lighting conditions. In the process we have contributed to the mounting evidence in support of the recognition potential of the human ear for biometrics. In our future work we hope to refine our new technique and develop its full potential. We also hope to promote the case for ears as a new and promising biometric capable of competing in a rapidly developing field.

References

1. Burge, M., and Burger, W., Ear biometrics in computer vision, *Proc. ICPR 2000,* pp. 822-826, 2002
2. Chang, K., Bowyer, K. W., Sarkar, S., and Victor, B., Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics, *IEEE Trans. PAMI,* **25**(9), pp.1160-1165, 2003
3. Moreno,B., Sanchez,a., Velez.,J.F., On the Use of Outer Ear Images for Personal Identification in Security Applications, *Proc. IEEE 33rd Annual International Carnahan Conference on Security Technology*, 5-7 Oct. 1999
4. Bhanu, B., Chen, H., Human Ear Recognition in 3D, workshop on Multimodal User Authentication, Dec 2003, Santa Barbara, CA, p91-98.
5. Iannarelli, A., *Ear Identification*, Paramount Publishing Company, Freemont, California, 1989
6. STATE v. David Wayne KUNZE, Court of Appeals of Washington, Division 2. 97 Wash. App. 832, 988 P.2d 977, 1999
7. Mark Dallagher Released, *News item: The Chambers of William Clegg QC*, 28-Jan-2004, available from <www.2bedfordrow.co.uk/NewsDetail.asp?NewsID=17>
8. Hurley, D. J., Nixon, M. S. and Carter, J. N., Force Field Energy Functionals for Image Feature Extraction*. Proc. 10th British Machine Vision Conference BMVC99* pp. 604-613, 1999
9. Hurley, D. J., Nixon, M. S. and Carter, J. N., Force Field Energy Functionals for Image Feature Extraction*, Image and Vision Computing*, **20**, pp. 311-317, 2002
10. Hurley, D. J., Nixon, M. S. and Carter, J. N., A New Force Field Transform for Ear and Face Recognition*, Proceedings IEEE International Conference on Image Processing ICIP2000*, pp. 25-28, 2000
11. Hurley, D. J., Nixon, M. S. and Carter, J. N., Force Field Feature Extraction for Ear Biometrics*, Computer Vision and Image Understanding,* 2005 (in press)
12. Luo, B., Cross, A. D., Hancock, E. R., Corner Detection Via Topographic Analysis of Vector Potential*, Pattern Recognition Letters* **20**(6), pp. 635-650, 1999
13. Ahuja, N., A Transform for Multiscale Image Segmentation by Integrated Edge and Region Detection, *IEEE Transactions on PAMI*, **18**(12), pp. 1211-1235, 1996
14. Xu, C., Prince, J. L., Gradient Vector Flow: A New External Force for Snakes, *Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR)*, pp. 66-71, 1997
15. Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G., XM2VTSDB: The Extended M2VTS Database, *Proc. AVBPA'99* Washington D.C., 1999
16. Sadiku, M.N.O., *Elements of Electromagnetics*, Saunders College Publishing, Second Ed., 1989.
17. Daugman, J., Biometric decision landscapes, *Technical Report TR482*, University of Cambridge Computer Laboratory, 1999
18. Turk,M., Pentland,A., Eigenfaces for Recognition, *Journal of Cognitive Neuroscience*, pp.71-86, March 1991

Towards Scalable View-Invariant Gait Recognition: Multilinear Analysis for Gait

Chan-Su Lee and Ahmed Elgammal

Department of Computer Science
Rutgers University
Piscataway, NJ, USA

Abstract. In this paper we introduce a novel approach for learning view-invariant gait representation that does not require synthesizing particular views or any camera calibration. Given walking sequences captured from multiple views for multiple people, we fit a multilinear generative model using higher-order singular value decomposition which decomposes view factors, body configuration factors, and gait-style factors. Gait-style is a view-invariant, time-invariant, and speed-invariant gait signature that can then be used in recognition. In the recognition phase, a new walking cycle of unknown person in unknown view is automatically aligned to the learned model and then iterative procedure is used to solve for both the gait-style parameter and the view. The proposed framework allows for scalability to add a new person to already learned model even if a single cycle of a single view is available.

1 Introduction

Human gait is a valuable biometric cue that can be used for human identification similar to other biometrics such as faces and fingerprints. Gait has significant advantages compared to other biometrics since it is easily observable in an unintrusive way and is difficult to disguise [4]. Therefore, gait recognition has a great potential for human identification in public spaces for surveillance and for security [4, 10, 11, 21]. A fundamental challenge in gait recognition is to develop robust recognition algorithms that can extract gait features that are invariant to the presence of various conditions which affect people appearance. As a challenging problem in gait recognition, different conditions such as view, clothing, walking surface, and shoe type were presented in the NIST dataset [21]. Many gait recognition algorithms assume constrained conditions to reduce various sources that influence recognition accuracy. Two typical assumptions are fixed view, especially side view, and constant speed.

Generally, appearance-based approaches have been favorable in gait recognition [3, 9, 10, 13, 16–21, 24, 26, 30, 31] because, in typical application scenarios, people might be at a distance from the camera which inhibits accurate fitting of 3D models. Therefore, many gait recognition research focus on extracting view-based invariant gait signature for use in identification. Several attempts have been made to achieve view-invariant gait recognition [2, 8, 12, 23], mainly based on synthesizing side-view images from multiple views. For example, Shakhnarovich [23] used image-based visual hull to render a side view from multiple cameras. Kale [12] also presented the view invariant gait recognition algorithm by synthesizing a side view using perspective projection methods and optical based structures.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 395–405, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper we introduce a novel approach for learning view-invariant gait signature that does not require synthesizing particular views and doesn't require any camera calibration. Instead, in the learning phase, multiple views are used to extract an invariant gait signature while in recognition phase, any single view can be used to extract the gait signature directly. Given walking sequences captured from multiple views for multiple people, first we learn a nonlinear generative model for each walking cycle which enables re-sampling the cycle into temporally aligned gait cycles. Then we fit a multilinear generative model using higher-order singular value decomposition (HOSVD) [14] that decomposes view factors, body configuration factors, and gait-style factors. Gait-style is a view-invariant, time-invariant, and speed-invariant gait signature that can then be used in recognition. In the recognition phase, a new walking of unknown person in unknown view is aligned to the learned model and then iterative procedure is used to solve for both the gait-style and the view. Related work in using mulitlinear analysis for gait includes [27]

One important feature of the proposed framework is it's scalability to include new people. Given a learned model we can add a new person to the model even if only single view is available for that person, i.e., one cycle gait sequence from one of the multiple possible views is need to include new people to the database. This is a very important feature since in realistic scenarios, it is not always possible to have multiple view sequences of each person to be included in the database. Experimental results using CMU Mobo gait database and NIST-USF database [21] are reported in this paper.

The organization of the paper is as follows: In Section 2, we introduce temporal normalization of gait and cycle detection. Section 3 describes decomposition of gait-style, view, and body configuration parameters, estimation of style, and its application to gait recognition. Experimental results are described in Section 4 prior to the conclusion in Section 5.

2 Temporal Normalization by Manifold Embedding and Re-sampling

2.1 Input Representation

The inputs to the training and recognition phases are sequences of human silhouettes detected using background subtraction. We represent each shape instance (silhouette) as an implicit function $y(x)$ at each pixel x such that $y(x) = 0$ on the contour, $y(x) > 0$ inside the contour, and $y(x) < 0$ outside the contour. We use a signed-distance function such that

$$y(x) = \begin{cases} d_c(x) & x \text{ inside } c \\ 0 & x \text{ on } c \\ -d_c(x) & x \text{ outside } c \end{cases}$$

where the $d_c(x)$ is the distance to the closest point on the contour c with a positive sign inside the contour and a negative sign outside the contour. Such representation impose smoothness on the distance between shapes. Given such representation, each input silhouette is represented as a d-dimensional vector, i.e., a point $y \in R^d$ where d is the the dimensionality of the input space. Implicit function representation is typically used in level-set methods.

2.2 Temporal Normalization and Re-sampling

In order to achieve the training and recognition we need to obtain temporally aligned input silhouettes, i.e, obtaining body poses in correspondence during the gait cycle given any input sequence captured at any frame rate with any walking speed. To achieve this task we use any given input sequence to learn a nonlinear generative model that can be used to synthesize silhouettes at any temporal instance within the gait cycle.

The human gait evolves along a one-dimensional manifold embedded in a high dimensional visual space. Only one degree of freedom controls the walking cycle, which corresponds to the constrained body pose as a function of time. Such manifold is nonlinear and can be twisted on the high dimensional space given viewpoint, person shape, and clothing [5, 6]. Therefore, we embed each gait cycle temporally on a unit circle, which is a topologically homeomorphic one-dimensional manifold embedded in a two-dimensional Euclidean space.

In order to obtain synthesized gait poses, we learn a nonlinear mapping function from the manifold embedded on a unit circle and the input silhouettes. Learning nonlinear mapping is necessary since the manifold is embedded nonlinearly and arbitrarily into a unit circle. We use generalized radial basis function (GRBF) [22] to learn this mapping as a collection of interpolation functions. Let N equally spaced centers along a unit circle be $\{t_j \in R^2, j = 1, \cdots, N\}$ and given a set input images $\mathsf{Y} = \{y_i \in R^d, i = 1, \cdots, M\}$ and let their corresponding embedding along the unit circle be $\mathsf{X} = \{x_i \in R^2, i = 1, \cdots, M\}$, we can learn interpolations in the form

$$f^k(x) = p^k(x) + \sum_{i=1}^{N} w_i^k \phi(|x - t_i|), \tag{1}$$

that satisfies the interpolation condition $y_i^k = f^k(x_i)$ where y_i^k is the k-th pixel of input silhouette y_i, $\phi(\cdot)$ is a real valued basic function, w_i^k are real coefficients, $p^k(\cdot)$ is a linear polynomial, and $|\cdot|$ is the norm on R^2. The mapping coefficients can be obtained by solving a linear system [5]. Such mapping can be written in the form of a generative model as

$$f(x) = B \cdot \psi(x), \tag{2}$$

which nonlinearly maps any point x from the two dimensional embedding space into the input space. Therefore, the model can be used to synthesize N intermediate silhouettes at N standard time instances equally spaced along the unit circle. Re-sampling gait from the embedding space enables us to find temporally well aligned gait poses invariant to different walking speed and frame rate using equally spaced N embedding points.

2.3 Gait Cycle Detection in Arbitrary Views

Detection of gait cycles is essential for training and recognition. Typically, for side or frontal views, cycles can be detected using features such as width or height of bounding box, correlation of image sequences, etc, [1, 3]. However, detecting cycles in arbitrary views are difficult. The generative model, described above, facilitates detecting accurate gait cycles in any view given that the model parameter is learned on that particular view.

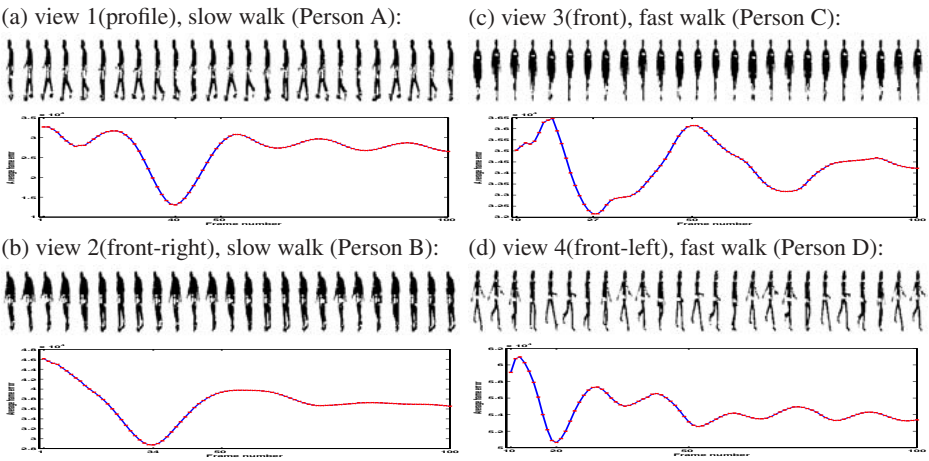


Fig. 1. Cycle detection in different view

Given an input sequence $y_i, i = 1, \cdots, M$ we need to find $k^* \leq M$ such that $y_1, \cdots, y_k$ corresponds to a full cycle. This can be achieved by finding k^* that minimizes the error between an input sequence of length k and model synthesized image sequence, of length k as well, starting from the same point. i.e., we need to find k^* such that

$$k^* = \arg \min_k = \frac{1}{k} \sum_{j=1}^{k} ||f(x_j^k) - y_j||$$

where x_j^k is a point on a unit circle with coordinate $x_j^k = [cos(2\pi \cdot j/k + \delta) \;\; sin(2\pi \cdot j/k + \delta)]$.

To show examples of generative model-based cycle detection, we used CMU Mobo gait data set which shows accurate detection of cycle in different views like side view, front-left view and front views within $1 \sim 2$ frames error. Fig. 1 shows four different view silhouette images (sampled at every 4th frames in the figure). Mean error are shown as a function of k in the range from 10 to 100. Even though we learned generative model for each view from one person, it performs accurately in segmenting cycles at different people.

3 Gait Style and View Decomposition

We model gait image sequences by three components: *gait style:* time-invariant and view-invariant personalized style of the gait which can be used for identification similar to in [16], *gait pose:* time-dependent factor representing body configuration during the gait cycle, and *gait view:* view-dependent factor representing variations of view.

3.1 Multilinear Model for Gait Analysis

Given different people walking sequences from different views, we detect gait cycles using gait cycle detection algorithm in Section 2.3. After cycle detection for every

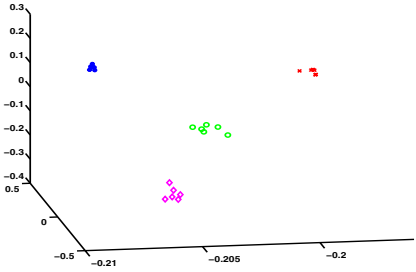
person, each cycle is used to learn the generative model described by equation 2 and re-sampled with the same number of temporally aligned poses. Therefore, the training data consists of N_s gait cycles[1], each captured from N_v different views, and each consists of N_p silhouette images representing aligned body poses. Each silhouette image is represented as a d dimensional vector using the representation described in section 2. The whole collection of aligned cycles for all different people and views is arranged into order four tensor (4-way array) $\mathcal{D}$ with dimensionality $N_s \times N_v \times N_p \times d$.

The data tensor $\mathcal{D}$ can be decomposed to parameterize orthogonal style, view, and pose factors using higher-order singular value decomposition (HOSVD). Higher-order singular value decomposition (HOSVD) is a generalization of SVD for multilinear model analysis by [14, 27, 28]. Multilinear model is a generalization of linear model (one-factor models) and bilinear model (two-factor models) [25] into higher-order tensor decomposition (multi-factor models). The data tensor $\mathcal{D}$ is decomposed to establish forth-order tensor using HOSVD which yields the decomposition

$$\mathcal{D} = \mathcal{Z} \times_1 \boldsymbol{S} \times_2 \boldsymbol{V} \times_3 \boldsymbol{P} \times_4 \boldsymbol{M}, \tag{3}$$

where $\boldsymbol{S}$, $\boldsymbol{V}$, $\boldsymbol{P}$, and $\boldsymbol{M}$ are orthogonal matrices with dimensionality $N_s \times N_s$, $N_v \times N_v$, $N_p \times N_p$, $d \times d$ corresponding to style, view, pose, and image orthogonal bases respectively. $\mathcal{Z}$ is a core tensor with the same dimensionality as the data tensor $\mathcal{D}$ which represents the interaction of the gait style, view, pose, and image pixel subspaces[2].

(a) Style clusters:



(b) View vectors:

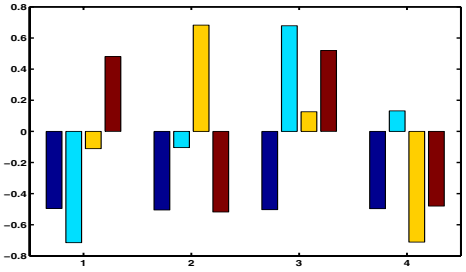


Fig. 2. Tensor analysis: 4 people with 6 cycles each from 4 different views. (a) First three style parameters for 6 gait cycles of each person. Each person's style shows good clustering within the person and good separation between different persons. (b)Four different view vectors, which are orthogonal to each others.

The orthogonal $N_s \times N_s$ matrix $\boldsymbol{S}$ spans the space of gait style parameters. In the style basis matrix $\boldsymbol{S} = [\boldsymbol{s}^1 \boldsymbol{s}^2 \cdots \boldsymbol{s}^s]^T$, each vector $\boldsymbol{s}^i$ represents a style parameter of person i as an N_s dimensional vector. This parameterization of the gait style independent of the view and body configuration is the basic feature we use in the recognition. Fig. 2 shows an example of the decomposition of gait style. We use 4 people from

[1] Each person can be represented by multiple cycles in the training data. So N_s represents the total number of cycles for all people.

[2] Reduced dimensional approximation can be achieved using higher-order orthogonal iteration method [15][29]

CMU-Mobogait data set with 6 cycles each from 4 different views to fit the model. As apparent in the figure, gait style parameters estimated from the different cycles of each person are clustered together in the style space.

Equation 3 can be rewritten as a generative model to synthesize gait cycles given any style vector $\boldsymbol{s}$ and view vector $\boldsymbol{v}$. This can be achieved by defining a new core tensor $\mathcal{B} = \mathcal{Z} \times_3 \mathcal{P} \times_4 \mathcal{M}$. Therefore, gait cycle images can be synthesized as $\boldsymbol{D}^{\boldsymbol{sv}}$ where

$$\boldsymbol{D}^{\boldsymbol{sv}} = \mathcal{B} \times_1 \boldsymbol{s} \times_2 \boldsymbol{v} \tag{4}$$

3.2 Gait Style Estimation from Unknown View and Style

Given images $y_1, \cdots, y_k$ representing a full gait cycle from unknown view with k frames, estimation of gait style is required for person identification. First, the sequence is used to learn a generative model in the form of Equation 2 and then the model is used to re-sample p gait images, $\boldsymbol{i}_1 \boldsymbol{i}_2 \cdots \boldsymbol{i}_p$, which are aligned with gait poses used in multilinear analysis. By stacking the gait images into a matrix $\boldsymbol{D} = [\boldsymbol{i}_1 \boldsymbol{i}_2 \cdots \boldsymbol{i}_p]$, the estimation of style and view can be formulated as solving for $\boldsymbol{s}$ and $\boldsymbol{v}$ that minimize error

$$E(\boldsymbol{s}, \boldsymbol{v}) = ||\boldsymbol{D} - \mathcal{B} \times_1 \boldsymbol{s} \times_2 \boldsymbol{v}||, \tag{5}$$

where $\boldsymbol{D}$ is $d \times N_p$ matrix. If the view vector $\boldsymbol{v}$ is known, we can obtain closed form solution for $\boldsymbol{s}$. This can be done by evaluating the product $\mathcal{H} = \mathcal{B} \times \boldsymbol{v}$ and unfolding the tensor $\mathcal{H}$ into a matrix by style-mode, i.e., $\boldsymbol{H}_{(\boldsymbol{1})} = unfolding(\mathcal{H}, 1)$. *Matrix unfolding* operation is explained in the appendix of this paper. The dimensions of $\boldsymbol{H}_{(\boldsymbol{1})}$ is $N_s \times (N_v \times N_p \times d)$. Solution for $\boldsymbol{s}$ can be obtained in closed form by solving the linear system $\boldsymbol{D} = \boldsymbol{H}_{(\boldsymbol{1})}{}^T \boldsymbol{s}$. Therefore estimation of $\boldsymbol{s}$ can be obtained by

$$\boldsymbol{s} = \left(\boldsymbol{H}_{(\boldsymbol{1})}{}^T\right)^+ \boldsymbol{D} \tag{6}$$

where $+$ is matrix pseudo-inverse operation using singular value decomposition (SVD). Similarly, we can analytically solve for $\boldsymbol{v}$ if the style vector $\boldsymbol{s}$ is known by forming a tensor $\mathcal{G} = \mathcal{B} \times_1 \boldsymbol{v}$ and forming its view-mode unfolding $\boldsymbol{G}_{(\boldsymbol{2})}$. Therefore, we can obtain the view as

$$\boldsymbol{v} = \left(\boldsymbol{G}_{(\boldsymbol{2})}{}^T\right)^+ \boldsymbol{D} \tag{7}$$

Iterative estimation of $\boldsymbol{s}$ and $\boldsymbol{v}$ using (6) and (7) leads to local minima for the error in (5). We can start initial style estimation by mean style $\boldsymbol{s} = (\sum_{i=1}^{N_s} s^i)/N_s$.

Given the estimated gait style vector $\boldsymbol{s}$, and different people's gait style vectors learned in the training, the recognition is a typical pattern classification problem. For the experimental results shown in this paper we used two simple classification approaches: Nearest Neighbor and Nearest class mean which shows very good recognition results. However, more sophisticated classification methods can be used to achieve even better results. The proposed framework can easily scale to include new people. Given a new person, theoretically, only one cycle from a single view is required to be able to solve for the person style parameter which can then be added to the trained database. In Section 4 we show experimental evaluation of the scalability and generalization of the model to learn style parameters from a single view and to recognize at different views.

(a) L_2 distance to each style vector in iteration: (b) L_2 distance to each view vector in iteration:

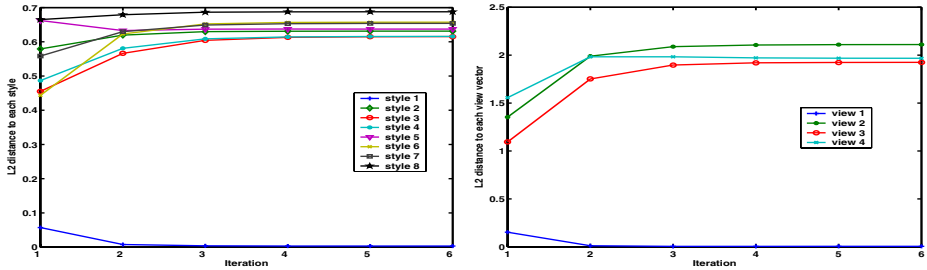


Fig. 3. Measurement of distance to style and view

Fig. 3 shows an example of the iterative estimation of view and gait style parameters. In this experiment we used 8 people with 4 different views from the Mobogait dataset to learn the model. The figure shows the change in the Euclidean distance to each mean style vector and mean view vector with the iterations. In this figure, a side view cycle for the first person was used for testing. It shows convergence to the correct style and view from the first iteration.

4 Experimental Results

We demonstrate the performance of the proposed algorithms on two databases: one is CMU mobo database and the other is USF-NIST gait database. In the preprocessing step, we applied median filter to remove noisy holes and spots. Bounding boxes which cover each person silhouettes were found and normalized to fixed size. Each silhouette shape is represented by a signed-distance function as described in section 2.

Experiment 1: Recognition of Gait in Different Speeds and Views: In this experiment we used CMU Mobo database, which has slow and fast walking sequences on a treadmill with six different views [7], to test gait recognition in different speeds and views. We chose a subset of 18 subjects which provided silhouettes for all different views and allowed finding proper bounding box for the subjects. Four different views (profile view, front-right view, front view, front-left view) were selected for multilinear gait analysis. Three cycles of slow walk for each person are used to learn the multilinear model parameters. In summary, the training data contains 18 people, 3 cycles each, from 4 views. Each person style is represented by the mean of the three style vectors obtained from three training cycles.

For evaluation we used three different slow-walk cycles and three fast-walk cycles for each of the 18 people with 4 views each. Overall there are 216 slow-walk evaluation cycles and 216 fast-walk evaluation cycles. For each evaluation cycle we estimate the view and the style of parmeteters of gait as described in Section 3.2. Finally, people are identified by finding closest style class mean. Table 1 shows the experiment result. For the slow-walk we achieve 100 % correct recognitions for all the views. For the fast-walk, we achieve around 90 % accuracy in average. The results shows fairly consistent recognition for all the different views. In both cases we achieve 100% view

Table 1. Gait recognition in different view and speed (CMU Data)

View class	slow walking sequences	fast walking sequences	Collins[3] (fast walking)
1(profile)	100%	88.9%	76%
2(front-right)	100%	88.9%	N/A
3(front)	100%	92.6%	100%
4(front-left)	100%	88.9%	N/A
Average	100%	90.0%	88%

identification. Even though we perform recognition for each cycle without knowing the view label, our results show better identification than template matching of key frames by Collins [3], shown in the forth column, which is tested for profile and front view separately using whole sequences.

Experiment 2: Generalization and Scalability Across Different Views: We evaluate the scalability of the proposed framework, i.e., Given a learned model, can we extend it to recognize a new person from different view points given that only one gait cycle from a single view is available for that person for training?

To evaluate this, we performed a new experiment by learning the model with a subset of subjects. Among 18 subjects, we learned the model using only eight subjects' slow walk sequences from 4 views. For the rest 10 subjects, only a single cycle data of slow walk from one view was given. We used this single view cycle to estimate gait style parameters. All the estimated style parameters are used as a database for recognition. The recognition is then evaluated using a test set consisting of 3 different slow-walk cycles and 3 fast-walk cycles from 4 views for all the 18 people.

Table 2 shows recognition results. We repeated the experiment by varying the view used in training for the 10 people with each single view cycle. Results show general identification capability to unknown views using style learned from a specific view. This clearly shows that the gait-style parameter is invariant to different view point. The identification performance varies across different views and the view used for training shows better performance on trained view class than others. Others, which do not learned style at all for the views,still, shows potentials for gait recognition. The performanace can be improved by using multiple cycles in the style estimation for given views.

Table 2. Gait recognition across different views(CMU Data)

View class	V1:slow	V2:slow	V3:slow	V4: slow	V1:fast	V2:fast	V3:fast	V4:fast
V1(profile)	96.3%	72.2%	53.7%	75.9%	53.7%	55.6%	40.7%	55.6%
V2(front-right)	72.2%	88.9%	59.3%	66.7%	53.7%	64.8%	48.2%	63.0%
V3(front)	51.9%	66.7%	90.9%	57.4%	50.0%	59.3%	92.6%	53.7%
V4(front-left)	59.3%	75.9%	70.7%	98.1%	46.3%	46.3%	55.6%	63.0%
Average(all)	69.9%	75.9%	68.7%	87.5%	50.9%	56.5%	59.3%	58.8%

Experiment 3: Recognition of Gait with Continuous Variation of Views (USF dataset): In this experiment we use NIST-USF Gait database [21] to evaluate perfor-

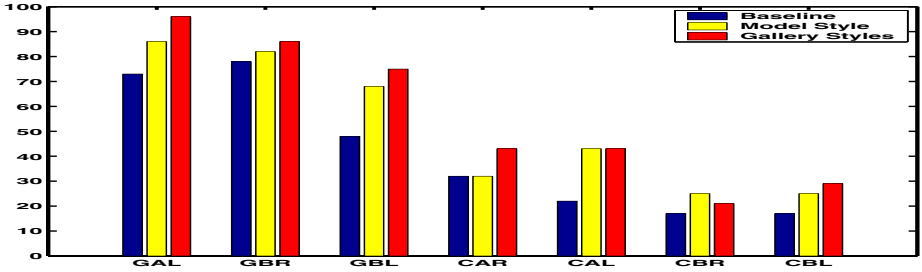


Fig. 4. Recognition result

Table 3. Comparison of Recognition with Baseline (USF Data)

Difference	Probe Set	Baseline	Nearest Mean	Nearest Neighbor	Kale [13]
View	GAL	73%	86%	96 %	89 %
Shoe	GBR	78%	82%	86 %	88 %
Shoe, view	GBL	48%	68%	75 %	68 %
Surface	CAR	32%	32%	43 %	35 %
Surface, shoe	CBR	22%	43%	43 %	28 %
surface, view	CAL	17%	25%	21 %	15 %
Surface, shoe, view	CBL	17%	25%	29 %	21 %

mance of gait recognition with continuous variation of the view due to the elliptical course that people used in capturing the database. We arbitrary select 28 people for a preliminary evaluation. We choose GAR, which is the gait sequence in grass surface, shoes type A, and right camera sequences, as a gallery set and tested by seven probe sets with variants in view, shoe and surface. Seven cycles were detected from the gallery sets and the probe sets. Three representative cycles of different views were selected from each sequence of gallery sets to learn the model.

For recognition we evaluated two classifiers for each estimated gait-style parameter for each test cycle: nearest style class mean (Model Style) and nearest neighbor style (Gallery styles). In both cases, we used majority vote from different test cycles to determine final person id. Results are shown in Table 3 and Fig. 4. Table 3 also shows recognition results reported in baseline evaluation [21] and recognition results reported using HMM by Kale *et al* in [13].

5 Conclusion

We presented a new framework to gait recognition which first uses a nonlinear generative model to re-sample gait sequences and then uses multilinear analysis to decompose view-invariant time-invariant gait parameters for identification. We showed promising human identification results in different views and speeds in CMU dataset. In USF dataset, which has continues view point variations within each probe set, also shows improvement in identification using the proposed view invariant iterative style estimation framework. We used very simple classification algorithms for identification from

the estimated gait style parameters. Recognition can be further improved by employing more sophisticated classification algorithms such as support vector machine (SVM) using style vectors. In the future we plan to report gait recognition for larger data sets.

Appendix

Matrix unfolding operation: Given an r-order tensor $\mathcal{A}$ with dimensions $N_1 \times N_2 \times \cdots \times N_r$, the mode-$n$ matrix unfolding, denoted by $unfolding(\boldsymbol{\mathcal{A}}, n)$, is flattening $\mathcal{A}$ into a matrix whose column vectors are the mode-n vectors [14, 27]. Therefore, the dimension of the unfolded matrix $\boldsymbol{A_{(n)}}$ is $N_n \times (N_1 \times N_2 \times \cdots N_{n-1} \times N_{n+1} \times \cdots N_r)$.

References

1. C. BenAdbelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Proc. FGR*, pages 357–363, 2002.
2. B. Bhanu and J. Han. Individual recognition by kinematic-based gait analysis. In *Proc. ICPR*, volume 3, pages 343– 346, 2002.
3. R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *Proc. FGR*, pages 351–366, 2002.
4. D. Cunado, M. S. Nixon, and J. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90:1–41, 2003.
5. A. Elgammal. Nonlinear generative models for dynamic shape and dynamic appearance. In *Proc. Int. Workshop GMBV*, 2004.
6. A. Elgammal and C.-S. Lee. Separating style and content on a nonlinear manifold. In *Proc. CVPR*, pages 478–485, 2004.
7. R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.
8. J. Han and B. Bhanu. Statistical feature fusion for gait-based human recognition. In *Proc. CVPR*, pages 842–847, 2004.
9. Q. He and C. Debrunner. Individual recognition from periodic activity using hidden markov models. In *In IEEE Workshop on Human Motion*, pages 47–52, 2000.
10. P. Huang, C. Haris, and M. Nixon. Recogising humans by gait via parametric canonical space. *Artificial Intelligence in Engineering*, 13:359–366, 1999.
11. A. Y. Johnson and A. F. Bobick. A multi-view method for gait recognition using static body parameters. In *Proc. AVBPA*, pages 301–311, June 2001.
12. A. Kale, A. K. R. Chowdhury, and R. Chellappa. Towards a view invariant gait recognition algorithm. In *Proc. on Advanced Video and Signal Based Surveillance*, pages 143–150, 2003.
13. A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa. Identification of human using gait. *IEEE Trans. Image Processing*, 13(9):1163–1173, 2004.
14. L. D. Lathauwer, B. de Moor, and J. Vandewalle. A multilinear singular value decomposiiton. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1253–1278, 2000.
15. L. D. Lathauwer, B. de Moor, and J. Vandewalle. On the best rank-1 and rank-(r1, r2, ..., rn) approximation of higher-order tensors. *SIAM Journal On Matrix Analysis and Applications*, 21(4):1324–1342, 2000.
16. C.-S. Lee and A. Elgammal. Gait style and gait content: Bilinear model for gait recogntion using gait re-sampling. In *Proc. FGR*, pages 147–152, 2004.

17. L. Lee, G. Dalley, and K. Tieu. Learning pedestrian models for silhouette refinement. In *Proc. ICCV*, pages 663–670, 2003.
18. Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frienze patterns. In *Proc. ECCV*, pages 657–671, 2002.
19. Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *Proc. ICPR*, pages 211–214, 2004.
20. H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysisand lip reading. *Pattern Recognition Letters*, 17:155–162, 1996.
21. P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer. Baseline results for the challenge problem of human id using gait analysis. In *Proc. FGR*, pages 137–142, 2002.
22. T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.
23. G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Proc. CVPR*, pages 439–446, 2001.
24. R. Tanawongsuwan and A. Bobick. Modelling the effects of walking speed on appearance-based gait recognition. In *Proc. CVPR*, pages 783–790, 2004.
25. J. B. Tenenbaum and W. T. Freeman. Separating style and content with biliear models. *Neural Computation*, 12:1247–1283, 2000.
26. D. Tolliver and R. T. Collins. Gait shape estimation for identification. In *Proc. AVBPA*, pages 734–742, 2003.
27. M. A. O. Vasilescu. Human motion signatures: Analysis, synthesis, recogntion. In *Proc. ICPR*, volume 3, pages 456–460, 2002.
28. M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. ECCV*, pages 447–460, 2002.
29. M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Proc. CVPR*, 2003.
30. L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. PAMI*, 25(12):1505–1518, 2003.
31. G. Zhao, R. Chen, G. Liu, and H. Li. Amplitude spectrum-based gait recognition. In *Proc. FGR*, pages 23–28, 2004.

Combining Verification Decisions in a Multi-vendor Environment*

Michael Beattie, B.V.K. Vijaya Kumar, Simon Lucey, and Ozan K. Tonguz

Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Abstract. Building access control represents an important application for biometric verification but often requires greater accuracy than can be provided by a single verifier. Even as algorithms continue to improve, poor samples and environmental factors will continue to impact performance in the building environment. We aim to improve verification accuracy by combining decisions from multiple verifiers spread throughout a building. In particular, we combine verifiers along the path traced out by each subject. When combining these decisions, the assumption of conditional independence simplifies implementation but can potentially lead to suboptimal performance. Through empirical evaluation of two related algorithms, we show that the assumption of conditional independence does not significantly impact verification accuracy. We argue that such a small reduction in accuracy can be attributed to the relative accuracy of each verifier. As a result, decisions can be combined using low complexity fusion rules without concerns of degraded accuracy.

1 Introduction

Combining decisions from a small number of biometric verifiers is a common strategy for improving verification accuracy [1–3]. In most contexts, the set of verifiers being combined is carefully selected during the design of a system, and all verifiers share a single location. For example, a system might use face and fingerprint as in [3]. In the following, we investigate the slightly different problem of combining biometric verifiers distributed throughout a secure building. One can envision a scenario in which a biometric verifier is placed at each secure door to assist in deciding whether the door should be unlocked for a particular subject.

As a subject passes through multiple secure doors, he traces a path of verification attempts. A building access control system is able to observe this path by forcing each subject to claim his or her identity with a key card. Rather than relying only on local decisions from each verifier, the access control system can construct a more informed decision using information from all verifiers in a path. This scenario differs from the standard combination of multiple biometric verifiers because the path differs from subject to subject and varies over time.

* This work has been supported in part by the National Institute of Standards and Technology (NIST) Building and Fire Research Laboratory

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 406–415, 2005.
© Springer-Verlag Berlin Heidelberg 2005

A path may include multiple modalities, multiple algorithms, or even multiple instances of identical devices. If verifiers from different vendors are combined, they might provide different types of verification results (e.g., match scores as opposed to binary decisions). If multiple instances of a single verifier appear in a path, their decisions may be correlated. We expect these identical verifiers to be common in buildings, where there is a significant motivation for access control systems to rely on a small number of verifier models. In particular, this simplifies purchase, installation and continued support of the system.

To ensure verification results of the same type, we restrict our analysis to decision level fusion [4], and assume that each verifier is assigned a pair of error rates based on testing by its respective vendor. These error rates correspond to the false accept rate (FAR), the rate at which an imposter is accepted as authentic, and false reject rate (FRR), the rate at which an authentic claimant is rejected. In general, joint characterizations of the error rates will not be available for the ensemble of verifiers. For example, the rate at which verifier A accepts a claimant and verifier B rejects that same claimant are not known because A and B are tested by different vendors.

Previously, we have described a technique for combining the decisions made along a path using these marginal error rates and an assumption of conditional independence [5]. In the following, we evaluate the impact of this assumption on the error rates of fused decisions. First, we show that multiple instances of the same verifier can be moderately correlated. Then, we evaluate two strategies that differ only in whether or not conditional independence is assumed. This evaluation suggests that the independence assumption does not significantly alter verification accuracy when combining verifiers at the decision level.

Domingos and Pazanni have previously suggested that scenarios exist in which a classifier that assumes conditional independence is optimal even for highly correlated data [6]. In contrast, we have found evidence that, for our problem, moderate correlation has a negligible impact on accuracy even when the conditions set forth in [6] are not met. This scenario occurs when combining decisions from relatively accurate verifiers rather than raw features. While it is possible for the conditional independence assumption to lead to suboptimal decisions, such decisions become increasingly rare with more accurate verifiers and longer paths. This observation permits a significant simplification of combination strategies while maintaining high accuracy.

2 The Building Environment

Combining biometric verifiers along a subject's path differs from a traditional multi-biometric system in several ways. One major difference is that verifiers may be purchased from multiple vendors, making joint training infeasible prior to system installation. As a result, joint characterizations of verifier error rates are not available. The internal operation of verifiers might also be opaque to protect the intellectual property of each vendor. In many cases, this concern leads to verifiers that provide only a decision – sometimes without even providing a configurable verification threshold.

When installing biometric verifiers throughout a building, system integrators are likely to use only a small number of distinct verifier models. This implies that any given path is likely to include multiple instances of the same verifier. Intuitively, two identical verifiers are likely to make errors on the same subject, violating the conditional independence assumption employed in [5] and elsewhere. For this reason, we expect that combination strategies assuming conditional independence will be outperformed by those that do not rely on such an assumption. Contrary to this expectation, our evaluation demonstrates only a marginal difference in verification accuracy.

In this paper, we assume the following: each verifier emits only a decision, and both FAR and FRR measurements are provided for each verifier. We choose this model based on the fact that vendors are not likely to expose the internals of their product. Some devices might provide match scores, and this information could be exploited using class conditional score models. We do not explore this possibility under the assumption that vendors are not likely to provide compatible or even accurate score models.

3 Correlation Measurements

To assess the degree of independence among decisions from multiple biometric verifiers, we measured the correlation coefficient for decisions from pairs of biometric verifiers. Our data consists of match scores that were generated by several algorithms against the XM2VTS database [7]. We refer to the scores from a single algorithm as a *score set*. Each score set represents a different algorithm entered in the face verification competition held at ICPR2000 [8]. Because all score sets were generated from the same database, we can use this data to estimate the degree of correlation between multiple verifiers on the same subjects. To construct decisions from these scores, we applied a maximum likelihood (ML) criterion to Gaussian Mixture Models for imposter and authentic score distributions. The associated thresholds and error rates for each verifier are listed in Table 1.

Table 1. Local thresholds and corresponding error rates

Algorithm	Threshold	FAR	FRR
AUT1	0.5359	0.0352	0.0825
EPFL	0.4334	0.0321	0.1225
USYD1	0.6785	0.1004	0.1525
SURREY2	0.1677	0.0422	0.0775

There are three scenarios where we expect to find correlation between verifier decisions. The first scenario is when two different verifiers observe the same subject (Fig. 1a). In this case, a particular authentic claimant might be difficult to verify or an imposter might resemble one of the valid subjects. The second scenario is when two identical verifiers observe the same subject (Fig. 1b). Each verifier captures a different sample so that different decisions are possible. In this case, errors inherent to this specific verification algorithm will occur at

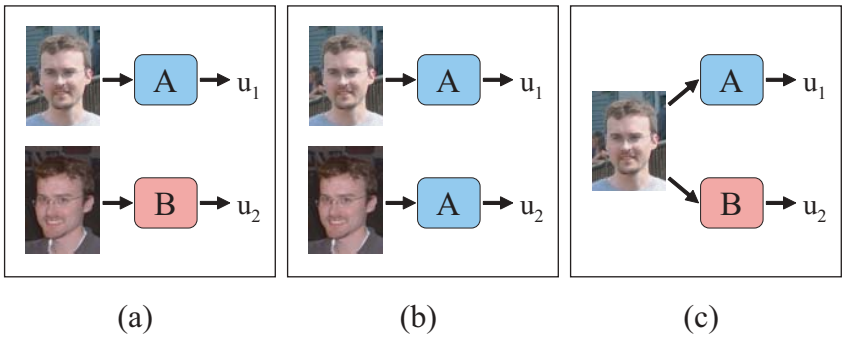


Fig. 1. Scenarios in which two verifiers construct two distinct decisions

both verifiers. The third scenario is when two different verification algorithms observe the same image (Fig. 1c), but this scenario will not occur in the building access context that we are investigating because each verifier will capture a new biometric sample. We do not expect to find correlation between two different modalities, and, for this reason, focus our attention here on face verification.

Without any additional processing, the XM2VTS score sets represent the third scenario described above. If the scores from each set are stored in the same order, then the n^{th} decision from two different sets will correspond to the same image. To generate decision pairs corresponding to scenarios (a) and (b) in Fig. 1, we observe that the XM2VTS database contains multiple face images for each subject. By carefully permuting the order of scores such that the n^{th} decision from two different sets correspond to different images of the same subject, we can construct scenarios (a) and (b). Specifically, a pair of score sets with one set in the original order and one set in the permuted order represents the combination of two decisions made from different samples of the same subject. Scenario (a) can be emulated when the pair is formed from an original set and the permuted variant of another set. Scenario (b) can be emulated when the pair is formed from an original set and its own permuted variant. It is also possible to construct independent verifiers by permuting the order such that each subject is aligned with scores from another subject. We call both procedures *score permutation*.

With score permutation defined, it is straightforward to calculate the correlation coefficient between two different biometric verifiers in each correlation scenario. We choose this metric over others presented in [9] because it is normalized according to individual verifier accuracy – allowing for comparison across different pairs of verifiers. Following [9], we construct the correlation coefficient using (1). In this equation, u_i represents the decision made by verifier i.

$$a = P\{\, u_1 = \text{accept}, u_2 = \text{accept}\,\}, \qquad b = P\{\, u_1 = \text{accept}, u_2 = \text{reject}\,\}$$
$$c = P\{\, u_1 = \text{reject}, u_2 = \text{accept}\,\}, \qquad d = P\{\, u_1 = \text{reject}, u_2 = \text{reject}\,\}$$

$$\rho = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{1}$$

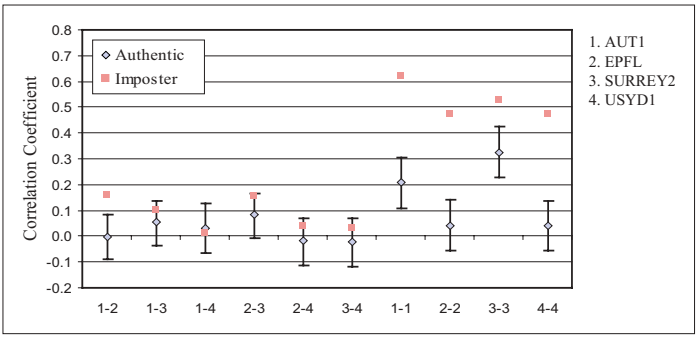


Fig. 2. Decision correlation for pairs of verifiers

There are actually two correlation coefficients for each pair of verifiers: one each from the imposter and authentic distributions. Both correlation coefficients are presented in Fig. 2, where the values of a, b, c, and d have been calculated using relative frequency estimates over the training data. Due to the relatively small number of authentic images, error bars have been included for the 95% confidence interval surrounding each authentic correlation coefficient.

It is clear from Fig. 2 that significant correlation exists between decisions from two identical verifiers, but correlation between two different verifiers is either small or statistically insignificant. Also apparent from this figure is that correlation among imposter decisions is stronger than for authentics in the case of identical verifiers – possibly indicating some repeatable impersonations.

4 Fusion Strategies

We have implemented an evaluation tool that combines decisions using two different algorithms: one that assumes conditional independence and another that does not. Both of these operate by estimating the class conditional probability (i.e., likelihood) for a path of decisions and emitting a maximum likelihood estimate of the underlying class (authentic or imposter). The only difference between the two algorithms is the use of a conditional independence assumption in the first. Both algorithms rely on the likelihood ratio test in (2) below.

$$\frac{P\{\,\mathbf{u}\,|\,\omega_1\,\}}{P\{\,\mathbf{u}\,|\,\omega_0\,\}} \underset{u_0=0}{\overset{u_0=1}{\gtrless}} W_{SEC} \tag{2}$$

In (2), $\mathbf{u}$ is a vector of local verifier decisions, ω_1 designates the authentic class and ω_0 designates the imposter class. The output u_0 is the path decision, and W_{SEC} defines the relative importance of each type of error. If both false accept and false reject errors are equally important, then W_{SEC} has a value of 1. A large W_{SEC} value indicates that false accept errors are more costly than false reject and vice versa.

If we assume conditional independence, then the likelihood in (2) can be expanded into a product of marginal likelihoods (3). Each of these marginal likelihoods can in turn be constructed from FAR and FRR estimates for each verifier. We note that this is simply a Naive Bayes Classifier that operates on local verifier decisions and refer to it hereafter as NB.

$$P\{\,\mathbf{u} \mid \omega_i\,\} = \prod_{i=1}^{N} P\{\,u_i \mid \omega_i\,\} \tag{3}$$

$$P\{\,u_i \mid \omega_1\,\} = \begin{cases} \mathrm{FAR}_i & u_i = 0 \\ 1 - \mathrm{FAR}_i & u_i = 1 \end{cases}$$

$$P\{\,u_i \mid \omega_0\,\} = \begin{cases} 1 - \mathrm{FRR}_i & u_i = 0 \\ \mathrm{FRR}_i & u_i = 1 \end{cases}$$

Without the conditional independence assumption, we are forced to estimate the joint likelihood for the vector of decisions $\mathbf{u}$. For the purposes of evaluation, it is feasible to operate on one path at a time and measure the relative frequency of each possible decision vector. This approach is possible because the set of possible decision vectors is manageable given the length of our test paths. For long paths, the number of possible vectors increases exponentially and training data is usually insufficient to estimate the likelihood of each vector. We refer to this estimation of the full likelihood expression as Full Bayes or simply FB. This strategy is defined precisely in (4), where $N_{\mathbf{u} \mid \omega_i}$ represents the number of times the vector $\mathbf{u}$ is produced by class ω_i and N_{ω_i} represents the total number of vectors from class ω_i.

$$P\{\,\mathbf{u} \mid \omega_i\,\} = \frac{N_{\mathbf{u} \mid \omega_i}}{N_{\omega_i}} \tag{4}$$

The FB approach enables decision combination without the assumption of conditional independence for evaluation, but it does not represent a viable strategy for combining decisions in a building environment. Recall that verifiers may be purchased from multiple vendors and evaluated against different test databases. Furthermore, training is required for each possible path through a building, making this solution intractable in practice. The primary purpose of this approach is to evaluate the performance impact of estimating a joint probability distribution rather than assuming conditional independence.

5 Evaluation

We have evaluated the accuracy of decision fusion both with and without the conditional independence assumption using the score sets introduced in Sect. 3. For reference, we have also evaluated the accuracy of a Support Vector Machine (SVM) [2, 10] and a majority vote. The SVM operates on decisions that are represented as ± 1, and the majority vote accepts a subject as authentic when more than half the verifiers in a path accept that subject.

Noting that the most significant correlation occurs when combining two identical verifiers, we have constructed 9 example paths in which the first two steps use score sets from the same verifier and the second two steps use score sets from two distinct verifiers. The paths are enumerated in Table 2. The intent is to construct a scenario in which knowledge of the correlation between the first two verifiers will change the path decision at the fourth step.

Table 2. Example paths for evaluating fusion strategies

ID	Step 1	Step 2	Step 3	Step 4
1	EPFL	EPFL	USYD1	AUT1
2	EPFL	EPFL	USYD1	SURREY2
3	EPFL	EPFL	AUT1	SURREY2
4	AUT1	AUT1	USYD1	EPFL
5	AUT1	AUT1	USYD1	SURREY2
6	AUT1	AUT1	EPFL	SURREY2
7	USYD1	USYD1	EPFL	AUT1
8	USYD1	USYD1	EPFL	SURREY2
9	USYD1	USYD1	AUT1	SURREY2

Only two decisions are available for each authentic subject in the test set, so only two permutations are possible while maintaining realistic correlation (scenarios (a) and (b) in Fig. 1). We are thus forced to permute each of these pairs such that the first pair is completely independent of the second. We believe that this is a reasonable transformation given the observation that the correlation between distinct verifiers is relatively small to begin with.

To calculate local thresholds and estimate likelihoods, we have used training data from each score set (called the "evaluation set" in [7]). Verification accuracy is then calculated over the test set. In this paper, we present accuracy in terms of the Weighted Error Rate (WER) [11], which is simply a weighted average of FAR and FRR based on the security parameter W_{SEC} appearing previously in (2). The WER is defined as

$$\mathrm{WER} = \frac{W_{SEC} \cdot \mathrm{FAR} + \mathrm{FRR}}{W_{SEC} + 1}. \tag{5}$$

Error rates for path decisions at the final step of each path are presented in Fig. 3 with path numbers presented along the abscissa. For visual clarity, a line has been drawn to connect the error rates from each combination strategy. This line does not imply any logical connection from one path to the next.

Fig. 3 indicates that FB does, in fact, outperform NB for some paths; however, the difference in accuracy is negligible. In several paths (1, 7, 8, and 9) there is no difference in accuracy between the two algorithms. This occurs because the USDY1 algorithm performs significantly worse than the others. As a result, both fusion strategies give scores from USYD1 a small weight and favor the decisions of the second two verifiers. Correlation only serves to further reduce the weight

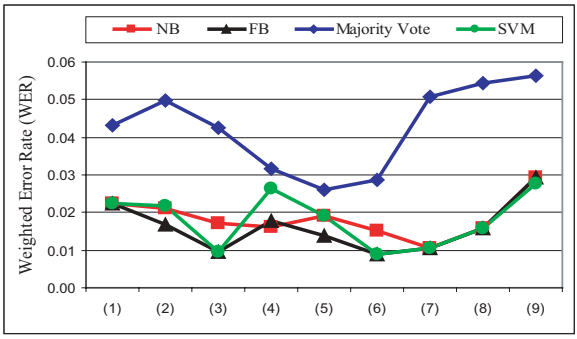


Fig. 3. Verification accuracy at the fourth step of example paths

of the USYD1 pair, so the second two verifiers continue to dominate. The remaining paths show an average difference in error rates of 0.5%. In these paths, the difference results from an increase in FAR and a corresponding decrease in FRR. Because one type of error is traded off for another, there is little change in the observed WER.

6 Discussion

In Sect. 5, we showed that even for significantly correlated decisions, the conditional independence assumption has a minimal impact on the weighted error rate. This result is rather surprising, but it can be explained based on two key observations: 1) only a small number of vectors are affected by the conditional independence assumption and 2) those vectors are ambiguous in the sense that they may have been generated by either authentic claimants or imposters.

For both the FB and NB strategies described in Sect. 4, the likelihood ratio test in (2) implicitly specifies a function mapping from local decision vectors $\mathbf{u}$ to a path decision u_0. We call this function the fusion rule and define $F_f(\mathbf{u})$ and $F_n(\mathbf{u})$ as the fusion rules specified by FB and NB, respectively.

For a given path, there may be decision vectors on which $F_f(\cdot)$ and $F_n(\cdot)$ disagree. We label the set of all such decision vectors D as defined in (6) and refer to D as the disagreeing set. For W_{SEC} set to 1 and similar error rates at each verifier, the vectors mostly likely to be in D are those ambiguous vectors containing the same number of accept and reject decisions. Because each individual verifier has a relatively low error rate, such vectors are unlikely to occur.

$$D = \{\, \mathbf{u} \,:\, F_f(\mathbf{u}) \neq F_n(\mathbf{u}) \,\} \tag{6}$$

To demonstrate the small size of D for W_{SEC} set to 1, we calculated the relative frequency of vectors in D for the paths from Table 2. These rates are presented in Fig. 4, which lists path numbers along the abscissa. As can be seen in the figure, vectors in D represent less than 2% of all vectors in both the training and test sets. Furthermore, we found that for the paths evaluated, D contained at most one unique vector – specifically the (accept, accept, reject, reject) vector.

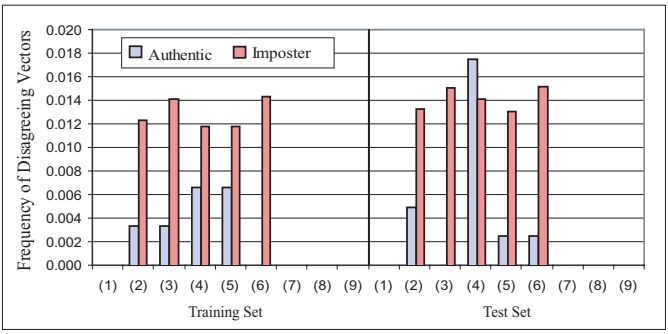


Fig. 4. Relative frequency of vectors in D

Fig. 4 also demonstrates that even with the joint characterization used by FB, the problem is not separable – making some number of errors unavoidable. For the correlation structure imposed in Sect. 5, we see that FB favors imposters for the (accept, accept, reject, reject) decision vector. We conclude that NB disagrees because it ignores the correlation between the first two decisions. Clearly, deciding to accept this vector was a poor choice for Path 4. Even for paths 2, 5, and 6, deciding to accept this vector will result in some number of false accept errors. These results are skewed by rather extreme the correlation scenario we have selected. In general, we expect any decision for vectors in D to lead to a tradeoff between a similar number of false accept and false reject errors.

The conclusion to be taken from Fig. 4 is that accounting for correlation between decisions minimally impacts verification accuracy. This is true because the conditional independence assumption only changes the fusion rule for ambiguous decision vectors (i.e., those that are generated by imposters and authentics with a similar likelihood). For relatively accurate verifiers, such ambiguous vectors are rare. The verifiers we evaluate here have a measured WER of between 0.06 and 0.13 ($W_{SEC} = 1$), and we expect such ambiguous decision vectors to become increasingly rare with improved verifier accuracy and longer paths.

7 Conclusion

In the preceding, we have defined two closely related strategies for combining decisions from multiple biometric verifiers. One (Naive Bayes) assumes conditional independence, while the other (Full Bayes) does not. Evaluating each against score sets from the XM2VTS database, we show that the conditional independence assumption does not significantly impact accuracy. Further analysis of the fusion rules generated by each strategy indicates that the decision vectors for which the two approaches disagree are both unlikely to occur and are generated by imposters and authentics with near equal likelihood. This trend can be attributed to the relatively high accuracy of each individual verifier. Based on this observation, we claim that decision fusion strategies can safely ignore moderate levels of correlation without significantly impacting accuracy.

References

1. Kittler, J., Messer, K.: Fusion of multiple experts in multimodal biometric personal identity verification systems. In: Proc. 12th IEEE Workshop on Neural Networks for Signal Processing. (2002) 3–12
2. Ben-Yacoub, S.: Multi-modal data fusion for person authentication using SVM. In: Proc. of the Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99). (1999) 25–30
3. Hong, L., Jain, A.: Integrating faces and fingerprints for personal identification. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 1295–1307
4. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters **24** (2003) 2115–2125
5. Beattie, M., Kumar, B.V.K. Vijaya, Lucey, S., Tonguz, O.: Building access control using coordinated biometric verification. In: Biometrics: Challenges arising from Theory to Practice (BCTP) Workshop. Proceedings. (2004)
6. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning **29** (1997) 103–130
7. Luettin, J., Maitre., G.: Evaluation protocol for the extended M2VTS database (XM2VTSDB). Technical Report IDIAP-COM 05, Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) (1998)
8. J. Matas et al: Comparison of face verification results on the XM2VTS database. In: 15th International Conference on Pattern Recognition, 2000. Proceedings. Volume 4. (2000) 858–863
9. L.I. Kuncheva, C.W.: Ten measures of diversity in classifier ensembles: limits for two classifiers. In: Intelligent Sensor Processing, A DERA/IEE Workshop on. (2001)
10. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm (2001)
11. E. Bailly-Bailliere et al: The BANCA database and evaluation protocol. In: Proc. 4th International Conference in Audio- and Video-Based Biometric Person Authentication (AVBPA). (2003)

Gait Recognition by Combining Classifiers Based on Environmental Contexts

Ju Han and Bir Bhanu

Center for Research in Intelligent Systems
University of California, Riverside, CA 92521, USA
{bhanu,jhan}@cris.ucr.edu

Abstract. Human gait properties can be affected by various environmental contexts such as walking surface and carrying objects. In this paper, we propose a novel approach for individual recognition by combining different gait classifiers with the knowledge of environmental contexts to improve the recognition performance. Different classifiers are designed to handle different environmental contexts, and context specific features are explored for context characterization. In the recognition procedure, we can determine the probability of environmental contexts in any probe sequence according to its context features, and apply the probabilistic classifier combination strategies for the recognition. Experimental results demonstrate the effectiveness of the proposed approach.

1 Introduction

Current image-based individual human recognition methods, such as fingerprints, face or iris biometric modalities, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods can not reliably recognize non-cooperating individuals at a distance in the real world under changing environmental conditions. Gait, which concerns recognizing individuals by the way they walk, is a relatively new biometric without these disadvantages. However, gait also has some limitations, it can be affected by clothing, shoes, or other environmental contexts. Moreover, special physical conditions such as injury can also change a person's walking style. The large gait variation of the same person under different conditions (intentionally or unintentionally) reduces the discriminating power of gait as a biometric and it may not be as unique as fingerprint or iris, but the inherent gait characteristic of an individual still makes it irreplaceable and useful in many visual surveillance applications.

In traditional biometric paradigms, individuals of interest are represented by their biometric examples in the gallery data. In general, the gallery examples are obtained under the similar environmental condition (context) and the number of examples for each individual is limited to one. This setup is good for strong biometrics such as iris and fingerprint, where the inherent discriminating features are abundance. Even if the context changes, there are still enough features to distinguish one individual from others.

This setup may not be appropriate for gait recognition. Human gait properties can be affected by various environmental contexts such as walking surface, carrying objects and environmental temperature. The change of an environmental context may introduce a large appearance change in the detected human silhouette, which may lead to

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 416–425, 2005.
© Springer-Verlag Berlin Heidelberg 2005

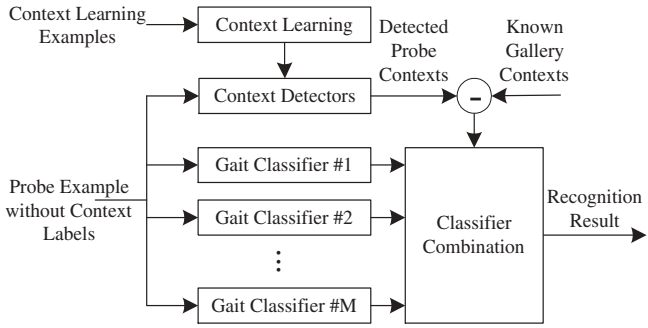


Fig. 1. Context-based classifier combination

a failure in recognition. The large gait variation of the same individual under different contexts requires more gallery examples of all individuals from all possible different environmental contexts. However, this requirement is unreal due to the complexity of real-world situations. Due to the difficulty of gait data acquisition, gait gallery examples are generally obtained under one or several environmental conditions and the number of examples for each individual is also very limited. Moreover, The environmental contexts are too rich in the real world to be entirely included in a gallery dataset.

Different gait recognition approaches (classifiers) character gait properties from different aspects. It is difficult to find a single classifier to effectively recognize individuals under all environmental contexts without gallery examples from these contexts. One classifier may be insensitive to the change of one context, while another classifier may be insensitive to the change of another context. If we can detect the environmental contexts of a given probe gait example, it is possible to combine these classifier to improve the recognition performance.

In this paper, we propose a context-based human recognition approach by probabilistically combining different gait classifiers under different environmental contexts. The basic idea is illustrated in Figure 1. First, context properties are learned from context training examples to construct context detectors. The contexts of a given probe gait examples are then obtained by these context detectors. Assuming that all gait gallery examples are obtained under the similar environmental contexts, the context changes between the probe example and gallery examples are obtained. With the gait classifiers designed for individual recognition under different environmental context changes, these classifiers are probabilistically combined to recognize the probe individual based on the detected context changes.

2 Related Work

In recent years, various approaches have been proposed for human recognition by gait. These approaches can be divided into two major categories: model-based approaches and model-free approaches.

Model-based gait recognition approaches focus on recovering a structural model of human motion. Niyogi and Adelson [1] find the bounding contours of the walker, and

fit a simplified stick model on them. A characteristic gait pattern in spatiotemporal volume is generated from the model parameters for recognition. Yoo et al. [2] estimate hip and knee angles from body contour by linear regression analysis. Then trigonometric-polynomial interpolant functions are fitted to the angle sequences, and the parameters so-obtained are used for recognition. Bhanu and Han [3] propose a kinematic-based approach to recognize individuals by gait. The 3D human walking parameters are estimated by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. Human gait signatures are generated by selecting features from the estimated parameters.

Model-free approaches make no attempt to recover a structural model of human motion. Little and Boyd [4] describe the shape of the human motion with a set of features derived from moments of a dense flow distribution. Shutler et al. [5] include velocity into the traditional moments to obtain the so-called velocity moments (VMs). BenAbdelkader et al. [6] use height, stride and cadence for to identify human. Sundaresan et al. [7] proposed a hidden Markov models (HMMs) based framework for individual recognition from their gait. Huang et al. [8] propose a template matching approach by combining transformation based on canonical analysis, with eigenspace transformation for feature selection. Similarly, Wang et al. [9] generate boundary distance vector from the original human silhouette contour as the template, which is used for gait recognition via eigenspace transformation. Phillips et al. [10] propose a direct template matching approach to measure the similarity between the gallery and probe sequences by computing the correlation of corresponding time-normalized frame pairs. Similarly, Collins et al. [11] first extract key frames from a sequence, and the similarity between two sequences is computed from the normalized correlation on key frames only. Tolliver and Collins [12] cluster human silhouettes of each training sequence into k prototypical shapes. Silhouettes in a testing sequence are also classified into k prototypical shapes that are used to compare with those in training sequences.

3 Technical Approach

In this section, we describe the proposed context-based classifier combination for individual recognition by gait. The context investigated in this paper is the walking surface type, but the approach could be extended to other contexts. The system diagram is shown in Figure 2.

3.1 Gait Representation

We assume that silhouettes have been extracted from original human walking sequences. A silhouette preprocessing procedure [10] is then applied on the extracted silhouette sequences. It includes size normalization (proportionally resizing each silhouette image so that all silhouettes have the same height) and horizontal alignment (centering the upper half silhouette part with respect to its horizontal centroid). In a preprocessed silhouette sequence, the time series signal of lower half silhouette size from each frame indicates the gait frequency and phase information. We estimate the gait frequency and phase by maximum entropy spectrum estimation [4] from the obtained time series signal.

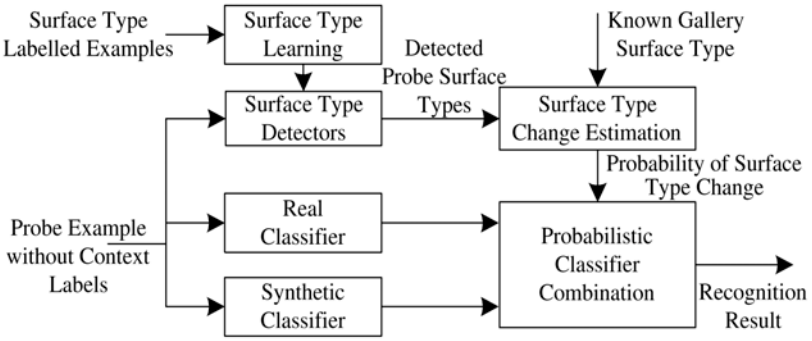


Fig. 2. Diagram of context-based classifier combination for individual recognition by gait. The context investigated in this diagram is the walking surface type

Given the preprocessed binary gait silhouette images $B_t(x, y)$ at time t in a sequence, the grey-level gait energy image (GEI) is defined as follows [13]

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B_t(x, y) \tag{1}$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number in the sequence (moment of time), x and y are values in the 2D image coordinate. It reflects major shapes of silhouettes and their changes over the gait cycle. We refer to it as gait energy image because: (a) each silhouette image is the space-normalized energy image of human walking at this moment; (b) GEI is the time-normalized accumulative energy image of human walking in the complete cycle(s); (c) a pixel with higher intensity value in GEI means that human walking occurs more frequently at this position (i.e., with higher energy). In comparison with binary silhouette sequence, GEI representation saves both storage space and computation time for recognition and is less sensitive to silhouette noise in individual frames. We use GEI as the gait representation for individual recognition in this paper.

3.2 Walking Surface Type Detection

Various environmental contexts have effect on silhouette appearance: clothing, shoes, walking surface, camera view, carrying object, time, etc. Among these contexts, slight camera view changes may be neglected. Irregular changes in clothing, shoe, carrying object and time generally cannot be detected. When the same person walks on different surface types, the detected silhouettes may have large difference in appearance. For example, silhouettes on the grass surface may miss the bottom part of feet, while silhouettes on the concrete surface may contain additional shadows. In these cases, silhouette size normalization errors occur, and silhouettes so-obtained may have different scales with respect to silhouettes on other surfaces. Figure 3 shows the GEI examples of three people walking on grass or concrete surfaces in USF HumanID database.

Considering the lower body part difference in silhouettes of people walking on grass and concrete surface, we use the silhouette energy in the lower body part as the indicator

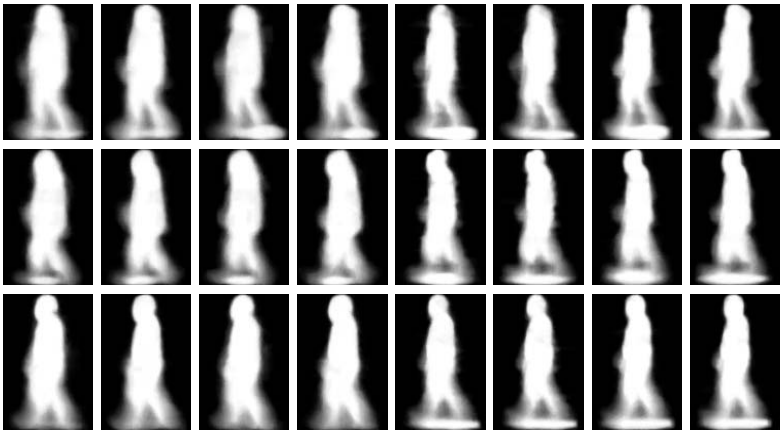


Fig. 3. GEI examples of three people (rows) waling on different surface types. First four examples in each row are on the grass surface, and the others are on the concrete surface

of the walking surface type. Let the bottom row be the first row and leftmost column be the first column in the image coordinate, the surface type indicator is defined as

$$s(G, N_{TOP}) = \frac{\sum_{i=1}^{N_{TOP}} \sum_{j=1}^{N_{COL}} G(i,j)}{\sum_{i=1}^{N_{ROW}} \sum_{j=1}^{N_{COL}} G(i,j)}, \tag{2}$$

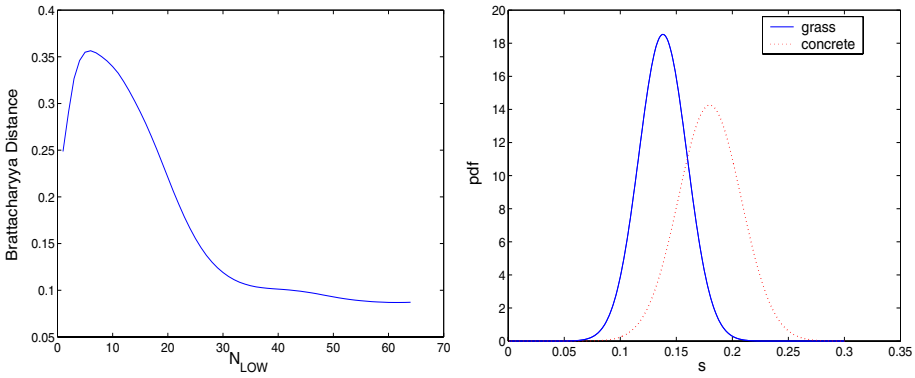
where G is a GEI example with the size of $N_{ROW} \times N_{COL}$, and N_{TOP} is the number of rows from the bottom. Assuming s has a Gaussian distribution for both grass GEI examples and concrete GEI examples, the class-conditional probability functions are estimated from the context training examples as follows

$$\begin{aligned} p(s|grass) &= \frac{1}{\sqrt{2\pi}\sigma_{grass}} \exp\left\{-\frac{(s-\mu_{grass})^2}{2\sigma_{grass}^2}\right\} \\ p(s|concrete) &= \frac{1}{\sqrt{2\pi}\sigma_{concrete}} \exp\left\{-\frac{(s-\mu_{concrete})^2}{2\sigma_{concrete}^2}\right\} \end{aligned} \tag{3}$$

where μ_{grass} and σ_{grass} are the sample mean and sample standard deviation of s for training examples on the grass surface, and $\mu_{concrete}$ and $\sigma_{concrete}$ are the sample mean and sample standard deviation of s for training examples on the concrete surface. These distributions are different for different N_{TOP} values. The optimal N_{TOP} for discriminating these two surface types is estimated by maximizing the Bhattacharyya distance with respect to N_{TOP}:

$$B = \frac{(\mu_{grass} - \mu_{concrete})^2}{4(\sigma_1^2 + \sigma_2^2)} + \frac{1}{2} \ln \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}. \tag{4}$$

The Bhattacharyya distance is used as a class seperability measure here. The Bhattacharyya distance of the two distribution with respect to different N_{TOP} values is shown in Figure 4(a). The estimated distribution for optimal $N_{TOP} = 6$ is shown in Figure 4(b).



(a) (b)

Fig. 4. (a) The Bhattacharyya distance of the two distribution with respect to different N_{TOP} values. (b) The estimated distributions of $p(s|grass)$ and $p(s|concrete)$ for $N_{TOP} = 6$

According to the Bayes rule, we have the following probabilities for probabilistic classifier combination

$$P(grass|s) = \frac{p(s|grass)P(grass)}{p(s)}$$
$$P(concrete|s) = \frac{p(s|concrete)P(concrete)}{p(s)}. \quad (5)$$

3.3 Classifier Design

In this paper, we use the real gait classifier for recognizing probe examples having no surface type change with respect to gallery examples, and synthetic gait classifier for recognizing probe examples having the surface type change [13].

The real GEI templates for an individual are directly computed from each cycle of the silhouette sequence of this individual. They are used as the input of real classifier for recognizing probe examples having no surface type change with respect to gallery examples.

A statistical feature extraction method by combining PCA and MDA is used for learning real gait features from training real templates. Let m_{ri} be the mean of real feature vectors belonging to the ith class (individual) in the gallery set. Given a probe example P, $\{R_j\}$, $j = 1, \ldots, n$, are its real gait templates. The corresponding real feature vector set is obtained as follows

$$\{\hat{R}_P\}: \quad \hat{\mathrm{r}}_j = T_r R_j, \quad j = 1, ..., n$$

where T_r is the learned transformation matrix for real feature extraction. The dissimilarity between the probe example and each gallery class is then measured by

$$D(\hat{R}_P, \omega_i) = \frac{1}{n}\sum_{j=1}^{n} ||\hat{\mathrm{r}}_j - \mathrm{m}_{ri}||, \quad i = 1, ..., c \quad (6)$$

where c is the number of classes in the gallery set. The real classifier is

$$\text{Decide} \quad P \in \omega_k \quad \text{if} \quad D(\hat{R}_P, \omega_k) = \min_{i=1}^{c} D(\hat{R}_P, \omega_i). \tag{7}$$

Although real gait templates provide cues for individual recognition, all the templates from the same sequence are obtained under the "same" physical conditions. If the condition changes, the learned features may not work well for recognition. Let R_0 be the GEI template computed all cycles of a given silhouette sequence. Assume that k bottom rows of R_0 are missed due to some kind of environmental conditions. According to the silhouette preprocessing procedure in Section 3.1, the remaining part needs to be proportionally resized to fit to the original height. In the same way, we can generate a series of new synthetic GEI templates corresponding to different lower body part distortion with the different values of k. The synthetic templates expanded from the same R_0 have the same global shape properties but different bottom parts and different scales. Therefore, they provide cues for individual recognition that are less sensitive to surface type changes.

A similar statistical feature extraction method by combining PCA and MDA is used for learning synthetic gait features from synthetic templates. Let m_{si} be the mean of synthetic feature vectors belonging to the ith class (individual) in the gallery set. Given a probe example P, $\{S_j\}$, $j = 1, \ldots, m$, are its synthetic gait templates. The corresponding synthetic feature vector set is obtained as follows

$$\{\hat{S}_P\}: \quad \hat{\mathrm{s}}_j = T_r S_j, \quad j = 1, ..., m$$

where T_s is the learned transformation matrix for synthetic feature extraction. The dissimilarity between the probe example and each gallery class is then measured by

$$D(\hat{S}_P, \omega_i) = \frac{1}{m} \sum_{j=1}^{m} ||\hat{\mathrm{s}}_j - \mathrm{m}_{si}||, \quad i = 1, ..., c \tag{8}$$

where c is the number of classes in the gallery set. The synthetic classifier is

$$\text{Decide} \quad P \in \omega_k \quad \text{if} \quad D(\hat{S}_P, \omega_k) = \min_{i=1}^{c} D(\hat{S}_P, \omega_i). \tag{9}$$

3.4 Probabilistic Classifier Combination

Given a probe example, the probabilities of different surface types are obtained in Equation (5). The dissimilarities of the probe example of each class in the gallery set are obtained in Equation (6) and (8), respectively. Notice that the real classifier is designed for recognizing probe examples having no surface type change with respect to gallery examples, and the synthetic gait classifier is designed for recognizing probe examples having the surface type change. If walking surface of gallery examples are grass, the combined dissimilarly is measures as follows

$$\begin{aligned} D(P, \omega_i) &= P(grass|s)\bar{D}(\hat{R}_P, \omega_i) + P(concrete|s)\bar{D}(\hat{S}_P, \omega_i) \\ &= P(grass|s)\frac{D(\hat{R}_P, \omega_i)}{\sum_{j=1}^{c} D(\hat{R}_P, \omega_j)} + P(concrete|s)\frac{D(\hat{S}_P, \omega_i)}{\sum_{j=1}^{c} D(hatS_P, \omega_j)} \end{aligned}$$

Table 1. Twelve experiments designed for human recognition in USF HumanID database

Experiment Label	Size of Probe Set	Difference between Gallery and Probe Sets
A	122	View
B	54	Shoe
C	54	View and Shoe
D	121	Surface
E	60	Surface and Shoe
F	121	Surface and View
G	60	Surface, Shoe and View
H	120	Briefcase
I	60	Shoe and Briefcase
J	120	View and Briefcase
K	33	Time, Shoe and Clothing
L	33	Surface and Time

for $i = 1, ..., c$, where $\bar{D}$ is the normalized dissimilarity. Assuming $P(grass) = P(concrete)$, we have

$$D(P, \omega_i) = P(s|grass)\frac{D(\hat{R}_P, \omega_i)}{\sum_{j=1}^{c} D(\hat{R}_P, \omega_j)} + P(s|concrete)\frac{D(\hat{S}_P, \omega_i)}{\sum_{j=1}^{c} D(\hat{S}_P, \omega_j)} \quad (10)$$

for $i = 1, ..., c$. The combined classifier based on surface context is

$$\text{Decide} \quad P \in \omega_k \quad \text{if} \quad D(P, \omega_k) = \min_{i=1}^{c} D(P, \omega_i). \quad (11)$$

4 Experimental Results

Our experiments are carried out on the USF HumanID gait database [10]. This database consists of people walking in elliptical paths in front of the camera. For each person, there are up to 5 covariates: viewpoints (left/right), shoe types (A/B), surface types (grass/concrete), carrying conditions (with/without a briefcase), and time and clothing. Twelve experiments are designed for individual recognition as shown in Table 1. The gallery set contains 122 sequences. Individuals are unique in the gallery and each probe set, and there are no common sequence among the gallery set and all probe sets. The walking surface type in the gallery set is grass.

Phillips et al. [10] propose a baseline approach to extract human silhouette and recognize people in this database. For comparison, they provide extracted silhouette data which can be found at the website http://marathon.csee.usf.edu/GaitBaseline/. Our experiments begin with these extracted binary silhouette data (version 2.1) that are updated on September 5, 2003. The performance of their baseline algorithm are shown in Table 2. In this table, rank1 means that only the first subject in the retrieval rank list is recognized as the same subject as the query subject, and rank5 means that the first five subjects are all recognized as the same subject as the query subject. The performance in the table is the recognition rate under these two definitions.

Table 2. Comparison of recognition performance among different approaches on silhouette sequence version 2.1 (Legends: baseline - USF baseline algorithm [10]; real - real classifier; synthetic - synthetic classifier; context - proposed context-based approach, this paper)

	Rank1 Performance				Rank5 Performance			
	baseline	real	synthetic	context	baseline	real	synthetic	context
A	73%	89%	84%	90%	88%	93%	93%	93%
B	78%	87%	93%	91%	93%	93%	96%	94%
C	48%	78%	67%	80%	78%	89%	93%	89%
D	32%	36%	53%	56%	66%	65%	75%	81%
E	22%	38%	55%	57%	55%	60%	71%	76%
F	17%	20%	30%	27%	42%	42%	53%	53%
G	17%	28%	34%	36%	38%	45%	53%	50%
H	61%	62%	47%	60%	85%	87%	79%	90%
I	57%	59%	57%	62%	78%	79%	81%	84%
J	36%	58%	40%	57%	62%	81%	65%	84%
K	3%	3%	21%	9%	3%	6%	33%	18%
L	3%	6%	24%	12%	15%	9%	42%	27%

We carry out experiments of human recognition by the real classifier, the synthetic classifier and the combined classifier based context according to rules in (7), (9), and (10), respectively. Table 2 shows the recognition performance of USF baseline algorithm and our proposed approaches. Note that the rank1 and rank5 performance of proposed classifiers is better than or equivalent to that of baseline algorithm on all experiments.

The performance of the synthetic classifier is significantly better than that of the real classifier on experiments D-G and L, where the surface type of probe examples is different from that of gallery examples. In other experiments where the surface type of probe examples is the same as that of gallery examples, the performance of the real classifiers is better on A, C and G-J, but a little worse on B and K. These results demonstrate the our designed real and synthetic classifiers is suitable for their desired contexts.

The combined classifier based on the surface context achieves better performance than individual real feature classifier and synthetic classifier in most experiments. It is shown that the combined classifier takes advantage of merits in individual classifiers based on the detected context information. In this paper, we only detect and use the specific context information about the walking surface type, and only design two classifiers for it. If we can detect or obtain more context information such as carrying objects, clothing and time, and design the corresponding classifiers, we expect further improved combination results.

5 Conclusions

In this paper, we propose a context-based human recognition approach by probabilistically combining different gait classifiers with different environmental contexts. First, context properties are learned from context training examples to construct context detectors. The contexts of a given probe gait examples are then obtained by these context

detectors. With the gait classifiers designed for individual recognition under different environmental context changes, these classifiers are probabilistically combined to recognize the probe individual based on the detected context changes. Experimental results show that the combined classifier takes advantage of merits in individual classifiers based on the detected context information.

References

1. Niyogi, S., Adelson, E.: Analyzing and recognizing walking figures in xyt. Proc. IEEE Conference on CVPR (1994) 469–474
2. Yoo, J.H., Nixon, M., Harris, C.: Model-driven statistical analysis of human gait motion. Proc. IEEE International Conference on Image Processing **1** (2002) 285–288
3. Bhanu, B., Han, J.: Individual recognition by kinematic-based gait analysis. Proc. International Conference on Pattern Recognition **3** (2002) 343–346
4. Little, J., Boyd, J.: Recognizing people by their gait: the shape of motion. Videre: Journal of Computer Vision Research **1** (1998) 1–32
5. Shutler, J., Nixon, M., Harris, C.: Statistical gait recognition via velocity moments. Proc. IEE Colloquium on Visual Biometrics (2000) 10/1–10/5
6. BenAbdelkader, C., Cutler, R., Davis, L.: Person identification using automatic height and stride estimation. Proc. International Conference on Pattern Recognition **4** (2002) 377–380
7. Sundaresan, A., RoyChowdhury, A., Chellappa, R.: A hidden Markov model based framework for recognition of humans from gait sequences. Proc. ICIP **2** (2003) 93–96
8. Huang, P., Harris, C., Nixon, M.: Recognizing humans by gait via parameteric canonical space. Artificial Intelligence in Engineering **13** (1999) 359–366
9. Wang, L., Hu, W., Tan, T.: A new attempt to gait-based human identification. Proc. International Conference on Pattern Recognition **1** (2002) 115–118
10. Phillips, P., Sarkar, S., Robledo, I., Grother, P., Bowyer, K.: The gait identification challenge problem: data sets and baseline algorithm. Proc. ICPR **1** (2002) 385–388
11. Collins, R., Gross, R., Shi, J.: Silhouette-based human identification from body shape and gait. Proc. IEEE Intl. Conf. on Automatic Face and Gesture Recognition (2002) 351–356
12. Tolliver, D., Collins, R.: Gait shape estimation for identification. Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 2688 (2003) 734–742
13. Han, J., Bhanu, B.: Statistical feature fusion for gait-based human recognition. Proc. IEEE Conf. on Computer Vision and Pattern Recognition **2** (2004) 842–847

Addressing the Vulnerabilities of Likelihood-Ratio-Based Face Verification

Krzysztof Kryszczuk and Andrzej Drygajlo

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne (EPFL)
{krzysztof.kryszczuk,andrzej.drygajlo}@epfl.ch

Abstract. Anti-spoofing protection of biometric systems is always a serious issue in real-life applications of an automatic personal verification system. Despite the fact that face image is the most common way of identifying persons and one of the most popular modalities in automatic biometric authentication, little attention has been given to the spoof resistance of face verification algorithms. In this paper, we discuss how a system based on DCT features with a likelihood-ratio-based classifier can be easily spoofed by adding white Gaussian noise to the test image. We propose a strategy to address this problem by measuring the quality of the test image and of the extracted features before making a verification decision.

1 Introduction

The goal of all automatic biometric verification systems is to reliably establish if the identity claim comes from the real claimant or from an impostor. Attempts to impersonate a selected person in order to gain privileges otherwise reserved for the rightful claimant, otherwise known as spoofing, have been not an unusual threat since personal identity verification became a necessity. Only quite recently systems that automatically compare voices, fingerprints, faces, irises and signatures, left the laboratories and met the challenges of the real world. One of those challenges is, and probably will remain, spoofing.

Moreover, more and more frequent are the attempts to store personal biometric information in a digital form and to embed this information in identity documents – like identity cards, passports, visas, company access cards, etc. One of the common biometric modality choices for those applications is face image. Digitally stored face images or templates are likely to soon accompany a traditional photograph, to allow both human and automated verification procedures.

The objective of the work presented in this paper is to show that an estimation of the quality of the test image is necessary to assure the robustness of a face verification system to spoofing. As defined in [8,10], a complete face verification system consists of modules that perform: 1) *localization*, 2) *normalization*, 3) *feature extraction*, 4) *classification*. We postulate to add an additional step before classification: *quality assessment*.

We show that omitting the quality assessment step may actually compromise the impermeability of an automated face verification system to imposters. Using an example of the local Discrete Cosine Transform (DCT)-feature based system with likelihood-ratio-based classifier, we show how an acceptable set of features can originate

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 426–435, 2005.
© Springer-Verlag Berlin Heidelberg 2005

from an alien signal (noise), which can successfully spoof a face verification system. Consequently, we propose to put additional constraints on the input signal in order to prevent such non-eligible accesses.

The paper is organized as follows: Section 2 gives an overview of features used for face verification. Section 3 focuses on face verification based on DCTmod2 features and Gaussian Mixture Model (GMM) classifier. Section 5 deals on how a discussed face verification system can produce unreliable verification decisions. Section 6 proposes two complementary quality assessment methods and their combination. Conclusions and future work prospects are found in Sections 7 and 8.

2 Feature Extraction Routines for Face Images

The most popular features for face recognition from 2D images can be divided into holistic and local [8,11,12]. The holistic features are probably in widest use. However, their recognition accuracy suffers from scaling, rotation and translation of the input signal [8,9].

Another group of feature extraction techniques is made of algorithms that divide the input image into segments and extract features from those segments independently, producing a set of feature vectors. Subject literature reports the use of local PCA [10], Gabor wavelets [8] and 2-dimensional DCT-based features and their derivates [4,5,8,9,10]. Local features reportedly suffer less than their holistic counterparts from incorrect geometrical normalization of the input signal, which manifests itself in good performance of modified DCT-based features, particularly in the tests involving automatically localized faces [6]. For this reason, we have chosen the local feature extraction approach, and a GMM-based classifier, as a testbed for our experiments.

To overcome the disadvantages of using only local or global feature extraction schemes alone, successful attempts have been made to create hybrid systems that use both approaches [4]. Although the overall performance of those systems is reported to be superior in comparison with non-hybrid approaches, they are also bound to suffer from attacks which would confuse one feature extraction scheme, leaving only the second one in operation.

3 DCTmod2-GMM Face Verification

In our experiments we used a face verification scheme implemented in similar fashion as presented in [5,8,10]. Images from BANCA database [2] (French part) were used to build the world model (520 images, 52 individuals, 338 Gaussians in the mixture), while images from BANCA (English part) database were used to build client models using a recursive adaptation of the world model, as described in [7]. The adaptation relevance parameter was set to 16, and the number of iterations was set to 10. The images used in the experiments were cropped, normalized and rescaled to the size of 64×64 pixels. All faces were localized manually and normalized geometrically (eye position). Mean pixel intensity subtraction was used as the data normalization procedure before feature extraction. More sophisticated normalization schemes grant slightly better verification performance [4], but minute gains in performance was not the objective of the experiments reported here.

To verify a claim that a given test image belongs to the client C, a set of feature vectors, X, is extracted from the image. The verification decision is based on the likelihood ratio:

$$LR(X) = \frac{L(X \mid \lambda_C)}{L(X \mid \lambda_W)} \tag{1}$$

where $L(X|\lambda_C)$ and $L(X|\lambda_W)$ are the joint likelihoods of the set of vectors X given λ_C (the model of client C), and λ_W (the *world model*) [8]. The value of $LR(X)$ is compared to a threshold Θ, whose value is computed depending on the desired properties of the verification system [2].

Making a decision based on the likelihood ratio was proved to be an optimal strategy for biometric verification based on fixed-length feature vectors [1]. This holds assuming that both λ_C and λ_W were built using representative data sample from their respective populations. In the case of face verification it means that λ_C and λ_W have to account for every possible condition and degree of quality of the input face image. In the case of the performance estimation based on a standard evaluation protocol (e.g. XM2VTS, BANCA) this condition is met. It may not be the case in a real-world application where there is no closed set of images that can appear as an identity claim.

If a significant mismatch exists between the quality of the test image and the quality of the images employed in the training of λ_C and λ_W, using the likelihood ratio stops to be a meaningful way of making reliable verification decisions. In a classical verification scheme, the only possible outcomes of the decision process are acceptance or rejection of the hypothesis that the claimant is who he claims to be. An image, whose quality does not match at all the quality of images used to train the models, cannot be correctly represented by those models. Therefore one could expect that upon encountering such an image, the system will reject the claim. In the likelihood ratio scheme though, if the world model explains the incoming data from the claimant to an even smaller extent, the decision of the system will be positive, which is an obviously meaningless result. We show that such situation is possible and quite likely in a real application.

4 Tricking a DCTmod2-GMM System

DCT-based local features capture predominantly higher spatial frequency in the image [4]. Therefore, in order to depart from image quality comparable (by means of the DCTmod2 features) to the quality of images used during the training of λ_C and λ_W, we corrupt the test images with white Gaussian noise. Such noise contamination introduces alien spatial frequencies to the image, and since the mean image intensity remains unchanged, the energy distribution between frequencies in the image alters.

We choose noise contamination as the way to depart from the initial image quality conditions because it is a likely factor to corrupt images in real life. The corruption was followed by normalization identical to that performed on the images used for the training of world and client models. Example images with different level of added noise can be found in Figure 1. Percentages of noise contamination of images are equivalent to noise-to-signal ratio (reciprocal of SNR).

Corrupted versions of face images have been prepared for all images from Sets 02, 03 and 04 of the English part of the BANCA database (total of 1560 images). Following tests were performed:

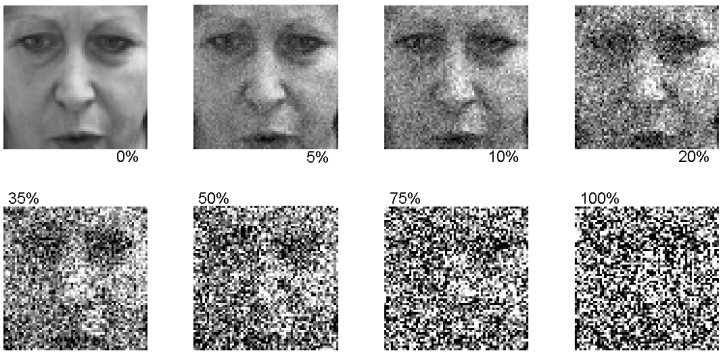


Fig. 1. Example face image from BANCA database (English part), corrupted with additive white Gaussian noise

- *Genuine client access tests*. Images corrupted with various amount of noise were tested using corresponding client models (corrupted images of client 1 against the model of client 1, etc.).
- *Impostor attack tests*. Images corrupted with various amount of noise were tested using client models created for another client. For simplicity, images coming from client n were tested against the model of client $n+1$. Face images of client 52 were attempting to impersonate client 1.

Genuine and impostor access attempts were scored using likelihood ratio $LR(X)$, as discussed in Section 3. The scores in terms of $LR(X)$ are presented in Figure 2. Gaussian approximations of their distributions are shown in Figure 3. For all tested images, $L(X|\lambda_C)$ was plotted against $L(X|\lambda_W)$ in Figure 4.

The influence of noise contamination on the likelihood scores is evident in Figure 4. For every X, the addition of noise causes a significant decrease of both $L(X|\lambda_C)$ and $L(X|\lambda_W)$, suggesting that the feature set X originating from the input image cannot be represented by neither the client, nor the world model. This information, however, is lost when $LR(X)$ is calculated (Figures 2 and 3). In this situation the verification system is bound to be confused.

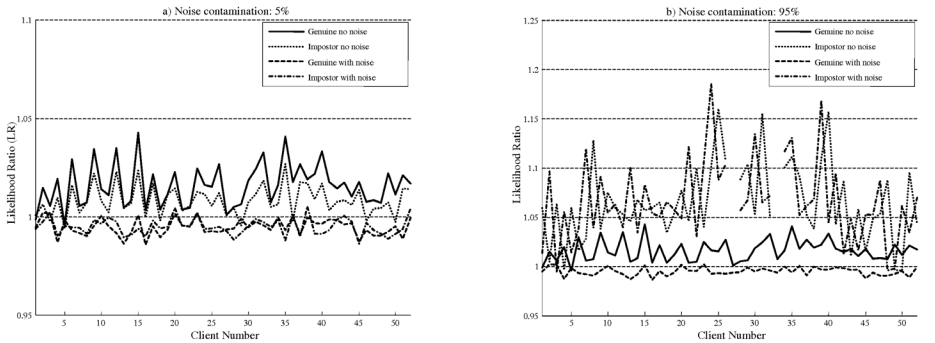


Fig. 2. Likelihood ratio scores for the verification tests on images from BANCA, English part, Session 03, for noise contamination 0% (no noise), and 5% and 95%

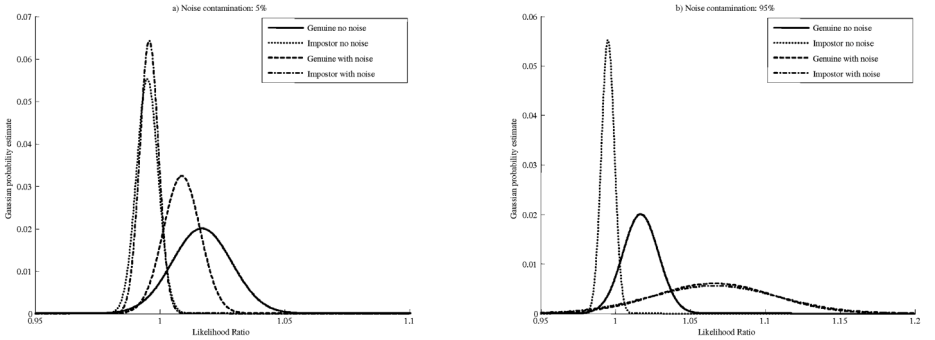


Fig. 3. Distributions of likelihood ratio scores for 0% (no noise), 5% and 95% of noise contamination

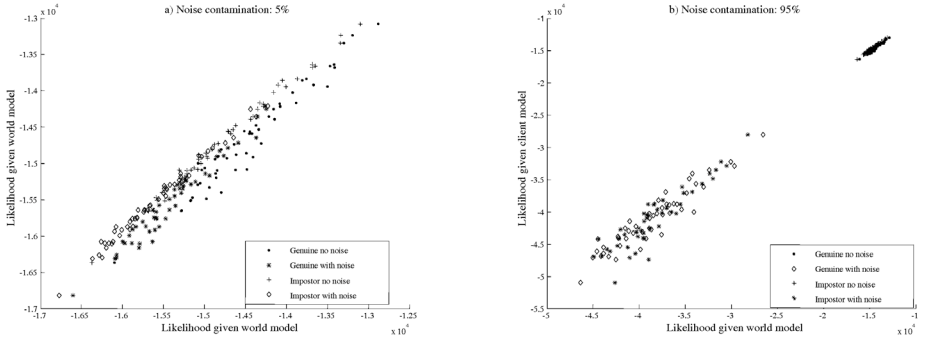


Fig. 4. Likelihood scores $L(X|\lambda_C)$ plotted against likelihood scores $L(X|\lambda_W)$

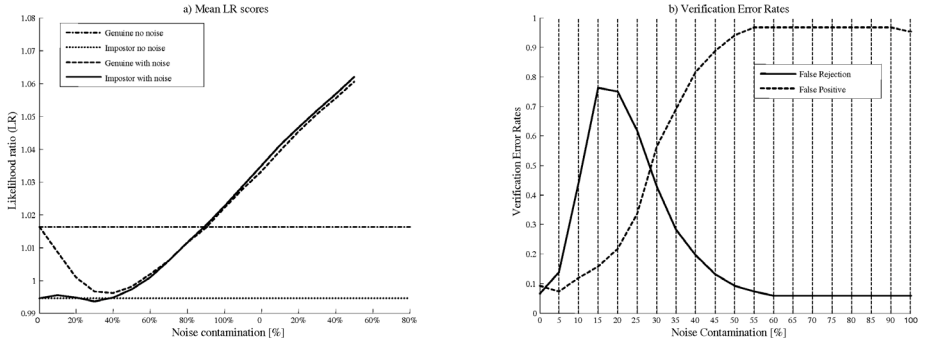


Fig. 5. a) Means of the score distributions of noise-contaminated genuine client and impostor claims. Scores for noise-free images left as a frame of reference, b) Mean verification results (BANCA, English, Sessions 02,03,04), as a function of noise contamination

The plots in Figure 5a represent the change of the mean distance between the genuine client and impostor likelihood ratio distributions, as a function of the noise contamination of face images.

As the presented results reveal, the automatic face verification system tends to reject impostors when the noise contamination is not significant. At those levels of

noise present in the input face images the average scores for real clients sink as well – a reasonable and desirable behavior which could be expected. For the same noise percentage, the impostor scores remain relatively constant. As the noise contamination of the input face image becomes large (above 25%) both genuine client and impostor scores grow rapidly, level up with the mean score for noise-free genuine clients at about 50% noise content, and continue growing. Figure 5b shows the verification error rates as a function of noise contamination (threshold Θ=1). Above 30% of noise contamination the system begins to favor acceptances over rejections, and above 50% of noise almost every claim is accepted.

5 Image Quality Assessment

Accepting every claim above certain level of image quality degradation is definitely an unacceptable behavior. Upon inspection of Figures 1 and 5b, it appears that the confused behavior of the system begins when the noise contamination begins to occlude the important facial features and the image bears less and less resemblance to a face. In order to address this vulnerability, it is necessary to introduce an intermediate step, which will automatically assess the quality of the input image. The goal of such assessment is:

1. To tell if the image presented to the system is indeed an image of a face.
2. To give a measure of the quality of the input image, relative to the quality of images used in the training of the system.

In order to meet those requirements, we consider two alternative approaches:

1. Quality assessment in the likelihood score domain.
2. Quality assessment independent of the features considered for verification.

5.1 Quality Assessment in the Likelihood Score Domain

The concept of likelihood-based verification, as expressed by Equation (1), is to find out if the feature vector is better represented by λ_C or by λ_W. This measure does not account for a situation when neither of the models represents the data adequately. We propose to compute a measure of how much the quality of the input matches either of the two models, or both simultaneously. For given feature set X originating from an image I we define the quality measure Q:

$$Q(I) = L(X \mid \lambda_C) + L(X \mid \lambda_W)\,. \tag{2}$$

The distribution of Q for N images I_T used in training of models λ_C and λ_W can be approximated using a mixture of 3 Gaussians, following identical model training procedure as during the training of λ_W. The distribution and the resulting model λ_Q are shown in Figure 6a. For given test image I we calculate its relative quality measure as:

$$R(I) = N \frac{L(Q(I) \mid \lambda_Q))}{\sum_{i=1}^{N} L(Q(I_T^i) \mid \lambda_Q)} \tag{3}$$

For every level of noise contamination of the n test images we calculate their corresponding mean relative quality measure $R_{mean}=(1/n)\cdot\sum R(I)$. Figure 6b shows R_{mean} as the function of the level of noise contamination.

The curve presented in Figure 6b is descending quickly from high relative quality values for clean and little noise-contaminated images, to arrive at values near zero for test images contaminated with more than 10% of white Gaussian noise.

The estimate is hence very sensitive to the degradation of the input image quality. At the same time, however, it depends heavily on the training conditions of λ_C and λ_W. Also, it really says nothing if the input image I is indeed a face image.

5.2 Quality Assessment Independent of the DCTmod2 Features

Upon inspection of Figure 1 one can conclude that gradual degradation makes first the individual facial features difficult to recognize, then even the rudimentary features stop to be obvious, until the image ceases to resemble a face at all. Since image quality should not be individual-dependent, it is desirable to have a measure of "face-likeness", in other words to estimate how much the input image resembles a face at all.

For this purpose, we propose to use normalized correlation of the input image with an average face template. We build the average face template T_F out of the same image set that was used before to build the world model λ_W, as described in [3]. The template can be seen in Figure 7a.

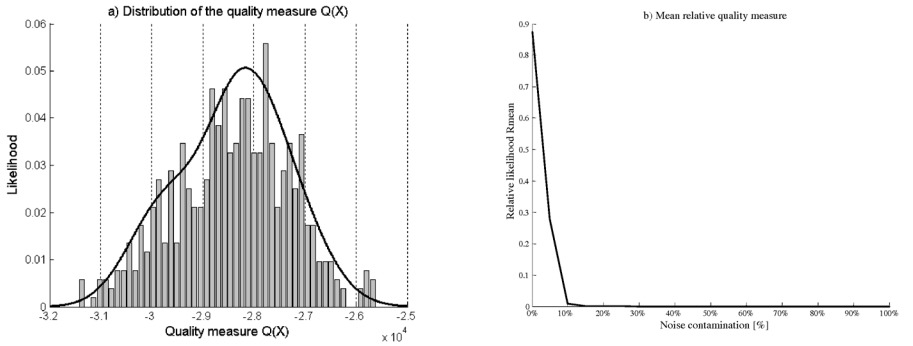


Fig. 6. a) Distribution of $Q(I)$ and its corresponding GMM, b) Relative mean quality measure R_{mean} as a function of noise contamination

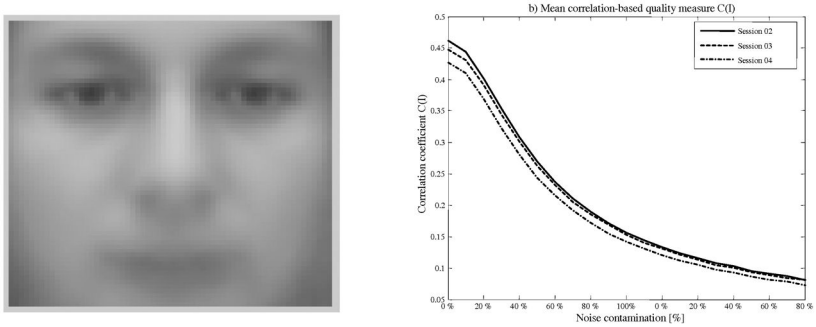


Fig. 7. Average face template T_F and b) mean correlation-based quality scores for images from BANCA (English), Sessions 02, 03 and 04

Since the average face template is a smooth reconstruction from first 8 principal components, it preserves only the facial features that are common to all faces from the training set. This makes it a good frame of reference to assess the "face-likeness" of an image. For given test image I we define its degree of resemblance to a face as:

$$C(I) = \max(corr(I, T_F)), \tag{4}$$

where $corr(I,T_F)$ is a normalized 2D correlation of T_F and I. Figure 7b shows how $C(I)$ changes as the function of noise contamination of the face image. The correlation-based quality measure gives a very good estimate if the input image indeed is a face image, independently of the features extracted for verification purposes.

Proposed correlation-based measure of "face-likeness" is one of the methods used in face detection [3]. Face detection, in general, is a way of assessing how much given object resembles a face. Therefore, in theory any face detection algorithm at some point does calculate some measure of "face-likeness" and this information can be used during the quality assessment step.

5.3 Combining Quality Measures for Increased Robustness

The relative quality measures $R(I)$ and $C(I)$ have complementary strengths and weaknesses. While $R(I)$ is more sensitive to the degradation of I in terms of features used for verification, $C(I)$ is providing information about how likely it is that I is an image of a face. Since both measures are computed independently, and it is required that an image of the rightful claimant is both an image of a face and that its quality is compatible with λ_C and λ_W, we define the combined quality measure $M(I)$ *a*s:

$$M(I) = R(I) \cdot C(I) \tag{5}$$

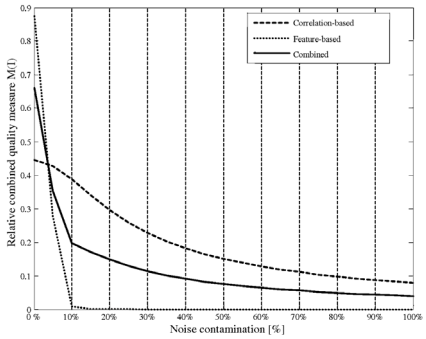


Fig. 8. Correlation-based, feature-based and combined relative quality measure of input face image, as a function of the percentage of noise contamination

Figure 8 presents the $M(I)$ as the function of the percentage of noise contamination.

Let's introduce a threshold Θ_R. For given test image I, if $M(I) < \Theta_R$, the quality assessment module rejects the image on the basis of its insufficient quality relative to the images used for the training of the verification system. The choice of Θ_R depends on the desired properties of the system. For example, if an increase of false acceptances is not desired, by comparison of curves in Figures 5b and 8, a threshold Θ_R=0.2 would be appropriate.

6 Summary and Conclusions

In this paper we have shown that a DCT-based face verification systems that uses likelihood-ratio-based classifier, can be vulnerable to spoofing attacks using face images contaminated with white Gaussian noise. We presented a method of obtaining an automatic assessment of the quality of face images which can help in preventing such attacks. The quality assessment method uses a combined measure that takes into account both the compatibility of the input image with world and client models, and the "face-likeness" of the image. The latter is particularly necessary in systems based on local features modeled with GMMs, since the spatial relations between facial features are not preserved in the models.

7 Future Work

We presented a combined quality assessment scheme for face images. The hybrid approach of this method gives a good global estimate of the quality of the image on the input of a face verification system. This estimate, however, gives no information as to why the quality is deteriorated, relative to the reference images. We are currently developing a set of techniques that allow a precise estimation of various quality measures of face images (localization, lighting, sharpness, etc.).

Acknowledgements

This work was supported in part by the Swiss National Science Foundation (SNSF) through the National Network of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.

References

1. A.M. Bazen and R.N.J.Veldhuis.:Likelihood-Ratio-Based Biometric Verification. In: IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, No.1, January 2004.
2. S. Bengio, F. Bimbot, J. Mariethoz, V. Popovici, F. Por´ee, E. Bailly-Balliere, G. Matas and B. Ruiz.: Experimental protocol on the BANCA database. Technical Report IDIAP-RR 02-05, IDIAP, 2002. (www.idiap.ch)
3. K. Kryszczuk and A. Drygajlo.: Color Correction For Face Detection Based on Human Visual Perception Metaphor. In: Proc. of the Workshop on Multimodal User Authentication, p. 138-143, Santa Barbara, CA USA, 2003.
4. S. Lucey.: The Symbiotic Relationship of Parts and Monolythic Face Representations in Verification. In: International Workshop on Face Processing in Video (FPIV), Washington D.C., 2004.
5. S. Lucey and T. Chen.: A GMM Parts Based Face Representation for Improved Verification through Relevance Adaptation. In: Proc. of the IEEE CSS Conf. on Computer Vision and Pattern Recognition, Vol. 2, pp. 855-861, Washington, USA, 2004.
6. K. Messer, J. Kittler, M. Sadeghi, M. Hamouz, A. Kostyn, S. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, N. Poh, Y. Rodriguez, K. Kryszczuk, J. Czyz, L. Vandendorpe, J. Ng, H. Cheung, and B. Tang.: Face authentication competition on the BANCA database. In: Proc. of the International Conference on Biometric Authentication, ICBA, Hong Kong, 2004.

7. D.A. Reynolds, T.F. Quatieri and R.B. Dunn.: Speaker Verification Using Adapted Gaussian Mixture Models. In: Digital Signal Processing, Vol. 10, 19-41, 2000.
8. C. Sanderson.: Automatic Person Verification Using Speech and Face Information. PhD Thesis, Griffith University, Australia, August 2002 (revised Feburary 2003).
9. C. Sanderson and S. Bengio.: Robust Features for Frontal Face Authentication in Difficult Image Conditions. Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA), Guildford, UK, 2003.
10. M. Saban and C. Sanderson.: On Local Features for Face Verificaton. Technical Report IDIAP-RR 04-36, IDIAP, 2004. (www.idiap.ch)
11. M. A. Turk and A. P. Pentland.: Eigenfaces for recognition. In: Journal of Cognitive Neuroscience, 3(1), pp. 71–86, 1991.
12. W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips.: Face Recognition: A Literature Survey. UMD CfAR Technical Report CAR-TR948, 2000.

Practical Biometric Authentication with Template Protection

Pim Tuyls[1], Anton H.M. Akkermans[1], Tom A.M. Kevenaar[1], Geert-Jan Schrijen[1], Asker M. Bazen[2], and Raymond N.J. Veldhuis[2]

[1] Philips Research Laboratories
Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands
{pim.tuyls,ton.h.akkermans,tom.kevenaar,geert.jan.schrijen}@philips.com
[2] University of Twente - EEMCS - SAS
P.O. box 217, 7500 AE Enschede, The Netherlands
{a.m.bazen,r.n.j.veldhuis}@utwente.nl

Abstract. In this paper we show the feasibility of template protecting biometric authentication systems. In particular, we apply template protection schemes to fingerprint data. Therefore we first make a fixed length representation of the fingerprint data by applying Gabor filtering. Next we introduce the *reliable components* scheme. In order to make a binary representation of the fingerprint images we extract and then quantize during the enrollment phase the reliable components with the highest signal to noise ratio. Finally, error correction coding is applied to the binary representation. It is shown that the scheme achieves an EER of approximately 4.2% with secret length of 40 bits in experiments.

1 Introduction

Biometrics identify/authenticate people on what they are rather than on what they have (tokens) or what they know (passwords). Since biometric properties can not be lost or forgotten in contrast to tokens and passwords, they offer an attractive and convenient alternative to identify and authenticate people.

When the reference information, captured during the enrollment phase, is not properly protected some privacy problems arise. The main risks are given by: i) Biometrics contain sensitive information about people [Bolling, P65]. ii) Once compromised, the templates are compromised forever and can not be reissued [S99]. iii) Biometric data stored without protection can be used to perform cross-matching between databases and track peoples behaviour. iv) Many biometric identifiers can be forged based on template information [MMJ03]. This problem received recently a lot of attention [JS02, TG04, LT03, DRS04, JW99, Sou98].

Two equivalent approaches, Helper Data and Fuzzy Extractors, were proposed to solve this privacy problem [TG04, DRS04]. In these papers the theory of template protection has been developed and some algorithms were proposed. In [TVI04] these algorithms were applied to ear identification and a satisfactory performance (EER=3%, secret length=100) was achieved.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 436–446, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we present an implementation of template protection for fingerprint based authentication. We present an algorithm based on helper data consisting of two parts. The first part identifies the *reliable* components with a high signal to noise ratio in the analog picture of a Gabor-filtered fingerprint. By applying quantization, a binary representation is made of the fingerprint. The second part of the helper data maps the binary representation onto a code word of an error-correcting code which is further used to correct the noise remaining after quantization.

2 Preliminaries

2.1 Biometric Verification

The biometric system that is considered in this paper is a *verification system*. As usual it consists of two phases. In the enrollment phase (executed at a Certification Authority (CA)), reference measurements are taken, the features are extracted, and the template is stored in e.g. a database in a properly protected way. During the verification phase, a live biometric measurement is compared to the template that is retrieved from the database using a claimed identity. Due to noise (caused by scratches, weather conditions, partial impressions, elastic deformations, etc.) the measurements taken during the enrollment and verification phase are different. This degrades the performance of a biometric verification system. In order to measure the performance, two different error rates are commonly used. The False Acceptance Rate (FAR) is the probability that an impostor is falsely accepted as a genuine user. The False Rejection Rate (FRR) is the probability that a genuine user is falsely rejected by the system. The Equal Error Rate (EER) is the error rate at the point of operation where FAR is equal to FRR.

2.2 Template Protection

Biometric data (and their extracted features) are modeled as k-dimensional random variables with entries in $\mathbb{R}$. The extracted features during the enrollment phase are denoted by $\boldsymbol{X}$ and those extracted during the verification phase by $\boldsymbol{X'}$. The data during the verification phase are modeled as a noisy version from those measured during the enrollment phase [TG04].

The core algorithm of a template protecting biometric system extracts a *secret* from the biometric data. Generally speaking such an algorithm is built on a Secret Extraction Code [TG04] or equivalently a Fuzzy Extractor [DRS04]. For the sake of simplicity we describe the algorithm in terms of a shielding function [LT03], which generates a special set of secret extraction codes [TG04] but has all necessary properties. A shielding function $G : \mathbb{R}^k \times \{0,1\}^k \to \{0,1\}^K$ extracts a secret of length K from the biometric as follows. Given a randomly chosen secret $S \in \{0,1\}^K$ and a biometric $\boldsymbol{X} \in \mathbb{R}^k$, *helper data* $W \in \{0,1\}^k$ is computed such that $G(X,W) = S$ (equivalently the equation $G(X,W) = S$ is solved for W). A shielding function is called δ-contracting if for all $\boldsymbol{X'}$ that lie

within a ball of radius δ of $\boldsymbol{X}$ we have $G(\boldsymbol{X}', W) = G(\boldsymbol{X}, W) = S$. The function G is called ϵ-revealing if the helper data W leaks less than ϵ bits on S (in the information theoretic sense), i.e. $\mathbf{I}(W; S) \leq \epsilon$. It is the goal to design the system such that W leaks also a minimal amount of information on $\boldsymbol{X}$; i.e. $\mathbf{I}(W; \boldsymbol{X})$ has to be minimized. It was shown in [LT03] that for a shielding function G, $\mathbf{I}(W; \boldsymbol{X})$ can not be made equal to zero.

During the *enrollment* phase the features $\boldsymbol{X}$ of Alice's biometric are extracted, a secret S is randomly choosen and the helper data W is computed. Then, a one-way hash function H is applied to S and the data $(Alice, W, H(S))$ is stored in a database.

During the *verification* phase, (at the sensor) a noisy version $\boldsymbol{X}'$ of Alice's biometric $\boldsymbol{X}$ is measured. When Alice claims her identity the helper data W is passed onto the sensor. The sensor computes $S' = G(\boldsymbol{X}', W)$ and $H(S')$. At the database $H(S')$ is compared to $H(S)$. If both are equal access is granted and if they are unequal no access is granted. Note that in contrast to usual practice in biometrics ("fuzzy matching") an exact match is performed. We stress that the helper data is sent over a public channel, i.e. W can be captured by an attacker. The system is however designed such that the knowledge of W provides a minimal amount of information on $\boldsymbol{X}$ and S [LT03, TG04, DRS04]. For basic examples of template protecting biometric verification systems, we refer to [TG04, DRS04].

2.3 Fingerprint Feature Extraction

In this section we present a fixed length feature vector representation, of which the elements can be compared one by one directly. The selected feature vector describes the global shape of the fingerprint by means of the local orientations of the ridge lines.

In order to allow for direct comparison of the feature vectors, without requiring a registration stage during matching, the feature vectors have to be prealigned during feature extraction. For this purpose, the core point (i.e. the uppermost point of the innermost curving ridge) is used. These core points are automatically extracted using a likelihood ratio-based algorithm that is described in [Baz04].

To describe the shape of the fingerprint, we extract two types of feature vectors from the gray scale fingerprint images. The first feature vector is the squared directional field that is defined in [Baz02]. It is evaluated at a regular grid of 16 by 16 points with spacings of 8 pixels, which is centered at the core point. At each of the 256 positions, the squared directional field is coded in a vector of two elements, representing the x- and y-values, resulting in a 512-dimensional feature vector. An example fingerprint and its directional field are shown in Figures 1a and 1b respectively.

The second feature vector is the Gabor response of the fingerprint, which is discussed in [BV04]. After substraction of the spatial local mean, the fingerprint image is filtered by a set of four complex Gabor filters, which are given by:

$$h_{\text{Gabor}}(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \cdot \exp\left(j2\pi f \cdot (x \sin\theta + y \cos\theta)\right) \qquad (1)$$

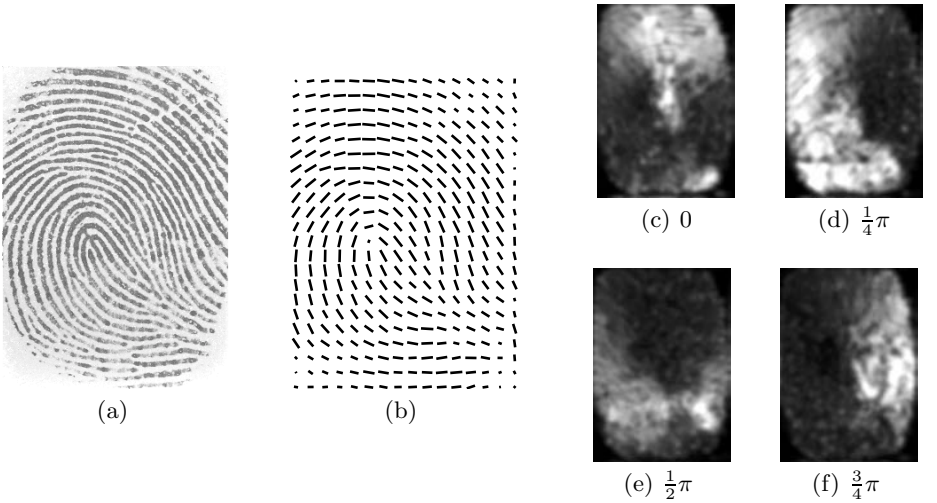


Fig. 1. (a) Fingerprint image, (b) its directional field and (c)-(f) the smoothened absolute values of Gabor responses for different orientations θ

The orientations θ are set to 0, $\pi/4$, $\pi/2$, and $3\pi/4$, the spatial frequency f is tuned to the average spatial ridge-valley frequency ($f = 0.11$), and the width of the filter σ is set such that the entire orientation range is covered ($\sigma = 3.5$). The absolute values of the output images are taken, which are subsequently filtered by a low-pass Gaussian window. The resulting images are shown in Figures 1c to 1f.

Again, samples are taken at a regular grid of 16 by 16 points with spacings of 8 pixels and centered at the core point. The resulting feature vector is of length 1024. This feature vector is inspired by FingerCode [Jai00], but it can be calculated more efficiently since a rectangular grid is used rather than a circular one, and it performs better.

The resulting feature vector that is used for matching is a concatenation of the squared directional field and the Gabor response. It describes the global shape of the fingerprint in 1536 elements.

3 Integration of Template Protection with Fingerprint Verification

From each user we use M measurements of his/her biometric for enrollment. The enrollment phase comprises five steps: *Feature Extraction*, *Statistical Analysis*, *Quantization*, *Selecting Reliable Components* and *Creating Helper Data*. These steps are described in detail in Section 3.1.

In the verification phase the biometric of a user is measured. Then, feature extraction and quantization are performed and using the helper data the noise is removed and the secret reconstructed. The details are explained in section 3.2. The complete scheme is shown in Fig. 2.

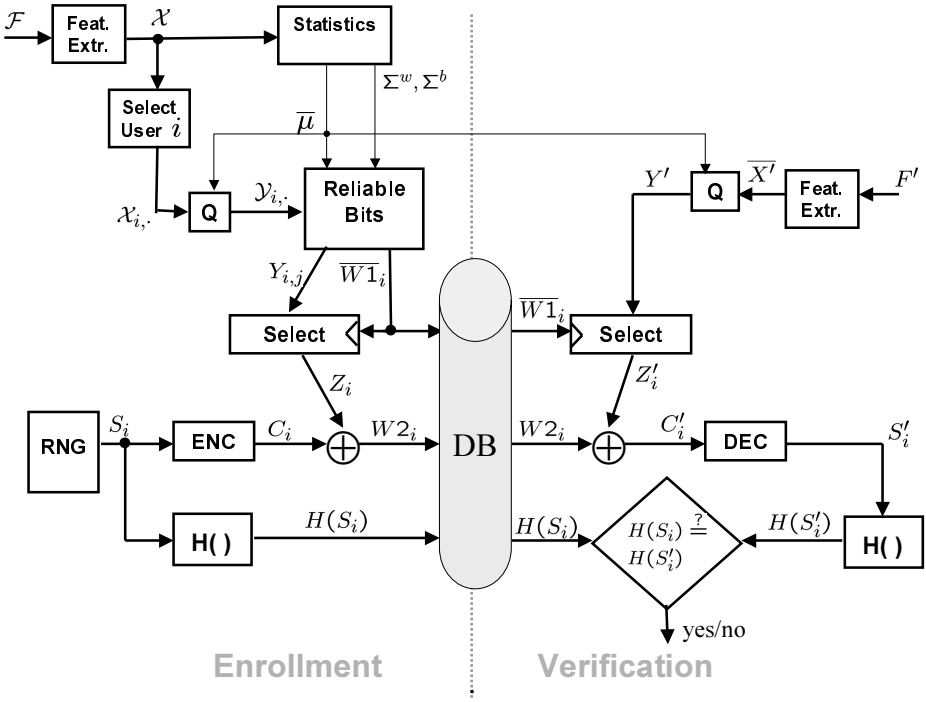


Fig. 2. Overview of the reliable components scheme

3.1 Enrollment

Feature Extraction. At the input of the scheme, we consider a set of biometric enrollment measurements $\mathcal{F} = \{F_{i,j}\}_{i=1..N,j=1..M}$ where a subscript i, j denotes the j-th enrollment measurement of the i-th user. Thus N is the number of users and M the number of enrollment measurements per user such that $\mathcal{F}$ consists of NM digital images of fingerprints. In the Feature Extraction block (depicted as 'Feat. Extr.' in Fig. 2) feature vectors $\boldsymbol{X}$ are extracted from these images, according to the method described in Section 2.3. The set of NM feature vectors is denoted as $\mathcal{X} = \{\boldsymbol{X}_{i,j}\}_{i=1..N,j=1..M}$, where $\boldsymbol{X}_{i,j} \in \mathbb{R}^k$ denotes the j-th feature vector of the i-th person with components $(\boldsymbol{X}_{i,j})_t$ where $t = 1 \ldots k$.

Statistical Analysis. Firstly, we compute the estimated mean feature vector $\boldsymbol{\mu}_i$ of person i and the mean $\boldsymbol{\mu}$ of all enrollment feature vectors as follows,

$$\boldsymbol{\mu}_i = \frac{1}{M}\sum_{j=1}^{M} \boldsymbol{X}_{i,j} \ , \quad \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\mu}_i \quad . \tag{2}$$

Secondly, we compute estimates of the within-class covariance matrix Σ^w and the between-class covariance matrix Σ^b,

$$\Sigma^w = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(\boldsymbol{X}_{i,j}-\boldsymbol{\mu}_i)(\boldsymbol{X}_{i,j}-\boldsymbol{\mu}_i)^T \ , \ \ \Sigma^b = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T. \tag{3}$$

Quantization. In Fig. 2, the quantization block is denoted by 'Q'. In this block a binary representation (bit string) is derived from the input feature vectors of person i denoted as $\mathcal{X}_i = \{\boldsymbol{X}_{i,j}\}_{j=1\ldots M}$. The 'Select User i' block in Fig. 2 selects these feature vectors from the total set $\mathcal{X}$. The quantization of $\mathcal{X}_i$ is based on the mean $\boldsymbol{\mu}$ [1] determined in the 'Statistical Analysis' block. A binary string $Q(\boldsymbol{X}_{i,j})$ is constructed from the feature vector $\boldsymbol{X}_{i,j}$ where each bit $(Q(\boldsymbol{X}_{i,j}))_t$ is defined as (for $t \in \{1, \ldots, k\}$)

$$(Q(\boldsymbol{X}_{i,j}))_t = \begin{cases} 0 & \text{if } (\boldsymbol{X}_{i,j})_t \leq (\boldsymbol{\mu})_t \\ 1 & \text{if } (\boldsymbol{X}_{i,j})_t > (\boldsymbol{\mu})_t \ . \end{cases} \tag{4}$$

Selecting Reliable Components. In this step we look for the reliable components in the M bit strings $Q(\mathcal{X}_i) = \{Q(\boldsymbol{X}_{i,j})\}_{j=1\ldots M}$ of user i. The block 'Reliable Bits' of Fig. 2 determines the K most reliable components (or bits) for user i and creates a first set of helper data $\boldsymbol{W1}_i$. K is a fixed parameter[2] that matches the length of the codewords that are going to be used in the 'Creating Helper Data' step. The reliable components are defined as follows.

The t-th component of $Q(\boldsymbol{X}_{i,j})$ for a fixed user $i = 1, \ldots, N$ is called *reliable*, if the values $(Q(\boldsymbol{X}_{i,j}))_t$ for $j = 1 \ldots M$ are *all equal*. The boolean vector $\boldsymbol{B}_i \in \{0,1\}^k$ denotes the reliable bits. Its t-th entry equals one if the t-th component of $Q(\boldsymbol{X}_{i,j})$ is reliable otherwise the t-th entry is zero. For user that have less than K reliable components, we additionally define *soft reliable* components. Define *p-soft reliable components* of user i as the values t for which $M - p$ of the values $(Q(\boldsymbol{X}_{i,j}))_t$ for $j = 1 \ldots M$ are equal. The boolean vector $\boldsymbol{B}_i^{(p)} \in \{0,1\}^k$ denotes these p-soft reliable bits.

Creating Helper Data. The helper data of our scheme consists of two parts. The first part, denoted by the vector $\boldsymbol{W1}$ is determined as follows. We define the Signal-to-Noise Ratio vector $\boldsymbol{\xi} \in \mathbb{R}^k$ by the following equation,

$$(\boldsymbol{\xi})_t = \frac{(\Sigma^b)_{t,t}}{(\Sigma^w)_{t,t}} \quad , \quad t \in \{1, \ldots, k\}. \tag{5}$$

1. For each user i we determine the K most reliable components with the highest Signal-to-Noise Ratio based on the vectors $\boldsymbol{\xi}$, $\boldsymbol{B_i}$ and $\boldsymbol{B_i^{(p)}}$: first the reliable components (indicated by $\boldsymbol{B_i}$) with the highest $\boldsymbol{\xi}_t$ value are chosen. If the chosen amount of components is less than K, the p-soft reliable components with the highest $\boldsymbol{\xi}_t$ value are added (for successively $p = 1, 2, \ldots$) until a total amount of K components is chosen. The positions of these chosen components are stored in the vector $\boldsymbol{W1}_i \in \mathbb{N}^K$.

[1] Instead of the mean, the median can be used too. This leads to the same results

[2] The value of K is chosen in such a way that the vast majority of users have more than K reliable components

2. For each user i, we select the bits indicated by helper data $\boldsymbol{W1}_i$ and combine these bits into a new vector Z_i. More precisely, $(Z_i)_t = (Q(\boldsymbol{X}_{i,j}))_{(\boldsymbol{W1}_i)_t}$. (This step corresponds to the 'Select' box in Fig. 2).
3. Let $\mathcal{C}$ be an ECC [3] with parameters (K, s, d) where K denotes the length of the code words, s the number of information symbols and d the number of errors that can be corrected. For each user i, a secret $S_i \in \{0,1\}^s$ is randomly choosen [4] and encoded into the codeword $C_i \in \mathcal{C}$. The second part of the helper data $W2_i$ is then given by $W2_i = C_i \oplus Z_i$ (where $\oplus$ stands for bitwise XOR).

Finally the secret S_i is hashed using a cryptographic (one-way) hash function H and the values $\boldsymbol{W1}_i$, $W2_i$ and $H(S_i)$ are stored in the database (indicated with 'DB' in Fig. 2), linked to user i. Note that the secret size equals the number of information symbols s in the (K, s, d) code $\mathcal{C}$.

3.2 Verification

During the verification phase a noisy biometric F_i' of user i is measured. On F_i' the following computations are performed. i) Features are extracted from F_i' and a feature vector $\boldsymbol{X'}_i$ is obtained. ii) In the quantization block a bit string is derived by comparing the value of each component $(\boldsymbol{X'}_i)_t$ with the mean value $(\boldsymbol{\mu})_t$ according to Eq. 4 (where $\boldsymbol{X}_{i,j}$ is replaced by $\boldsymbol{X'}_i$ and $Q(\boldsymbol{X}_{i,j})$ is replaced by $Q(\boldsymbol{X'}_i)$). iii) The first helper data vector $\boldsymbol{W1}_i$ from the database is used to select K components from $Q(\boldsymbol{X'}_i)$ which yields a bit string Z_i'. iv) Then, $Z_i' \oplus W2_i = C_i \oplus (Z_i \oplus Z_i')$ is computed and the errors are corrected such that C_i' is obtained. v) Finally S_i' is obtained by decoding C_i' and $H(S_i')$ is compared to $H(S_i)$ stored in the database. If both values match, user i is authenticated.

4 Results

4.1 Fingerprint Databases

To compare the performance of the matching algorithms with and without template protection, we applied the recognition algorithms to two fingerprint databases.

i) The first fingerprint database we used is the second FVC2000 [Mai00] database. This database contains 8 images of 110 different fingers. The 8-bit gray scale fingerprint images were captured using a capacitive sensor with a resolution of 500 dpi. The image size is 256 by 364 pixels. We use six fingerprints per person during enrollment, two fingerprints per person during verification.

ii) The second fingerprint database is collected at the University of Twente using an optical digitalPersona U.are.U sensor. This database contains 5 images of 500 different fingers. The resolution of the images is 500 dpi, the bit-depth is 8 bit, and the image size is 452 by 492 pixels. We used 4 fingerprints per person during enrollment, and one fingerprint per person for verification.

[3] ECC stands for Error Correcting Code

[4] This is indicated by the Random Number Generator (RNG) block in Fig. 2

4.2 Classification Without Template Protection

For comparison we implemented a likelihood ratio-based verification scheme. For the first database this yields an $EER = 1.4\%$ and for the second database an $EER = 1.6\%$ [5] The results are shown in Fig. 3.

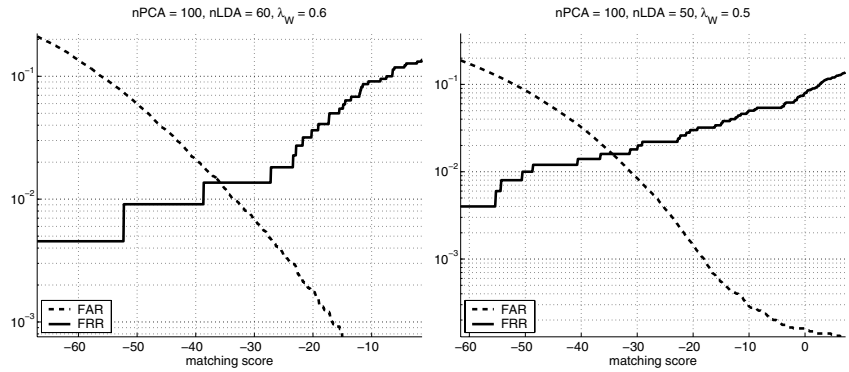


Fig. 3. Likelihood ratio-based results on databases 1 (left) and 2 (right)

4.3 Classification Results of the Reliable Component Scheme

In this section we give the results of the proposed Reliable Components Scheme for the databases described in Section 4.1.

The ECC we use is a binary BCH code described by the triplet (K, s, d). Since BCH codes do not exist for all triplets (K, s, d) we choose from the list of possible BCH codes the one that maximizes the performance. This choice is made as follows. For a set of test users, we investigate how the performance depends on the used BCH code. Fix K as explained in the enrollment procedure according to a valid BCH code. For our fingerprint databases, $K = 511$ is a good choice since the vast majority of users have more than 511 reliable components[6]. For this value of K consider the set of possible BCH codes $\mathcal{B}$ corresponding to all possible values of d (see also Figure 2).

i) For each possible value of d (according to K) choose a code $B(d)$ from $\mathcal{B}$.
ii) Perform enrollment i.e. determine $S, \boldsymbol{W1}, W2$.
iii) Perform the verification phase and compute the FAR(d) and the FRR(d) for that value of d.

[5] For database 1, we used $n_{\mathrm{PCA}} = 100$, $n_{\mathrm{LDA}} = 60$, $\lambda_W = 0.5$ and threshold of -36. For database 2, we used $n_{\mathrm{PCA}} = 100$, $n_{\mathrm{LDA}} = 50$, $\lambda_w = 0.5$ and a threshold of -35. (n_{PCA},$n_{\mathrm{LDA}} = 50$ stand for the dimension after the PCA and LDA transformation respectively and λ_w is a regularization constant)

[6] On average a user has $\approx$ 800 reliable components. When selecting $K = 511$, on average 13 users will have less than 511 reliable bits (in both databases) and hence their helper data vector $\boldsymbol{W1}_i$ also contains 1-soft reliable bits

As mentioned in section 4.1, we split the fingerprint database in a set of enrollment measurements and a set of verification measurements. The dependence of the FAR and the FRR on d is shown in Fig. 4 for one particular split. Clearly, when d is small the FAR will be small but the FRR will be rather high because the system is sensitive to noise. The results that we present here, are calculated by averaging over all possible splits: $\binom{8}{6} = 28$ different splits for database 1 and $\binom{5}{1} = 5$ splits for database 2. On average, the EER is achieved for $d \approx 86$ and $d \approx 102$ for database 1 and 2 respectively.

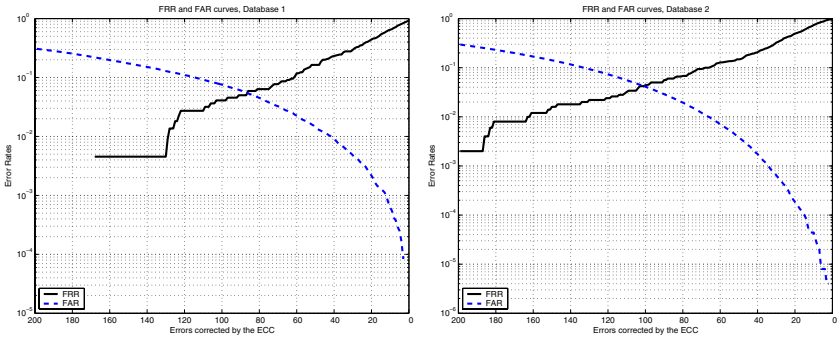


Fig. 4. Left: FAR and FRR as a function of d for database 1, where the training set consists of measurements $\{2, 3, 4, 5, 6, 7\}$ and the verification set of measurements $\{1, 8\}$. Right: FAR and FRR as a function of d for database 2, where the training set consists of measurements $\{1, 2, 3, 4\}$ and the verification set of measurement $\{5\}$

The BCH code that is closest to our (on average) required error correcting capability has parameters $(511, 76, 85)$ for database 1 and parameters $(511, 40, 95)$ for database 2. Fig. 5 summarizes the resulting FRR and FAR that can be achieved using these codes in columns 3 and 4. Furthermore the results for a few other codes (with error correcting capability close to the required average) are displayed. The figure shows that the average EER that can be achieved for database 1 is close to 5.3% and for database 2 around 4.5%. It turns out that many false acceptances and false rejections occur for certain people that have some low quality pictures in the original fingerprint database. For example, some

	ECC (K,s,d)	**FRR**	**FAR**	**FRR***	**FAR***
Database 1	(511,85,63)	0.099	0.025	0.069	0.029
	(511,76,85)	0.054	0.052	0.035	0.058
	(511,67,87)	0.052	0.055	0.034	0.061
Database 2	(511,49,93)	0.054	0.032	0.048	0.033
	(511,40,95)	0.054	0.035	0.048	0.036
	(511,31,109)	0.041	0.055	0.035	0.056

Fig. 5. Summary of the results for the two databases, for several selections of ECCs. (*): The columns FRR* and FAR* show the results if badly enrolled users are not taken into account

users have measurements where the core is on the edge of the picture or where no ridges can be distinguished. Users with such pictures amongst their enrollment data, will often have less than 511 reliable bits (and soft reliable bits are added). In a practical situation, these low quality pictures can easily be avoided during enrollment by visually checking the image quality of each enrollment measurement and repeating a measurement if the quality is too low. We tested this idea by leaving out users for which $\boldsymbol{W1}_i$ contains also *soft reliable* bits (see section 3.1). The results in terms of FAR and FRR are printed in the last two columns of Fig. 5. The performance for database 1 has improved, achieving an EER of about 4.5%. For database 2 the result is only slightly better with an EER of about 4.2%.

It follows from the results that the Reliable Components Scheme degrades the classification performance when compared to the likelihood ratio based scheme but performance is still of the same order (from a security point of view).

5 Security Analysis

The helper data consists of two parts ($\boldsymbol{W1}$ and $W2$) which are used for reliable feature extraction and noise correction on discrete data respectively, we discuss the information leakage by both parts. We present the analysis under the assumption that the quantized strings $Q(\boldsymbol{X})$ are randomly distributed over $\{0,1\}^{511}$ [7]. It follows from results in [TG04], that $\mathsf{H}(S|W2) = \mathsf{H}(S)$, i.e. $W2$ leaks no information on S. It follows from the assumption on the distribution of $Q(\boldsymbol{X})$ that $\boldsymbol{W1}$ does not provide information on S. Hence, the scheme is 0-revealing. It follows from the results in [TG04] that for the discrete case $\mathsf{H}(Q(\boldsymbol{X})|W) \geq \mathsf{H}(Q(\boldsymbol{X})) - (K - s)$ when a (K, s, d) BCH code is used. For database 1 using a (511,76,85) code, this implies that the helper data $W2$ reveals 435 bits and for database 2 using a (511,40,95) code it reveals 471 bits. We note however that given the helper data $W2$, the space of quantized fingerprints $Q(\boldsymbol{X})$ is still sufficiently large (2^{76} and 2^{40} respectively) to make an attack exploiting the helper data infeasible. Again from the assumption on the distribution of $Q(\boldsymbol{X})$ it follows that $W1$ does not increase the information leakage substantially.

6 Conclusions

We showed in this paper, that template protecting biometric authentication techniques can be efficiently implemented with a performance of EER $\approx 4.2\%$ and secret size ≈ 40 bits on fingerprints. The main idea consist of splitting the helper data in two parts, one part determines the reliable components and the other part allows for noise correction on the quantized representations.

[7] We can not prove this at the moment and need more data to compute the distribution of the strings $Q(\boldsymbol{X})$. The presented analysis gives however a good idea of how the security of the system has to be analyzed

References

[Baz02] A.M. Bazen and S.H. Gerez, Systematic Methods for the Computation of the Directional Field and Singular Points of Fingerprints *IEEE Trans. PAMI*,2002, 24, 7, 905-919.

[Baz04] A.M. Bazen and R.N.J. Veldhuis, Detection of cores in fingerprints with improved dimension reduction *Proc. SPS 2004*, 41-44, Hilvarenbeek, The Netherlands,

[BV04] A.M. Bazen and R.N.J. Veldhuis, Likelihood Ratio-Based Biometric Verification, *IEEE Trans. Circuits and Systems for Video Technology*, 2004, 14, 1, 86-94.

[Bolling] J. Bolling, A window to your health, In *Jacksonville Medicine*, 51, Special Issue: Retinal deseases.

[DRS04] Y. Dodis, L. Reyzin, A. Smith, Fuzzy Extractors: How to generate strong secret keys from biometrics and other noisy data, In *Advances in Cryptology - Eurocrypt'04*, LNCS 3027, 523–540, 2004.

[Jai00] A.K. Jain and S. Prabhakar and L. Hong and S. Pankanti, Filterbank-Based Fingerprint Matching, *IEEE Trans. Image Processing*, 2000, 9, 5, 846-859.

[JPP04] U. Uludag, S. Pankanti, S. Prabhakar, and A.K. Jain, Biometric Cryptosystems: Issues and Challenges, In *Proceedings of the IEEE*, Vol. 92, 6, June 2004.

[JS02] A. Juels, M. Sudan, A Fuzzy Vault Scheme *Proceedings of the 2002 International Symposium on Information Theory (ISIT 2002)*, Lausanne.

[JW99] A. Juels and M. Wattenberg, A fuzzy commitment scheme, *6th ACM Conference on Computer and Communication Security*, p. 28-36, 1999.

[LT03] J.-P. Linnartz and P. Tuyls, New shielding functions to enhance privacy and prevent misuse of biometric templates, *4th International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[Mai00] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman and A.K. Jain, FVC2000: Fingerprint Verification Competition, *IEEE Trans. PAMI*, 2002, 24, 3, 402-412.

[MMJ03] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*, Springer-Verlag New-York 2003.

[P65] L. Penrose, Dermatoglyphic topology, *Nature*, 205, (1965) 545-546.

[S99] B. Schneier, Inside risks: The uses and abuses of Biometrics, *Communications of the ACM*, 42, p136, 1999.

[Sou98] C. Soutar, D. Roberge, S.A. Stojanov, R. Gilroy and B.V.K. Vijaya Kumar, Biometric Encryption-Enrollment and Verification Procedures, Proc. of SPIE, Vol. 3386, 24-35, April 1998.

[TG04] P. Tuyls and J. Goseling, Capacity and Examples of Template Protecting Biometric Authentication Systems, *Biometric Authentication Workshop (BioAW 2004)*, LNCS 3087, 158–170, Prague, 2004.

[TVI04] P. Tuyls, E. Verbitskiy, T. Ignatenko, D. Schobben and T.H. Akkermans Privacy Protected Biometric Templates: Ear Identification Proceedings of SPIE, Vol. 5404, 176-182, April 2004.

A Study of Brute-Force Break-Ins of a Palmprint Verification System

Adams Kong[1,2], David Zhang[1], and Mohamed Kamel[2]

[1] Biometrics Research Centre, Department of Computing,
Hong Kong Polytechnic University, Kowloon, Hong Kong
adamskong@ieee.org
csdzhang@comp.polyu.edu.hk
[2] Pattern Analysis and Machine Intelligence Lab,
University of Waterloo, 200 University Avenue West, Ontario, Canada
mkamel@uwaterloo.ca

Abstract. Biometric systems are widely applied since they offer inherent advantages over traditional knowledge-based and token-based personal authentication approaches. This has led to the development of palmprint systems and their use in several real applications. Biometric systems are not, however, invulnerable. The potential attacks including replay and brute-force attacks have to be analyzed before they are massively deployed in real applications. With this in mind, this paper will consider brute-force break-ins directed against palmprint verification systems.

1 Introduction

Accurate automatic personal authentication does not only act as an important means for protecting our lives and properties, it is also an integral element in the ever rapidly expanding e-applications arena, playing in our everyday encounters such as e-banking, e-commerce, e-kiosks, etc. Traditional security systems which automatically identify individuals generally use either tokens of private possessions such as a physical key or private knowledge such as a password. Such tokens are insecure. They can be shared, duplicated, lost or stolen. In this respect, biometric systems that recognize individuals based on their physiological and behavioral characteristics such as the fingerprint, face, iris, palmprint or signature are much more secure. However, they are not invulnerable. For instance, the systems can be broken into using replay and brute-force attacks.

Fig. 1 shows a generic biometric system, where Points 1-8 are vulnerable points as identified by [2-3]. At Point 1, a system is able to be spoofed using fake biometrics e.g. face masks and artificial gummy fingerprints [4]. At Point 2, liveness detection countermeasures in the sensors can be avoided by using a pre-recorded signal such as iris image. This is a so-called replay attack. At Point 3, a Trojan horse can override the feature extraction process so that the original output features are replaced with a predefined feature. At Point 4, it is possible to use both replay and brute-force attacks, submitting on the one hand prerecorded templates or, on the other, numerous synthetic templates. At Point 5, the matching scores obtained can be replaced with preselected matching scores by using a Trojan horse. At Point 6, it is possible to modify templates in the database or to insert templates from unauthorized users into the data-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 447–454, 2005.
© Springer-Verlag Berlin Heidelberg 2005

base. At Point 7, replay attacks are once again possible. At Point 8, it is possible to directly override the decision output of the system.

In remote, unattended applications, such as web-based e-commerce applications, attackers may have enough time to make complex and numerous attempts to break in. Security and biometric researchers have recently proposed methods for detecting and preventing these attacks [2-3, 5-8]. Some researchers have analyzed specific attack types vis-à-vis specific biometrics, for instance, brute-force attacks at Point 4 of fingerprint systems [2-3, 5].

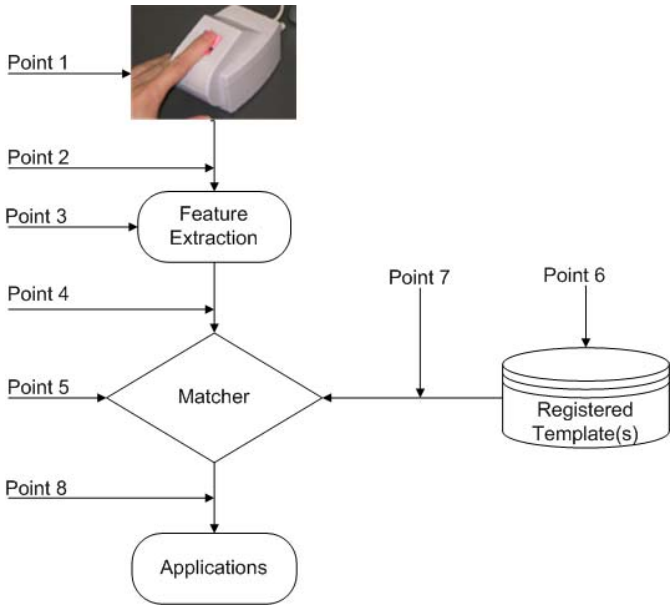


Fig. 1. Vulnerable points in a biometric system

Given the commercial potential of palmprint systems and the variety of capture devices, and preprocessing, feature extraction, matching and classification algorithms [9-16] that have been developed over the last several years, it is certainly the case that any security issues should be systematically addressed prior to their widespread deployment. In this paper, we concentrate on brute-force attacks at Point 4. As far as we know, this is the first paper that considers security issues in palmprint systems.

The rest of this paper is organized as the follows. Section 2 gives an overview of the palmprint system for this analysis. Section 3 provides a probabilistic model describing the relationship between number of attacks and false acceptance rates. Section 4 provides experimental results. Finally, Section 5 offers some concluding remarks and further research directions.

2 A Summary of the Palmprint System Using Competitive Code

In this Section, we introduce our palmprint system using a palmprint identification algorithm known as Competitive Code [13-14]. We select to study Competitive Code

in the context of brute-force attacks rather than other palmprint algorithms since it is the most accurate and fastest algorithm developed by us [10, 15, 17-19]. Precisely, Competitive Code can operate at a high genuine acceptance rate of 98.4% while the corresponding false acceptance rate is 3×10^{-6}% [14]. The computation speed of Competitive Code can be comparable with IrisCode [20] since the angular distance is implemented using Boolean operators. In addition to speed and accuracy, Competitive Code can effectively distinguish the palmprints of identical twins [16]. Our system using Competitive Code consists of the following parts.

Image acquisition: Transmit a palmprint image to a processor from a palmprint scanner [13]. Fig. 2(a) shows a palmprint scanner developed by the Biometrics Research Centre, The Hong Kong Polytechnic University and Fig. 2(b) shows a collected palmprint image.

Preprocessing: Determine the two key points between fingers to establish a coordinate system for aligning different palmprint images [13]. Then, extract the central parts on the base of the coordinate system. Fig. 2(c) illustrates the key points and the coordinate system and Fig. 2(d) shows a preprocessed palmprint image.

Feature extraction: The real parts of six Gabor filters with different orientations, $\psi_R(x,y,\theta_j)$, where θ_j represents the orientation of the filters are applied to a preprocessed palmprint image, *I(x,y)* [14]. The orientation of a sample point is estimated using a competitive rule, $k=arg(min_j(I(x,y)*\psi_R(x,y,\theta_j)))$, where *k* is called the winning index and *j=0, 1, 2, 3, 4,* and *5.* Combining the winning indexes at different sample points, we have the final feature, called Competitive Code.

Coding: For effective matching, the winning indexes are coded using Table 1. Three bits are used to represent one winning index.

Angular comparison: The difference between two Competitive Codes is measured using their angular distance. The bitwise representation of angular distance is defined as:

$$A_H(P,Q)=\frac{\sum_{y=1}^{N}\sum_{x=1}^{N}\sum_{i=1}^{3}(P_M(x,y)\cap Q_M(x,y))\cap(P_i^b(x,y)\otimes Q_i^b(x,y))}{3\sum_{y=1}^{N}\sum_{x=1}^{N}P_M(x,y)\cap Q_M(x,y)} \quad (1)$$

where $P_i^b\left(Q_i^b\right)$ is the i^{th} bit plane of Competitive Code *P(Q)*; $P_M(Q_M)$ is the mask of *P(Q)* used to denote the non-palmprint pixels; $\otimes$ is bitwise exclusive OR; $\cap$ is bitwise AND and N^2 is the size of Competitive Code. Obviously, A_H is between 0 and 1. Since the preprocessing algorithm is not perfect, one of the features must be translated horizontally and vertically and then the matching is carried out again. The ranges of both the horizontal and the vertical translations are –2 to 2.The minimum of the A_H's obtained by translated matching is regarded as the final angular distance, A_f.

Table 1. Bitwise representation of the Competitive Code

Winning index	Bit 1	Bit 2	Bit 3
0	0	0	0
1	0	0	1
2	0	1	1
3	1	1	1
4	1	1	0
5	1	0	0

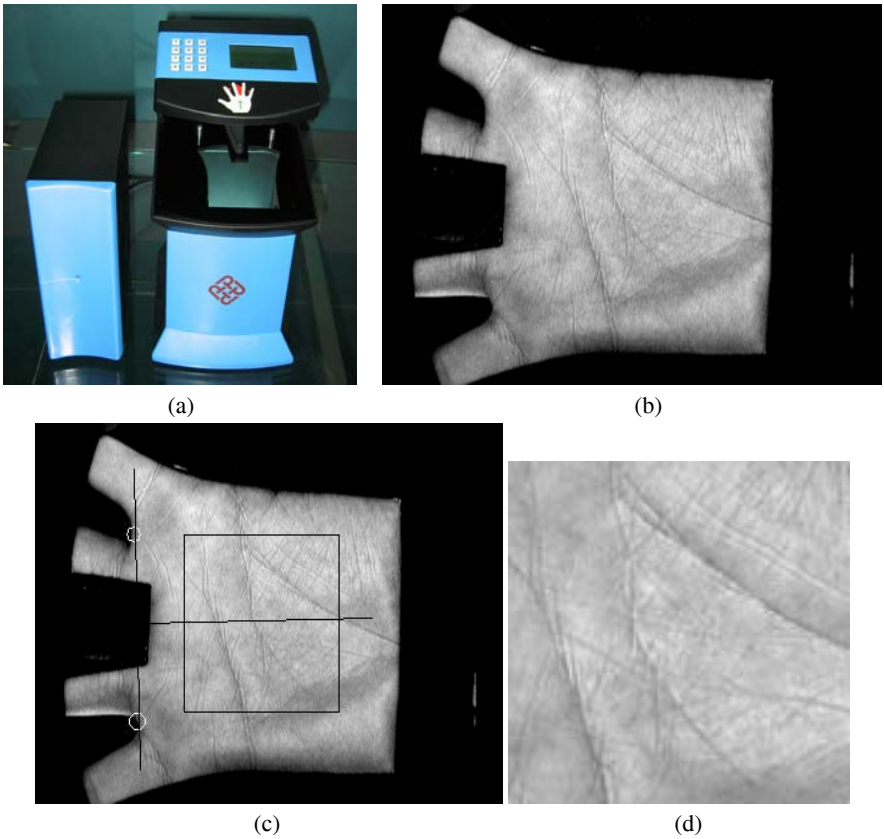


Fig. 2. Illustration of palmprint collection and pre-processing: a) a palmprint scanner developed by the Biometrics Research Centre, The Hong Kong Polytechnic University, b) a collected palmprint image, c) the key points and coordinate system for palmprint segmentation and alignment and d) the pre-processed palmprint image for the proposed framework.

3 A Probabilistic Model for Studying Brute-Force Break-Ins

To analyze brute-force break-ins, we have to develop a probabilistic model that describes the relationship between the probability of a false acceptance and the numbers of attacks. In other words, we require a probabilistic model for the angular distance described in Eq. 1. To simplify the model, we assume that all preprocessed palmprint

images contain no non-palmprint pixels. This will allow us to neglect the normalization constant and the masks. For the sake of convenience, we employ the integer representation of Competitive Code rather than the bitwise representation for the following analysis. As a result, the angular distance between two Competitive Codes is

$$A_H(P,Q) = \sum_{y=1}^{N}\sum_{x=1}^{N} A(P_{x,y}, Q_{x,y}), \tag{2}$$

where $A(P_{x,y}, Q_{x,y})$ is the angular distance between two winning indexes, $P_{x,y}$ and $Q_{x,y}$. Table 2 gives all possible angular distances.

Let $W = [w_o, w_1, w_2, w_3]$ be a random vector where w_i is the number of $A(P_{x,y}, Q_{x,y}) = i$ in Eq. 2 and let p_i be the probability of $A(P_{x,y}, Q_{x,y}) = i$. Consequently, we can rewrite the angular distance described in Eq. 2 as $A_H(P,Q) = WK^T$, where $K = [0, 1, 2, 3]$. We also assume that $A(P_{x,y}, Q_{x,y})$ is independent and p_i is stationary. By "stationary" we mean that p_i does not depend on the position *(x, y)*. Similar assumptions have been employed in analyzing brute-force break-ins of fingerprint systems [2-3]. Using these assumptions, we can infer that *W* follows multinomial distribution i.e.

$$f(w_0, w_1, w_2, w_3) = \frac{n!}{w_0!\, w_1!\, w_2!\, w_3!}\, p_0^{w_0} p_1^{w_1} p_2^{w_2} p_3^{w_3}, \tag{3}$$

where *n* is equal to N^2, the effective matching area. Therefore, the probability density of the angular distance is,

$$\Pr(A_H(P,Q) = t) = \sum_{W \ni WK^T = t} f(w_o, w_1, w_2, w_3). \tag{4}$$

So far, we have established a probability model in which the model parameter *n* depends on the effective matching area. This area changes according to the translated matchings. To simplify the following formulation, we treat all the translated matchings as having the same effective matching area, i.e., 900. It is the minimum matching area.

Let $\Pr(A_H(P,Q) < t) = F(t)$ and thus, $\Pr(A_H(P,Q) \geq t) = 1 - F(t)$. The probability of the final angular distance A_f being greater than the threshold *t* is

$$\Pr(A_f(P,Q) \geq t) = (1 - F(t))^m, \tag{5}$$

where *m*, the number of translated matchings is 25. If we make *z* independent comparisons, the probability of all the angular distances being greater than or equal to *t* is

$$\Pr(A_f(P_i, Q_i) \geq t \mid \forall i = 1,..., z) = (1 - F(t))^{mz}, \tag{6}$$

where P_i and Q_i represent different Competitive Codes. Finally, the probability of at least one of final angular distances being shorter than *t* is

$$\Pr(A_f(P_i, Q_i) < t) = 1 - (1 - F(t))^{mz}. \tag{7}$$

Now, we are able to analyze brute-force attacks against our system using Eq. 7. For verification, each submitted templates, P_i as a brute-force attack, is matched with the templates associated with a particular user. We assume that each user only has one template, Q in the database and the hackers submit z templates to attack the system. Therefore, the probability of a false acceptance for verification is

$$\Pr(A_f(P_i, Q) < t) = 1 - (1 - F(t))^{mz}, \tag{8}$$

the same as Eq. 7.

Table 2. All possible angular distances between different winning indexes, the elements of Competitive Code

Angular distance		Winning indexes					
		0	1	2	3	4	5
Winning indexes	0	0	1	2	3	2	1
	1	1	0	1	2	3	2
	2	2	1	0	1	2	3
	3	3	2	1	0	1	2
	4	2	3	2	1	0	1
	5	1	2	3	2	1	0

4 Parameter Estimation and Experimental Results

The use of the probabilistic model to investigate brute-force break-ins into our palmprint system requires us to make some assumptions to obtain the model parameters, p_i. We suppose that the attackers use uniform distributions to generate the winning indexes of their synthetic Competitive Codes. We also assume that the winning indexes of the template, Q, in database follow uniform distribution and all of their winning indexes are independent, we can infer that

$$p_0 = p_3 = 1/6 \tag{9}$$

and

$$p_1 = p_2 = 1/3 \tag{10}$$

from Table 2. Using these parameters and Eq. 7, we can estimate the probability of false acceptance at different thresholds and under different number attacks, z. Fig. 3 shows the experimental results but only provides the thresholds in the range between 0.34 and 0.4 since they associate with acceptable false acceptance (general case, not brute-force attack) and false rejection rates for our palmprint system. Our system generally operates at the threshold 0.37, at which threshold, it has a false acceptance rate of 0×10^{-6}% and a genuine acceptance rate of 97.7% [14]. Table 3 lists the probabilities of a false acceptance of brute-force attacks and the corresponding computation time when the threshold is set to 0.37. We assume that the system can make 1 million comparisons per second to estimate the computation time. Fig. 3 and Table 3 show that it is computationally infeasible to use a brute-force attack to break in the system.

Table 3. The probabilities of false acceptance under different number attacks, z when the threshold is set to 0.37 and the corresponding computation times

No of attacks z	Time	Probability of false acceptance
10^{11}	1.16 days	9×10^{-24}
10^{12}	11.5 days	9×10^{-23}
10^{13}	115 days	9×10^{-22}
10^{14}	3.17 years	9×10^{-21}
10^{15}	31years	9×10^{-20}

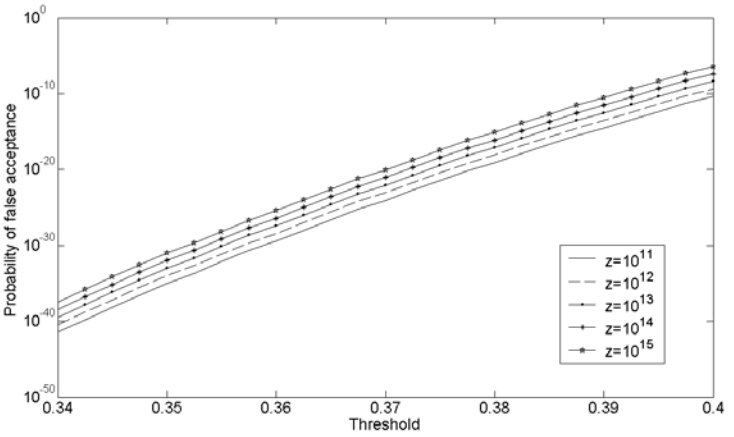


Fig. 3. A plot of the probability of false acceptance against threshold, where z represents the number of attacks

5 Conclusion and Further Research

This paper presents a study of brute-force break-ins directed against our palmprint system that uses Competitive Code as the features and angular distance as the matching scheme. We set up a probabilistic model to describe the relationship between the number of attacks and the probability of false acceptance. According to our analysis, when the system threshold is set to lower than 0.37, it is computationally infeasible to break in our palmprint system using brute-force attacks.

In our previous paper [14], we have developed a bitwise angular distance for matching Competitive Codes. In this paper, we derive a projected multinomial distribution to model the distribution of the angular distance. IrisCode, a well-known biometric recognition method, exploits bitwise hamming distance for comparing two iris features and its imposter distribution is modeled by binomial distribution [20]. The relationships between IrisCode and Competitive Code call for further investigation.

References

1. A. Jain, R. Bolle and S. Pankanti (eds.), *Biometrics: Personal Identification in Networked Society*, Boston, Mass: Kluwer Academic Publishers, 1999.
2. R.M. Bolle, J.H. Connell and N.K. Ratha, "Biometric perils and patches", *Pattern Recognition*, vol. 35, pp. 2727-2738, 2002.

3. N.K. Ratha, J.H. Connell and R.M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems", *IBM Systems Journal*, vol. 40, no. 3, pp. 614-634, 2001.
4. T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial gummy fingers on fingerprint systems", *Proc, SPIE*, vol. 4677, pp. 275-289, San Jose, USA, Feb, 2002.
5. N.K. Ratha, J.H. Connell and R.M. Bolle, "Biometrics break-ins and band-aids", *Pattern Recognition Letters*, vol. 24, pp, 2105-2113, 2003.
6. L. O'Gorman, "Comparing passwords, tokens, biometrics for user authentication", *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021-2040, 2003.
7. U. Uludag, S. Pankanti, S. Prabhakar and A.K. Jain, "Biometric cryptosystems: issues and challenges", *Proceedings of the IEEE*, vol. 92, no. 6, pp. 948-960, 2004.
8. A.K. Jain and U. Uludag, "Hiding biometric data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1494-1498, 2003.
9. W. Shu and D. Zhang, "Automated personal identification by palmprint", *Optical Engineering*, vol. 37, no. 8, pp.2359-2363, 1998.
10. X. Wu, D. Zhang, K. Wang and B. Hung, "Palmprint classification using principal lines", *Pattern recognition*, vol. 37, no. 10, pp. 1987-1998, 2004.
11. C.C. Han, H.L. Cheng, K.C. Fan and C.L. Lin, "Personal authentication using palmprint features", *Pattern Recognition*, vol. 36, no 2, pp. 371-381, 2003.
12. C.C. Han, "A hand-based personal authentication using a coarse-to-fine strategy", *Image and Vision Computing*, vol. 22, pp. 909-918, 2004.
13. D. Zhang, W.K. Kong, J. You and M. Wong, "On-line palmprint identification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1041-1050, 2003.
14. A.W.K. Kong and D. Zhang, "Competitive coding scheme for palmprint verification", *in Proceedings of International Conference on Pattern Recognition*, vol. 1, pp. 520-523, 2004.
15. A.W.K. Kong and D. Zhang, "Feature-level fusion for effective palmprint authentication" *in Proceedings of International Conference on Biometric Authentication*, pp. 761-767, 2004.
16. A. Kong, D Zhang and G. Lu, "A study of identical twins palmprint for personal verification", *To appear in Pattern Recognition.*
17. W. K. Kong and D. Zhang, "Palmprint texture analysis based on low-resolution images for personal authentication", *in Proceedings of International Conference on Pattern Recognition*, pp. 807-810, 2002.
18. X. Wu, D. Zhang and K. Wang, "Fisherpalms based palmprint recognition", *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2829-2838, 2003.
19. L. Zhang and D. Zhang, "Characterization of palmprints by wavelet signatures via directional context modeling", *IEEE Transactions on Systems, Man and Cybernetics, Part B*, vol. 34, no. 3, pp. 1335-1347, 2004.
20. J. Daugman, "High confidence visual recognition of persons by a test of statistical independence", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148-1161, 1993

Modification of Intersession Variability in On-Line Signature Verifier

Yasunori Hongo, Daigo Muramatsu, and Takashi Matsumoto

Department of Electrical Engineering and Bioscience, Waseda University,
3-4-1 Ohkubo, Shinjuku-ku, Tokyo, Japan
{hongo03,daigo}@matsumoto.elec.waseda.ac.jp
takashi@mse.waseda.ac.jp
http://www.matsumoto.elec.waseda.ac.jp/

Abstract. For Pen-input on-line signature verification algorithms, the influence of intersession variability is a considerable problem because hand-written signatures change with time, causing performance degradation. In our previous work, we proposed a user-generic model using AdaBoost. However, this model did not allow for the fact that features of signatures change over time. In this paper, we propose a template renewal method to reduce the performance degradation caused by signature changes over time. In our proposed method, the oldest template is replaced with a new one if the new signature data gives rise to an index which exceeds a threshold value. No further learning is necessary. A preliminary experiment was conducted on a subset of the MCYT database.

1 Introduction

Personal identity verification has a variety of applications including electronic commerce, access control for buildings and computer terminals, and credit card verification. The algorithms used to verify personal identity can be classified into the four groups described in Fig. 1, depending on whether they are static, dynamic, biometric, or physical/knowledge-based.

For example, algorithms for fingerprints, the iris, the retina, DNA, palm prints, the face, and the blood vessels are static and biometric. Algorithms classified as biometric and dynamic involve lip movements, body movements, the voice, and on-line signatures. Schemes that use passwords are static and knowledge-based, whereas methods using IC cards, magnetic cards, or keys are physical. Due to the rapidly increasing use of Tablet PCs and PDAs, on-line signature verification is a promising technique for personal identity verification.

A variety of algorithms have been proposed for on-line signature verification. Research results continue to be reported, indicating that this problem is difficult and challenging.

2 The Algorithm

2.1 Feature Extraction

The raw data from our readily available tablet (WACOM INTUOS A6 USB) consists of the five-dimensional time series data set:

$$(x(j), y(j), p(j), \gamma(j), \varphi(j)) \in R^2 \times \{0,1,...,1023\} \times R^2 \quad j = 1,2,...,J \tag{1}$$

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 455–463, 2005.
© Springer-Verlag Berlin Heidelberg 2005

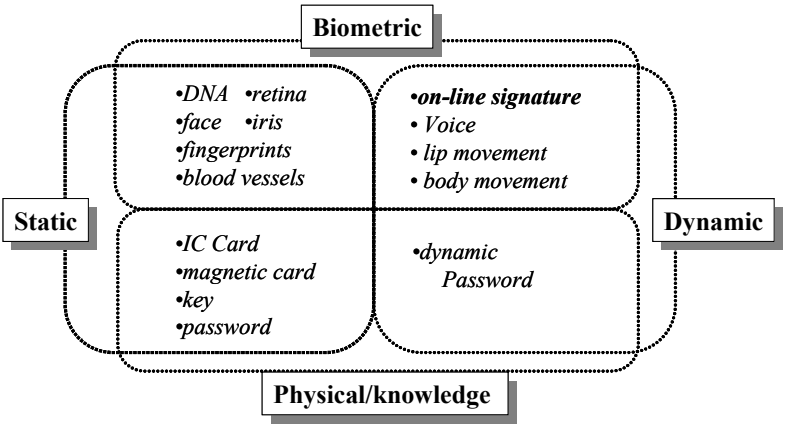


Fig. 1. Authentication methods

where $(x(j), y(j)) \in R^2$ is the pen position at time j, $p(j) \in \{0,1,\ldots,1023\}$ represents the pen pressure, $\gamma(j)$ is pen azimuth angle and $\varphi(j)$is pen altitude angle.

Define

$$X_g = \frac{\sum_{j=1}^{J} x(j)}{J} \tag{2}$$

$$Y_g = \frac{\sum_{j=1}^{J} y(j)}{J} \tag{3}$$

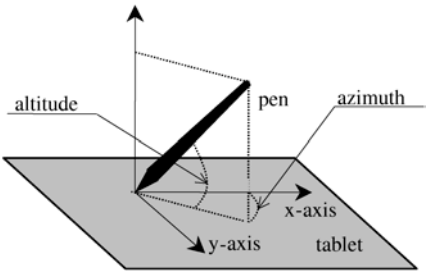


Fig. 2. Raw data from tablet

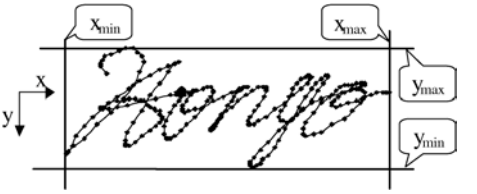


Fig. 3. x_{min}, x_{max}, y_{min} and y_{max} of signature

Let

$$(dx(j), dy(j)) := \left(\frac{x(j) - X_g}{x_{max} - x_{min}} \times L, \frac{y(j) - Y_g}{y_{max} - y_{min}} \times L\right) \qquad j = 1,2,\ldots,J \tag{4}$$

be the relative pen position with respect to (2) and (3) where L is a scaling parameter.

The length $f(j)$ and the angle $\theta(j)$ of each pen position are given by

$$f(j) = \sqrt{dx(j)^2 + dy(j)^2} \qquad j = 1,2,\ldots,J \tag{5}$$

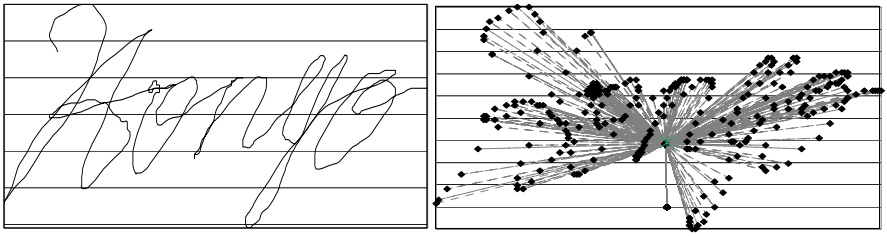
$$\theta(j)=\begin{cases}\tan^{-1}\dfrac{dy(j)}{dx(j)} & (dx(j)>0)\\ sign(dy(j))\times\dfrac{\pi}{2} & (dx(j)=0)\\ \tan^{-1}\dfrac{dy(j)}{dx(j)}+\pi & (dx(j)<0,dy(j)\geq 0)\\ \tan^{-1}\dfrac{dy(j)}{dx(j)}-\pi & (dx(j)<0,dy(j)<0)\end{cases} \qquad j=1,2,...,J \tag{6}$$

Feature vectors that we use consist of the following five-dimensional data elements:

$$(\theta(j),f(j),p(j),\gamma(j),\varphi(j))\in R^2\times\{0,1,...,1023\}\times R^2 \quad j=1,2,...,J \tag{7}$$

where J is the number of sample points.

A typical original signature trajectory given by Fig. 4(a) is converted into the relative trajectory given by Fig. 4(b).



(a) (b)

Fig. 4. (a) Original signature trajectories, (b) Relative trajectories

2.2 Distance Calculation

Let

$$(\eta(k),g(k,)q(k),\delta(k),\phi(k))\in R^2\times\{0,1,...,1023\}\times R^2 \quad k=1,2,...,K \tag{8}$$

be the feature trajectory of a template signature.

We calculate the following six kinds of distance using each feature value along the number of sample points:

$$d_1 := \min_{\substack{i_s\le i_{s+1}\le i_s+1\\ k_s\le k_{s+1}\le k_s+1}}\sum_{s=1}^{S}\left|\theta(j_s)-\eta(k_s)\right| \tag{9}$$

$$d_2 := \min_{\substack{i_s\le i_{s+1}\le i_s+1\\ k_s\le k_{s+1}\le k_s+1}}\sum_{s=1}^{S}\left|p(j_s)-q(k_s)\right| \tag{10}$$

$$d_3 := \min_{\substack{i_s\le i_{s+1}\le i_s+1\\ k_s\le k_{s+1}\le k_s+1}}\sum_{s=1}^{S}\left|f(j_s)-g(k_s)\right| \tag{11}$$

$$d_4 := \min_{\substack{i_{s'}\le i_{s'+1}\le i_{s'}+1\\ k_{s'}\le k_{s'+1}\le k_{s'}+1}}\sum_{s'=1}^{S'}\left|\gamma(j_s)-\delta(k_s)\right| \tag{12}$$

$$d_5 := \min_{\substack{i_{s''}\le i_{s''+1}\le i_{s''}+1\\ k_{s''}\le k_{s''+1}\le k_{s''}+1}}\sum_{s''=1}^{S''}\left|\varphi(j_s)-\phi(k_s)\right| \tag{13}$$

$$d_6 := |J - K| \tag{14}$$

where $j_1=k_1=1$, $j_s=J$, $k_s=K$. J and K denote the size of the data (the number of sampled points).

Dynamic Programming can be used for computing $d_1,\ldots,d_5$ because of the sequential nature of the distance function.

$$\begin{aligned} &D_1(0,0) = 0 \\ &D_1(j_{s+1},k_{s+1}) = |\theta(j_s) - \eta(k_s)| + \min \begin{cases} D_1(j_s-1,k_s-1) \\ D_1(j_s-1,k_s) \\ D_1(j_s,k_s-1) \end{cases} \end{aligned} \tag{15}$$

2.3 Authentication Method

To distinguish genuine signatures from forged signatures, we use the six-feature vectors. We choose the Boosting algorithm for separation because its generalization error is small, and it has no free parameter affecting the threshold values when used for signature verification. AdaBoost can thus provide a good classifier.

2.4 AdaBoost

AdaBoost, originally proposed by Freund and Schapire [2], is a methodology which provides a highly accurate classifier by combining many weak classifiers.

We begin with training data, $(u_1, v_1),\ldots, (u_N, v_N)$, where u_i is a vector-valued feature and $v_i = \{-1,+1\}$. The training data has distribution $D_t(i)$, and $D_1(i)$ is uniform. At round t, a weak classifier defines a weak hypothesis $h_t(u_i)$ by a learning scheme that has moderate accuracy. When the classifier defines $h_t(u_i)$, we calculate the error.

$$\varepsilon_t = \sum_i D_t(i) I(h_t(u_i), v_i) \tag{16}$$

$$I(h_t(u_i), v_i) = \begin{cases} 1 & if \quad sign(h_t(u_i)) \neq sign(v_i) \\ 0 & otherwise \end{cases} \tag{17}$$

Using ε_t, we define the classifier's confidence.

$$\alpha_t = \ln \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \tag{18}$$

After defining the classifier's confidence, we change the distribution D using the following update rule:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t v_i h_t(u_i))}{Z_t} \tag{19}$$

where Z_t is a normalization factor.

After finishing the learning stage, we are ready to calculate $F(u)$ and final hypothesis $H(u)$.

$$F(u) = \sum_t \alpha_t h_t(u) \tag{20}$$

$$H(u) = sign(F(u)) \tag{21}$$

2.5 Weak Classifier

In this paper, we use a two-layer perceptron as a weak classifier described by

$$h(u_m;\mathbf{w}_{weak}) := \sum_{j=1}^{U}(\mathrm{b}_j u_m + \mathrm{c}) \tag{22}$$

where

$$\mathbf{w}_{weak} = (\{b_j\},\{c\}) \tag{23}$$

U is the number of dimensions of input data. We use this as a weak classifier because it is simple and easy to calculate. Model parameter w is drawn from a Gaussian distribution.

2.6 Algorithm for Signature Verification

To effectively employ AdaBoost, many signatures belonging to both classes are necessary for training. However, we could use only a few genuine signatures, and there were no skilled forgeries available, so we could not generate a good user-specific model. Therefore, we will propose a user-generic model $Model(w_s)$ which is created by using available database (MCYT database [1] in the present study) where w_s is a parameter vector. The model does not use signature data from the person to be tested. Overall algorithm is described in Fig.5.

In the learning phase, we compute the parameter vector w_s. In the testing phase, we first calculate $u(t,sig_{test})$ defined by (24) below where D_i is defined in (9)-(15). And we use $I_1,\ldots,I_6$ defined by (24) in addition to $D_1,\ldots,D_6$ so in this paper, we use twelve feature vectors (U=12). $temp_i$ is the ith template signature.

$$\begin{aligned} &u(t,s) = (D_1, D_2, \ldots, D_6, I_1, I_2, \ldots, I_6) \\ &D_{i,j} = D_i(temp_j, sig_{test}) \qquad i = 1,\ldots,6,\ j = 1,\ldots,M \\ &I_i = \frac{1}{M^2}\sum_{j=1}^{M}\sum_{k=1}^{M} D_i(temp_j, temp_k) \end{aligned} \tag{24}$$

M is the number of the template signatures.

Secondly we calculate the score as described in (25) and make decision using (26)

$$Score(sig_{test}) = \frac{\sum_{i=1}^{M} F(u(temp_i, sig_{test}))}{\frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{M} F(u(temp_j, temp_k))} \tag{25}$$

$$sig_{test} \text{ is } \begin{cases} \text{genuine} & \text{if } c_{verf} \le Score(sig_{test}) \\ \text{forgery} & \text{if } c_{verf} > Score(sig_{test}) \end{cases} \tag{26}$$

where c_{verf} is a threshold value.

2.7 Template Renewal Method

The intersession variability of hand-written signature causes performance degradation. We propose an algorithm that solves the problem by changing template signatures.

We use the following scheme.

$$\begin{aligned} &\text{if } c_{renew} \le Score(sig_{test}) \quad \text{One of the template signatures is replaced by the } sig_{test} \\ &\text{if } c_{renew} > Score(sig_{test}) \quad \text{All the template signatures are hold} \end{aligned} \tag{27}$$

c_{renew} is the threshold value for changing the template signature.

Reference [3] also proposed a template renewal method, where all signatures that accepted as genuine signature were added to template signatures.

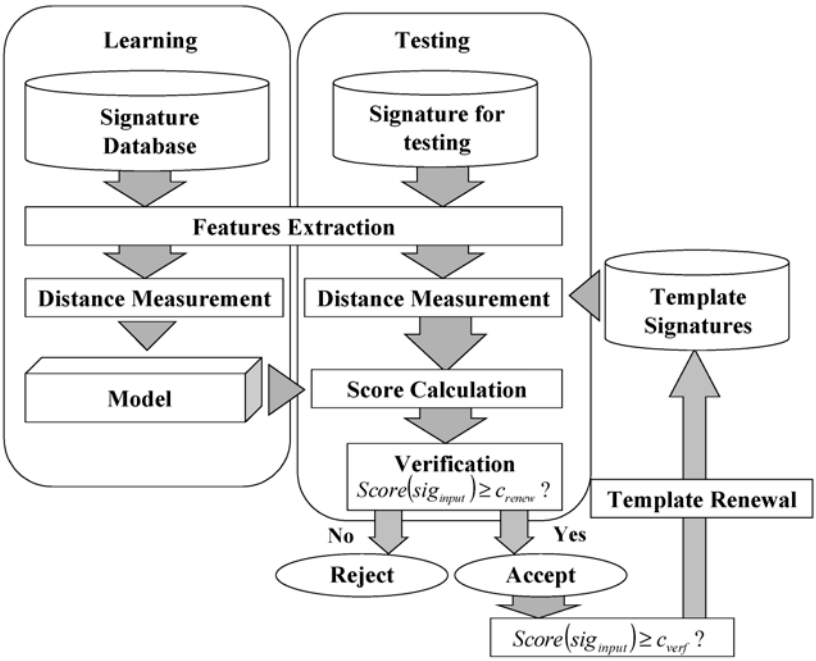


Fig. 5. Overall algorithm

3 Experiment

3.1 An Experiment with the Proposed Algorithm on Data Set 1

This section reports on an experiment using our algorithm for a data set consisting of 1000 signatures from 50 Europeans, with ten genuine signatures and ten skilled forged signatures associated with each individual. We used five genuine signatures as a template (M=5).

In this experiment, we applied a 50-fold cross-validation method. For each experiment, we used 980 signatures for training data (representing 49 individuals, excluding one test person), 5 signatures as template and 15 signatures for test data (5 genuine signatures and 10 skilled forgery signatures).

In this experiment, we continue 1,000 rounds for learning (i.e., T=1000) and set c_{renew}=0.7. Table 3.2 shows the error rates for our algorithm. In order to report an Error tradeoff curve, we tested several values of c_{verf}, although c_{verf}=0 is selected in original AdaBoost. Fig.6 displays the Error tradeoff curve.

Table 3.1. Data Set 1 (for one experiment)

Signatures for Learning		Template Signatures	Signatures for Testing	
Genuine	Forgery	Genuine	Genuine	Forgery
490	490	5	5	10

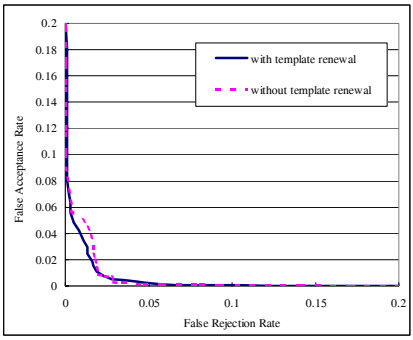


Fig. 6. Error tradeoff curve for Data Set 1

Table 3.2. Verification Error Rate on Data Set 1

	With Template Renewal	Without Template Renewal
EER	1.66%	1.82%
FR(@c_{verf} =0)	1.70%	1.90%
FA(@c_{verf} =0)	1.54%	1.80%
FR(@FA=1%)	1.85%	2.10%

3.2 Experiment Using the Proposed Algorithm on Data Set 2

This section reports on the experimental results of our algorithm for the second data set consisting of 5000 signatures from 100 Europeans, with 25 genuine signatures and 25 skilled forged signatures associated with each individual. About 70% of Data Set 1 is included in Data Set 2. It corresponds to 14% of Data Set 2.

Table 3.3. Data Set 2 (for one experiment)

Signatures for Learning		Template Signatures	Signatures for Testing	
Genuine	Forgery	Genuine	Genuine	Forgery
2475	2475	5	20	25

To show that our algorithm reduces the influence of intersession variability, we divided the genuine signatures used for testing into four groups. Each group consists of five signatures as follows:

Group 0 (template signatures): 1st-5th genuine signatures
Group 1: 6th-10th genuine signatures
Group 2: 11th-15th genuine signatures
Group 3: 16th-20th genuine signatures
Group 4: 21st-25th genuine signatures

In this experiment, we applied a 100-fold cross-validation method. For each experiment, we used 4900 signatures for training data (representing 99 individuals, excluding one test person), 5 signatures as template and 45 signatures for test data (20 genuine signatures and 25 skilled forgery signatures). We continued for 2,000 rounds for learning (i.e., $T = 2000$) and set $c_{renew} = 0.7$.

Table 3.4 shows the error rates for our algorithm. Figure.7 displays the Error tradeoff curve.

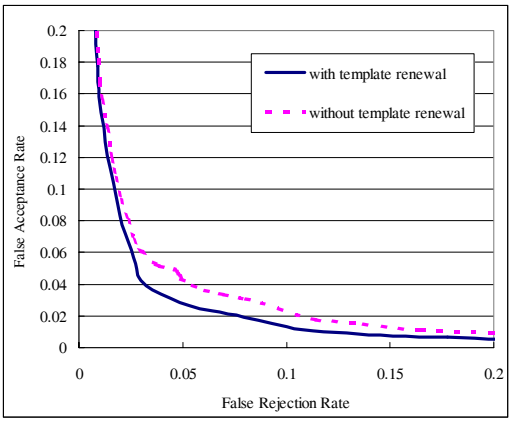


Fig. 7. Error tradeoff curve on Data Set 2

Table 3.4. Verification Error Rate on data set 2

		Total	Group 1	Group 2	Group 3	Group 4
With Template Renewal	EER	3.60%	3.14%	3.43%	3.22%	4.64%
	FR(@c_{verf}=0)	3.55%	2.40%	3.20%	2.60%	6.00%
	FA(@c_{verf} =0)	3.63%	3.63%	3.63%	3.63%	3.63%
	FR(@FA=1%)	11.85%	7.2%	12.7%	11.8%	15.7%
Without Template Renewal	EER	4.72%	3.44%	4.31%	5.06%	5.32%
	FR(@c_{verf} =0)	3.70%	1.60%	3.60%	4.20%	5.40%
	FA(@c_{verf} =0)	5.28%	5.28%	5.28%	5.28%	5.28%
	FR(@FA=1%)	16.15%	12.10%	19.40%	17.40%	23.70%

4 Conclusion

We proposed a verification algorithm using template renewal to reduce the influence of intersession variability. We improved the verification rate by using this algorithm. Considering that no fine tuning was done, this algorithm looks very promising.

References

1. J.Ortega-Garcia and J. Fierrez-Aguilar and D.Simon, J. Gonzalez and M. Faundez-Zanuy, V. Espinosa and A. Satue and I. Hemaez: MCTY Baseline corpus: a bimodal biometric database, IEE Proceeding Vision, Image and Signal Processing, vol. 150, No. 6, 2003

2. Yoav Freund and Robert E. Schapire: A decision-theoretic generalization of on-line learning and an applicationto boosting. In Computational Learning Theory: Eurocolt '95, pages 23–37. Springer-Verlag, 1995.
3. S.Yamanaka, Masato Kawamoto, T.Hamamono, and S.Hangai,: Signature Verification Adapting to Intersession Variability: IEEE International Conference on Multimedia and Expo 2001, Tokyo Japan (2001) 88-91.
4. Y. Hongo, D. Muramatsu, and T. Matsumoto: AdaBoost-based on-line signature verifier: SPIE Defense and Security Symposium, Orlando, Florida USA (2005) Proceeding vol. 5779 Biometric Technology for Human Identification II 373-380.

MOC via TOC Using a Mobile Agent Framework

Stefano Bistarelli[1,2], Stefano Frassi[2], and Anna Vaccarelli[2]

[1] Dipartimento di Scienze, Università "G. D'Annunzio" di Chieti-Pescara, Italy
bista@sci.unich.it
[2] Istituto di Informatica e Telematica, CNR, Pisa, Italy
{stefano.bistarelli,stefano.frassi,anna.vaccarelli}@iit.cnr.it

Abstract. A novel protocol is proposed to address the problem of user authentication to smartcards using biometric authentication instead of the usual PIN. The protocol emulates expensive *Match On Card* (MOC) smartcards, which can compute a biometric match onboard, by using cheap *Template on Card* (TOC) smartcards, which only store a biometric template. The biometric match is performed by a module running on the user's workstation, authenticated by a mobile agent coming from a reliable server. The protocol uses today's cryptographic tokens without requiring any HW/SW modifications.

1 Introduction

Smartcards are currently used as a secure and tamper-proof device to store sensitive information such as digital certificates and private keys. Access to smartcards has historically been regulated by a trivial means of authentication: the Personal Identification Number (PIN). A user gains access to a card if he/she enters the right PIN. Experience shows that PINs are weak secrets in the sense that they are often poorly chosen, and that they are easy to forget.

Biometric technologies have been proposed to strengthen authentication mechanisms in general by matching a stored biometric template to a live biometric template [1, 2]. In the case of authentication *to* smartcards, intuition imposes the match to be performed by the smartcard chip. However, this is not always possible because of the complexity of biometric information such as fingerprints or iris scans, and because of the still limited computational resources offered by currently available smartcards.

In general, three strategies of biometric authentication can be identified.

Template on Card (TOC). The biometric template is stored on a hardware security module (smartcard or USB token). It must be retrieved and transmitted to a different system that matches it to the live template acquired from the user by special scanners. Cheap memory-cards with no or small operating systems are generally sufficient for this purpose.

Match on Card (MOC). The biometric template is stored on a hardware security module, which also performs the matching with the live template. Therefore, a microprocessor smartcard is necessary, which must contain an operating system running a suitable match application.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 464–473, 2005.
© Springer-Verlag Berlin Heidelberg 2005

System on Card (SOC). This is a combination of the two previous technologies. The biometric template is stored on a hardware security module, which also performs the matching with the live template, and includes the biometric scanner to acquire, select, and process the live template.

Clearly, the third of the strategies sketched out above is the best in terms of security as everything takes place on card. Embedding a biometric reader on a smartcard offers all the privacy and security solutions but, unfortunately, it is expensive and presents more than one realization problem.

The benefits derived from MOC cards are valuable in themselves: using its own processing capabilities the smartcard decides if the live template matches the stored template closely enough to grant access to its private data. Nevertheless this scheme presents a danger: we have no certainty that a biometric reading has been collected through live-scan and there is the risk of an attacker's sniffing the biometric and later using it to unlock the card in a replay attack.

In the present setting, how can we implement biometric authentication on smartcards that are already commercially available?

We address this issue by developing a novel protocol that employs inexpensive TOC cards as if they were MOC cards and that counterbalances the MOC technology's drawbacks; the requirements of the present work are to employ common crypto smartcards without modifying the code inside them and without asking the user directly for the PIN.

This paper is organized in the following way: Section 2 illustrates the security problems using TOC for authenticating a user to a smartcard. Section 3 sketches out the adopted solution. The protocol is introduced in Section 4, while implementation details are described in Section 5. Section 6 illustrates the solved/unsolved security problems. Finally, Section 7 concludes the paper and proposes possible future works.

2 Security Problems Using TOC Technology

Before describing the protocol, we would like to explain some problems related to the use of TOC technology for authenticating a user to a smartcard (SC). There are several points of attack in the use of TOC technology without securing the data transmission between the biometric device, the smartcard reader and the local host that carries out the biometric match. Consequently, we have to consider some aspects before designing our secure protocol.

The idea of using TOC technology to authenticate a user to a SC (without security concerns) is:

1. A cryptographic application asks the user to authenticate himself to the SC via a specific API call.
2. Verification Module reads the biometric template from the SC.
3. A real time template is acquired from the user using a biometric scanner.
4. A biometric match between the two templates is performed on the local host.
5. If the biometric match is successful then the actual PIN is submitted to the card to unlock the crypto chip.

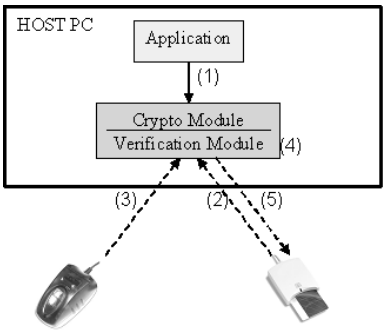


Fig. 1. A TOC Protocol

A diagram of the protocol is illustrated in Fig. 1.

The major problems of the above protocol are:

- The enrollment template stored inside the smartcard could be eavesdropped at step 2.
- The real time template could be eavesdropped at step 3.
- The smartcard doesn't trust the module performing the match at step 4.
- Where to store the secret login data (PIN) used to actually log the user into the cryptographic chip (we don't want to ask the user for it).

The first two points are critical if it is possible for an attacker to use the smartcard (after he/she has stolen it) for a replay attack sending again the eavesdropped data directly to the verification module. In this case, there is no security mechanism to verify that the biometric verification data are derived from an actual live presentation to the biometric sensor. To solve these problems we could encrypt the data exchanged between the devices and the crypto library or find a way to trust the module that acquires the live template.

SC is unable to authenticate the verification module. Maybe using a kind of challenge-response protocol: the module requests a random to the SC and this is returned encrypted with a shared secret key. Now the dilemma is where to store the key on the host.

The only method to unlock the private area of the chip is to supply the exact PIN to the SC. If the PIN is correct then the SC trusts the module that has performed the local biometric match.

Therefore, the crucial point is the last one: where to store the secret PIN[1]. An obvious method is to store the PIN inside the compiled Crypto Module. This is not a good solution because a malicious user might do reverse engineering on the library and find the secret. Every place inside the user's file system is not secure if a malicious program has manipulated the host, so the safest place is inside a protected remote Server.

[1] In a previous work [3], a similar problem has been investigated developing a comparable protocol. In that case, the user was asked directly for the PIN and there was the need to install a piece of code into the smartcard to carry out the protocol

3 The Adopted Solution

Now the problem to solve is how the remote Server can send critical data to an unknown remote host. This is like a black box and the server does not know if a malicious entity is running on that client.

A mutual authentication by establishing a SSL connection between the client and the server is a good solution, but like the PIN, there is always the recurrent problem of where to store the certificate/private key of the client (we don't want to use another smartcard [11] and we want to avoid asking the user for another PIN to unblock this private key).

We chose to adopt another solution: using a Mobile Agent framework. If the server cannot trust applications running on the client, it will trust the code that it launches to the client: a mobile agent.

A remote agent, launched from the secure server, will try to authenticate the module that executes the biometric match on the client; if the result of the authentication is positive then the server sends the secret PIN to the client via a previously opened secure connection.

The chosen framework was **SeMoA** [6] (Secure Mobile Agents). It is a runtime environment for Java-based mobile agents in development at the Fraunhofer Institute for Computer Graphics, with its main focus on security.

A Mobile Agent is a software entity that is not bound to the host where it begins execution, but has the unique ability to travel across a network and perform tasks on machines that provide agent-hosting capability. Unlike remote procedure calls, where a process invokes procedures of a remote host, process migration allows executable code to travel autonomously and to interact with the hosting machine's resources, including other mobile agents. Therefore, a Mobile Agent framework has to cope with various security threats [8]: malicious agents might try to break into the server in order to harm other agents or to gain unauthorized system access. A malicious host could tamper with agents. Agents might be sniffed while they are transferred over the network.

Many open source agent development frameworks are available on the internet: we decided to adopt SeMoA because it focuses on security and tries to solve the above-mentioned problems.

4 Protocol Description

This section presents the protocol. It illustrates the interactions among the main entities (details will be introduced in section 5). Description of the entities:

- Client: the user's workstation
  - Application: the user application that requires the access to the smartcard through the Crypto Module (for instance a digital signature application).
  - Crypto/Verification Module: the main entity used to access the smartcard and to perform the local biometric match (it implements the client-side protocol).

  - SeMoA Framework: the runtime environment for the Mobile Agent coming from the Server.
- Server: the secure host where the secret login data is contained
  - SeMoA Framework: the runtime environment for the Service implementing the protocol.
  - MOC Service: the Service which accepts connections and implements the protocol.
  - MOC Agent: the Mobile Agent that is launched by Moc Service, gets to the Client and comes back with the result of the authentication.

A diagram can be defined as in Fig. 2.

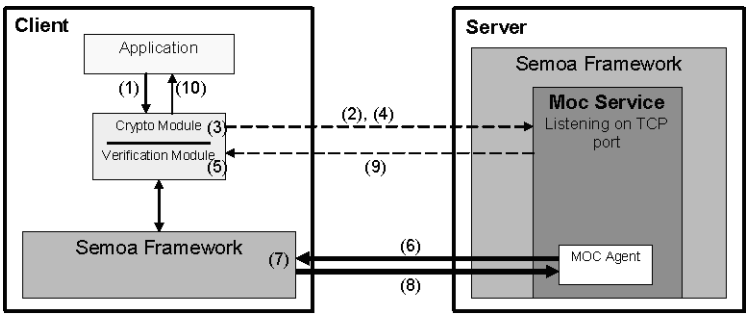


Fig. 2. The Protocol

1. User application requires access to the private space of the smartcard through a particular Crypto API call.
2. The Crypto Module (CM) opens an encrypted connection to the Moc Service ($MocS$) running on the Server ($MocS$ is authenticated via SSL).
3. After the connection has been established, CM generates a large Random value and stores it inside itself (value used by the mobile agent at step 7).
4. Finally, CM sends the Random and the smartcard Serial Number to $MocS$: if the subsequent controls succeed, CM will receive the secret login data at step 9, otherwise $MocS$ will close the connection.
5. In the meantime, the Verification Module executes a biometric match between the template stored inside the smartcard and the live acquired one.
6. $MocS$ stores the Random previously received within a MOC Mobile Agent and launches it to the Client address over a new encrypted channel; after this, $MocS$ begins to wait for the return of the Agent.
7. Now the MOC Agent is migrated on the client:
   - It tests the validity of the modules residing on the user's workstation checking their digital signature.
   - It checks that the random inside itself has the same value as the random contained in CM.
   - It ensures that the biometric match executed at step 5 is successful.

   If all the previous controls are positive then we can trust the CM module that has started the protocol.

8. The Agent comes back to the Server and returns the result to $MocS$.
9. If the result is positive then $MocS$ sends, on the same connection opened at step 2, the secret login data (PIN)[2] correlated to the Serial Number received at step 4. Otherwise, it closes the connection with the client.
10. In the last step, CM unlocks the private area using the received PIN and confirms to the user application the success of the Crypto API call made at step 1.

5 Implementation

The protocol and all the entities have been developed and deployed on a Windows 2000 Professional workstation, so some details are particular to this architecture. As regards the hardware, the biometric scanner employed is an FX2000 produced by Biometrika srl [12], while the smartcard used is a Cyberflex e-gate produced by Schlumberger [13]. (As we will see later, it is possible to employ any kind of biometric device or smartcard without modifying the protocol by only changing the respective library.)

The principal technology employed, besides SeMoA, is the **PKCS#11** standard [7], which has been used as the Crypto module. We have chosen this solution because this is the most widespread de-facto standard in today's cryptographic tokens. The PKCS#11 standard specifies an API, called "Cryptoki" (cryptographic token interface), to interface the devices that hold cryptographic information and that perform cryptographic functions. The Cryptoki is important because it isolates an application from the details of the cryptographic device.

The standard employed to perform all the required biometric operations is **BioAPI** [4]. This API is intended to provide a high-level generic biometric authentication model, covering the basic functions of Enrollment, Verification, and Identification.

Another technology we have used to implement this protocol has been the **Java Native Interface** (JNI) [9]. This was used to exchange data between the MOC Mobile Agent (which runs in a Java virtual machine) and the dlls (which are native libraries) at step 7 of the protocol.

Figure 3 describes the protocol in more detail. The previous client's Crypto-Verification module has been separated in four different dynamic link libraries (dll):

1. PKCS#11 module: this is the library provided by the smartcard manufacturer. This dll permits user authentication to the smartcard using the normal PIN. Therefore it is possible to switch from a smartcard brand to another by only changing this module.

[2] The PIN can be delivered directly by the Agent at the end of step 7 (only if all the checks are positive). We have avoided this solution because even though it is the fastest, it is the least safe too: the agent might be tampered with by a malicious entity, to extract the PIN

2. Crypto Wrapper: this is a dll wrapper to the PKCS#11 module; all the API calls are proxied to the manufacturer's dll except for the C_Login function: this is the point where the client-side protocol is implemented.
3. Verification module: it performs the local biometric match via the BioAPI library. Also in this case it is possible to use another Biometric device by only changing the Biometric Service Provider (BSP) dll.
4. JNI stub: this module is used by the Java Mobile Agent to access the Crypto Wrapper and the Verification module. It works with JNI.

Inside the server there is a certification authority (CA) which is used to issue certificates for the users. The CA also issues Attribute Certificates containing the enrolled biometric template [1]; they are stored in the smartcards along with the x509 user's certificates.

Every client's dll which performs the protocol is digitally signed [5] by the Server's private key, so that the components can mutually authenticate each other and the Mobile Agent can check their validity.

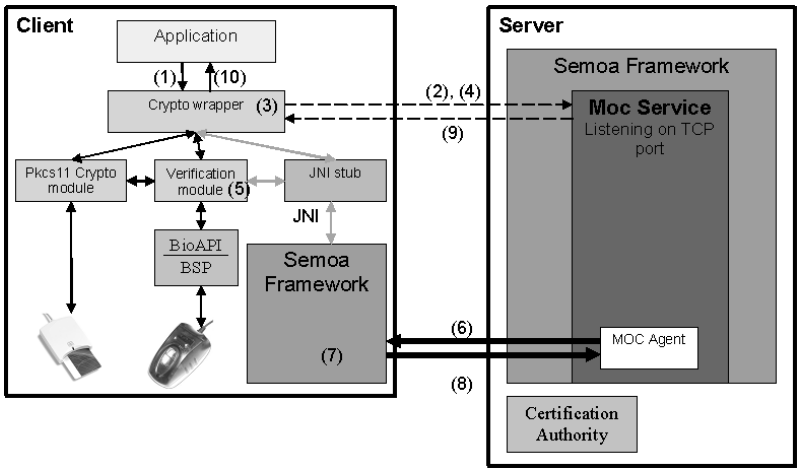


Fig. 3. The detailed Protocol

5.1 Detailed Protocol

1. The user application requires access to the smartcard via a C_Login call. (The application needs, for instance, to use the user's private key stored in the smartcard). The call corresponds to a C_Login(NULL), where NULL means that biometric authentication is requested (no PIN is given).
2. The Crypto Wrapper (CW) opens a SSL connection to the Moc Service ($MocS$) running on the Server. Naturally, an encrypted connection is used to avoid sniffing the data when the secret login PIN is sent over the channel. We use SSL server authentication to check the server's identity. (NO SSL client authentication, because there would be the recurrent problem of where to store the client's private key.)

3. After the connection has been established, the CW generates a large random value and stores it inside a Shared Data Section [10]: all the processes that will use this module, will access the same variable. In this way, we can detect possible malicious modules that try to start the protocol. (The only module that is permitted to start the protocol is the CW). The random will be checked by the mobile agent at step 7.
4. Finally, the CW sends the Random and the smartcard Serial Number[3] to $MocS$: if the subsequent controls succeed, the CW will receive the secret login data at step 9, otherwise $MocS$ will close the SSL connection.
5. In the meantime, the Verification Module (VM) executes a biometric match between the template stored in the smartcard and the live acquired one. VM reads the Attribute Certificate stored inside the smartcard, verifies its validity and extracts the biometric template. Then, VM acquires the live template from the scanner via the BioAPI module only if the BSP's digital signature is correctly verified.
6. $MocS$ generates a MOC Mobile Agent, signs it, and launches it to the Client address (the Random previously received has been stored inside the Agent); after this, $MocS$ begins to wait for the return of the Agent. Also in this case, the agent is sent over an encrypted channel.
7. Now the MOC Agent is migrated on the client. The SeMoA environment checks the digital signature of the Agent to see if it comes from the trusted Server; if so, then:
   - The Agent tests the validity of the dlls checking their digital signature (the dlls reside on the user's workstation at a precise path).
   - Using Java Native Interface, it checks that the random inside itself has the same value as the random contained in the CW. The Agent reads the random value using a new function created in the CW.
   - Using Java Native Interface, it ensures that the biometric match executed at step 5 is successful.

   The check of the random value has been employed to verify that the correct CW has started the protocol. If the random is different, it means that a malicious entity is trying to deceive the Server to steal the secret PINs.
   The digital signature of the modules is checked to ensure that only trusted dlls are carrying out the protocol. Finally, the last check verifies that the proper user is accessing the smartcard. If the above checks are positive only then we trust the CW module that has started the SSL connection.
8. The Agent comes back to the Server and returns the result to $MocS$.
9. If the result is positive then $MocS$ sends, on the same SSL connection opened at step 2, the secret login data (PIN) correlated to the Serial Number received at step 4. Otherwise, it closes the connection with the client.
10. In the last step, CW unlocks the private area using the received PIN (it performs a C_Login(PIN) calling the PKCS#11 module) and confirms to the user application the success of the C_Login call made at step 1.

[3] There is a DataBase installed inside the Server which contains the corresponding unique PIN for every serial number. The serial is used at step 9

6 Security Problems Resolved

If the attacker *has not stolen* the user's smartcard, he could try to get possession of the PINs residing in the safe Server:

- If no smartcard is inserted, the protocol will not start.
- If a wrong or malicious dll dealing with the communication with the SC is installed, the Mobile Agent will notice it.
- If a different module from CW tries to connect to $MocS$, the Mobile Agent will notice it using the random comparison.
- If he is using a brand new smartcard, with a proper serial number, the server will not return the PIN: the malicious user is not able to pass the local biometric match. (Inside the smartcard there is not an Attribute Certificate containing the fingerprint template issued by the Server's CA).

If the attacker *has stolen* the user's smartcard and the right dlls are installed:

- he could try to change the template stored in the SC: he cannot do this, because the biometric template is contained in an Attribute Certificate signed by the Server CA.
- Even if a biometric template has been previously sniffed, it is not possible to inject it within the Verification Module: before the VM acquires a template from the biometric device, it verifies that the Biometric Service Provider dll is the trusted one via digital signature (no more replay attacks).

6.1 The Two Feasible Attacks

The possible attacks to the implemented protocol concern how the PIN is transmitted in the final step, and the malicious host threats [8] in a mobile agent framework. In the first case, the PIN could be sniffed if the channel between the SC reader and the host is not protected. If the attacker has stolen the user's SC, he could bypass all the protocol and use only the manufacturer PKCS#11 library. We assume that this is not possible because we rely on the smartcard producer's Crypto Module (a trusted path between the SC and the host should be employed).

The other way to attack this protocol is modifying the SeMoA framework and/or the Java Virtual Machine on the client's workstation. This is a common problem in the Mobile Agent Systems field [8]. In our case, an attacker succeeds if he is able to alter the Mobile Agent's return value with a positive result even if the checks on the client are negative. This problem can be solved by employing a mutual authentication protocol between the client/server SeMoA environments, using a secure trusted hardware: it would be used to store the framework private key and to check the validity of the agent runtime environment.

7 Conclusions

Modern, inexpensive TOC smartcards cannot compute a biometric match like MOC smartcards. We have developed a protocol, which simulates the MOC strategy through the use of TOC cards. In practice, the actual match is delegated to a module of the card host after an authentication performed by a mobile agent coming from a secure server.

The design we have presented has been fully implemented using the SeMoA framework, which provides an open source mobile agent system, and through two de-facto standards such as PKCS#11 and BioAPI, respectively used to communicate with the crypto smartcard and to interact with the biometric functions. The use of these standards lead to an implementation where any smartcard and any biometric device can be used.

Potential future works will concern addressing the issues described in Section 6.1 (local PIN sniffing and malicious host attack), and adapting the protocol in other areas where the entity performing the biometric match is not trusted.

References

1. L. Bechelli, S. Bistarelli, F. Martinelli, M. Petrocchi, and A. Vaccarelli. Integrating biometric techniques with electronic signature for remote authentication. *ERCIM News*, (49), 2002.
2. L. Bechelli, S. Bistarelli, and A. Vaccarelli. Biometrics authentication with smartcard. *Technical Report 08-2002*, CNR, IIT, Pisa, 2002.
3. G.Bella, S. Bistarelli, and F. Martinelli. Biometrics to Enhance Smartcard Security (Simulating MOC using TOC). *Proc. 11th International Workshop on Security Protocols*, Cambridge, England, 2-4 April 2003.
4. BioAPI Consortium. BioAPI Specification Version 1.1. http://www.bioapi.org
5. Authenticode, *http://msdn.microsoft.com/workshop/security/authcode/signing.asp*
6. V. Roth and M. Jalali. Concepts and Architecture of a Security-centric Mobile Agent Server. *IEEE Proceedings of 5th International Symposium on Autonomous Decentralized Systems (ISADS01)*, pages 435-442, Dallas, Texas, March 2001.
7. RSA Laboratories. PKCS#11-cryptographic token interface standard.
8. E. Bierman and E. Cloete. Classification of Malicious Host Threats in Mobile Agent Computing. *Proceedings of SAICSIT 2002*, pages 141-148, 2002.
9. Java Native Interface, *http://java.sun.com/docs/books/jni/index.html*
10. How To Share Data Between Different Mappings of a DLL. Microsoft KB 125677
11. U. Waldmann, D. Scheuermann and C. Eckert. Protected transmission of biometric user authentication data for oncard-matching. *Proceedings of SAC 2004*, pages 425-430.
12. Biometrika srl, *http://www.biometrika.it*
13. Schlumberger, *http://www.axalto.com*

Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks

Norman Poh and Samy Bengio

IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
{norman,bengio}@idiap.ch

Abstract. This study investigates a new *confidence criterion* to improve fusion via a linear combination of scores of several biometric authentication systems. This confidence is based on the margin of making a decision, which answers the question, "after observing the score of a given system, what is the confidence (or risk) associated to that given access?". In the context of multimodal and intramodal fusion, such information proves valuable because the margin information can determine which of the systems should be given higher weights. Finally, we propose a *linear discriminative framework* to fuse the margin information with an existing *global* fusion function. The results of 32 fusion experiments carried out on the XM2VTS multimodal database show that fusion using margin (product of margin and expert opinion) is superior over fusion without the margin information (i.e., the original expert opinion). Furthermore, combining both sources of information increases fusion performance further.

1 Introduction

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. Compared to traditional authentication methods such as keys and PIN numbers, biometric authentication has the advantages that it is not susceptible to misplacement or forgetfulness. Unfortunately, its accuracy and reliability still need to be improved to make the system practical in day-to-day applications.

One way to increase its performance accuracy is to combine several biometric systems. In this paper, we show how multimodal or intramodal fusion BA system can be improved by using a new confidence measure based on margin. This quantity can be interpreted as "how confident we are that a given access is correct after observing the score". It is bounded between zero and one; when it is zero, a given access has 50% chance of being correctly classified. The greater the confidence, the higher the chance that the given access is correct. We show that this margin-derived confidence can be used in fusion of multimodal biometric systems. The margin-derived confidence can be used to *modify* the fixed decision boundary. This is done by a linear combination between the confidence-derived function and the fixed discriminative function. The former function is *adaptive*, i.e., it changes *after* observing the access scores. In contrast, the latter function is *fixed* once (hence non-adaptive) and applied to all accesses.

Improving fusion with quality has already been examined by several authors. Toh *et al.* [1] fused fingerprint and speech systems using a modified multivariate polynomial

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 474–483, 2005.
© Springer-Verlag Berlin Heidelberg 2005

regression function to take the quality information into account. Bigun *et al.* [2] also fused fingerprint and speech systems but using a statistical model (that reconciles expert opinions) modified to take the quality into account. Fierrez-Aguilar [3] fused fingerprint and speech systems, with quality derived from fingerprint, using a modified Support Vector Machine algorithm. Garcia-Romero *et al.* [4] considered quality in speaker authentication task using the first formant. Fusion is done so as to favour speech frames with high quality. Hence, instead of taking the average Log-Likelihood Ratio (LLR) over the entire utterance frames, a weighted LLR (by quality) is used. All these studies provide empirical evidences that *quality information can improve the performance* of single-modal and multimodal biometric systems.

We propose to derive a quality index based on margin. This margin is a function of False Acceptance and False Rejection Rates, which themselves are estimated from a set of expert scores. The main advantage of margin-derived quality is that no additional (and often independent) system is needed to estimate the quality, as compared to the previously mentioned approaches[1].

Section 2 presents the proposed idea of margin and compares it with existing margin definitions in the literature. Section 3 presents how confidence can be integrated with existing fusion functions. Section 4 presents briefly the 32 fusion problems based on the XM2VTS database and Section 5 discusses a pooled EPC curve as a performance visualisation tool. Experiments are reported in Section 6. This is followed by conclusions in Section 7.

2 Margin as Confidence

Given an acquired biometric feature $\mathbf{x}$, an opinion of a BA system $y(\mathbf{x})$ as a function of $\mathbf{x}$ and a preset threshold Δ, a biometric system makes its decision based on the following decision function:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise.} \end{cases} \tag{1}$$

Since $\mathbf{x}$ is present in $y(\mathbf{x})$ and variables derived from it, we simply write y instead of $y(\mathbf{x})$. The system may make two types of mistakes: false acceptance (FA) and false rejection (FR) as a function of threshold Δ. By tracing this function empirically from a development set, and normalising them using the total number of impostor and client accesses, respectively, one obtains the false acceptance rate (FAR) and false rejection rate (FRR) curve as a function of threshold Δ. FAR and FRR are defined as follows:

$$\text{FAR}(\Delta) = \frac{\text{number of FAs}(\Delta)}{\text{number of impostor accesses}}\,, \tag{2}$$

$$\text{FRR}(\Delta) = \frac{\text{number of FRs}(\Delta)}{\text{number of client accesses}}\,. \tag{3}$$

[1] The additional measurement system *may* provide additional degree of freedom to describe the biometric classes if the system output is *independent* of the original feature sets. However, in most situations, the additional system derives the quality information from the same feature sets as those used by the verification system, e.g., [1, 2]. Regardless of how the quality information is derived (from the feature sets or from the scores as proposed here), we conjecture that the quality information can provide better information regarding the separation decision

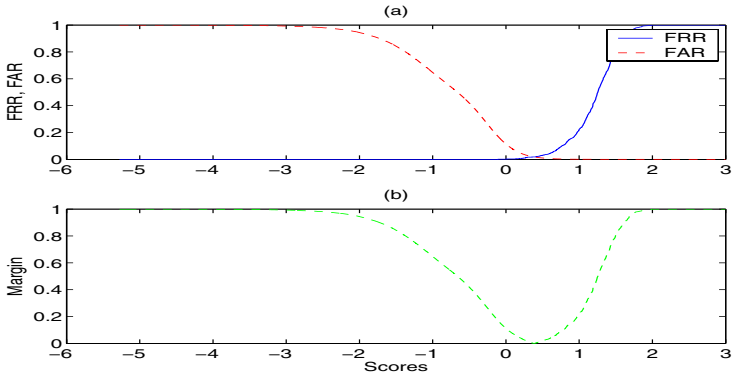


Fig. 1. (a) FAR and FRR as a function of the threshold in the score space. (b) The derived margin based on (a)

A commonly used point to examine the quality of performance is to evaluate the value FAR = FRR. This is the Equal Error Rate (EER) point and it assumes that the costs of FA and FR are equal, and that the class prior probabilities (of client and impostor distributions) are also equal.

The empirical procedure to find Δ that satisfies the EER criterion (on the training set) is:

$$\Delta^* = \arg\min_{\Delta} |\text{FAR}(\Delta) - \text{FRR}(\Delta)| \,. \tag{4}$$

We define the margin as:

$$\mathcal{M}(\Delta) = |\text{FAR}(\Delta) - \text{FRR}(\Delta)|. \tag{5}$$

By replacing Δ by y, we effectively evaluate the margin of the output y. FAR, FRR and margin are shown in Figure 1. The margin derived this way simply tells us how much confident we are given an opinion y. The further it is from the decision boundary Δ^*, the more confident we are. Note that because FAR and FRR are cumulative density functions, they are confined in the range $[0, 1]$. Hence, the margin defined here is also confined in the range $[0, 1]$. The additional scores that are needed to derive the margin function can either be obtained from *additional biometric data* or *cross-validated data* (not used to train the underlying systems) in case the additional data is not available.

Note that the margin defined here is different from the concept of margin in the boosting [5] or Vapnik's *margin slack variable* [6]. Several definitions of margin are defined in [7, Sect. 2]. Suppose that the target output is t_p and the output of a system is y_p for the p-th example. t_p takes on $\{-1, 1\}$, each representing a class (impostor or client here). Using this notation, margin in boosting for a given example p is:

$$margin(y_p) = \underbrace{(y_p - \Delta^*)}\, t_p, \tag{6}$$

whereas, Vapnik's margin slack variable for a given example p is:

$$\xi_p = \max(0, \gamma - \underbrace{(y_p - \Delta^*)}\, t_p), \tag{7}$$

where $\gamma > 0$ is known as *target margin* and is fixed *a priori*. Note that in our notation, the subtraction in the underbraced term $y_p - \Delta^*$ is to make sure that the decision boundary has a value of 0 (normally, the Δ^* has already been absorbed by the output of the system as a bias term; in our context, this bias term corresponds to $-\Delta^*$). Briefly, $margin(y_p)$ measures how far an example is from the decision boundary. The further it is, the better. Negative margin in this case implies wrong classification of example p. In Vapnik's margin, ξ_p measures how much example p fails to have a margin of γ from the hyperplane. If $\xi_p > \gamma$ then example p is misclassified by $y_p - \Delta^*$. The difference between Vapnik's margin slack variable and margin in boosting is that the former takes the target margin into account whereas the latter does not. Both of these margin definitions can only be calculated supposing that the target output (class-label) is known. In fact, they are used to select examples that are difficult to classify. They are only important during the training phrase. Our proposed definition of margin *does not* require the target output (although the margin function is constructed from a labeled training set). Furthermore, it is used exclusively during testing, which differs from the rest of the margin definitions. Perhaps the most remarkable difference is that this margin is based on FAR and FRR, with minimum at EER. The aforementioned margins are also valid but they do not optimise EER directly. Despite their different usages, one similarity among all these margins is that they all have to be derived from labeled (training) data.

In the next section, we will propose a method to incorporate the margin-derived confidence measure into an existing fusion function.

3 Combining *a Priori* Weights with Confidence

3.1 General Fusion Function

The most used form of fusion function in biometric authentication is perhaps a linear combination of several expert opinions passed through an activation function. Suppose y'_j is the j-th opinion and α_j is the weight associated to y'_j, respecting the constraint that $\sum_j \alpha_j = 1$. The combined opinion of M base experts, y_{COM} can be written as:

$$y_{COM} = f\left(\sum_{j=1}^{M} \alpha_j y'_j\right) \tag{8}$$

where f is an activation function. Suppose that there are N biometric systems but there are $M \geq N$ opinions. The number of opinions can be more than the number of systems because we assume here that each system can give more than one opinion, derived in one way or another. For instance, for the case of fusing two systems with output y_1 and y_2, we could have:

$$y'_j \in \{y_1, y_2, y_1^2, y_2^2, y_1 y_2, 1\}, \tag{9}$$

where 1 is a bias term, and

$$f(z) = \frac{1}{1 + \exp[-a(z-b)]}, \tag{10}$$

which yields a *polynomial logistic regression* function (with $a = 1, b = 0$). The full expansion of polynomial is exponential with respect to its degree. In [8], a reduced

polynomial expansion is used to reduce the complexity (the degree of freedom of the classifier) and to make it practical enough for fusion problems. When y'_j is defined as:

$$y'_j \in \{y_i | i = 1, \ldots, N\} \tag{11}$$

and using Eqn. (10) with $a = 1, b = 0$, one obtains a *logistic regression* function [9] In this study, we concentrate on the linear function f, i.e., $f(z) = z$ (a linear function) and establish a means to combine margin-derived confidence with a fixed discriminative function. We will show how the form of fusion in Eqn. (8) occurs naturally.

3.2 Fusion Function with Quality

In the literature, to the best of our knowledge, there are two forms to integrate the quality information with an *a priori* weight that modifies α_i in Eqn. (8). Suppose that w_j is the *a priori* weight (found by optimising Equal Error Rate, for instance) and q_j is the quality associated to y'_j. The two forms that incorporate the quality information are as follow:

$$\alpha_j \propto w_j + q_j \tag{12}$$

and

$$\alpha_j \propto w_j \times q_j \tag{13}$$

Note that in the absence of the quality information, we have $\alpha_j \propto w_j$. The usage of Eqn. (12) can be found in [1] using a reduced polynomial expansion of logistic regression function, i.e., using Eqn. (9) for the case of polynomial degree 2 and Eqn. (10). In the mentioned work, only polynomial up to degree 3 was examined. Experiments were conducted on fusion of fingerprint and speech biometrics with quality information obtained only from the fingerprint.

The usage of Eqn. (13) was found in [10, 11]. In [10], a speech expert ($j = 1$) and a lip expert ($j = 2$) were fused. Suppose that y_j^k is the j-th opinion given that the access is $k = \{C, I\}$, i.e., client or impostor. Suppose that y_j^k is generated from a normal distribution with mean μ_j^k and variance $(\sigma_j^k)^2$, i.e., $y_j^k \sim \mathcal{N}\left(\mu_j^k, (\sigma_j^k)^2\right)$. In [10], w_1 is defined as:

$$w_1 = \frac{\zeta_2}{\zeta_1 + \zeta_2} \tag{14}$$

where,

$$\zeta_j = \sqrt{\frac{(\sigma_j^C)^2}{NC} + \frac{(\sigma_j^I)^2}{NI}} \tag{15}$$

and NC is the total number of client accesses and NI is the total number of impostor accesses. By the summation constraint, $w_2 = 1 - w_1$. ζ_j is called the standard error. In [10], it was assumed that this error gives relative discrimination of an expert. High ζ_j indicates that expert j has high class dependent variance and hence, lower performance. As a result, its weight is lowered and the other expert's weight is increased[2]. q_j is defined as:

$$q_j \propto |\mathcal{M}_j^C(y_j) - \mathcal{M}_j^I(y_j)|, \tag{16}$$

[2] Although this criterion is valid, examining class-dependent variance is not sufficient; the mean difference is an important factor [12]

where

$$\mathcal{M}_j^k(y_j) = \frac{(y_j - \mu_j^k)^2}{(\sigma_j^k)^2} \tag{17}$$

for $k = \{C, I\}$ and $\sum_j q_j = 1$. Note that in this context, only the speech expert ($j = 1$) can be corrupted by noise whereas the lip expert ($j = 2$) stays intact. It was demonstrated experimentally [10] that under clean conditions, q_1 is relatively large (as compared to q_2) whereas under noisy conditions, q_1 is relatively small.

In [11], face and speech experts are fused and the speech expert is susceptible to noise whereas the face expert remains intact. The quality of the speech signal is estimated by using a statistical model (Gaussian Mixture Model) from the unvoiced part of speech frames. The unvoiced part of speech was obtained from the speech features right before an utterance begins. The output of the model (Log-Likelihood Ratio, LLR) is normalised into the range $[0, 1]$ by using a sigmoid function, as shown in Eqn. (10). a and b were tuned by heuristics, such that q_j is close to one for good quality speech and close to 0 for bad quality speech. According to the authors, the likelihood normalisation step is necessary because the normalised LLR is used directly to influence the *a priori* weight. $w_j|\forall_j$ are estimated using standard methods to minimise Equal Error Rate (EER), to be discussed in the later section.

We will use the method in Eqn. (12) because, as will be shown, it can be used to fuse different information sources. Furthermore, the multiplicative effect in Eqn. (13) can adversely influence α_j drastically as compared to Eqn. (12). To begin with, we consider a linear function of f, i.e., $f(z) = z$. We wish to fuse existing weight w_i with quality q_i for all $i = 1, \ldots, N$. Hence, α_i can be written as:

$$\alpha_i = \beta_{1,i} w_i + \beta_{2,i} q_i \tag{18}$$

where β_i control the contribution between the *a priori* weight w_i and the quality information q_i. Using $f(z) = z$, Eqns. (8) and (18), we obtain:

$$\begin{aligned} y_{COM} &= \sum_i (\beta_{1,i} w_j + \beta_{2,i} q_i) y_i \\ &= \sum_{m=1}^{N} \left(\underbrace{\beta_{1,m} w_m}\ \underbrace{y_m} \right) + \sum_{n=1}^{N} \left(\underbrace{\beta_{2,n}}\ \underbrace{q_n y_n} \right) \end{aligned} \tag{19}$$

where the four under-braces in Eqn. (19) can be written in the form of Eqn. (8). with y'_j defined by:

$$y'_j \in \{y_i, q_i y_i | i = 1, \ldots, N\}$$

Hence, fusion of *a priori* weight with the quality information can be performed by a linear combination of y_i and $q_i y_i$, for all i. The corresponding weights α_j can be found using standard methods such as Fisher-ratio or linear regression. The use of non-linear solutions is direct. For instance, one can use a Multi-Layer Perceptron with $y'_j|\forall_j$ as an input vector. Standard Support Vector Machine (SVM) algorithm with a polynomial kernel can also be used to classify the secondary features, thus, eliminating the need to create a dedicated classifier to fuse the quality information, as in [1] or to apply heuristics, as in [10, 11].

4 Database

The XM2VTS database [13] contains synchronized video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence. The database is divided into three sets: a training set, an evaluation set and a test set. The training set was used to build client models, while the evaluation set was used to compute the decision thresholds as well as other hyper-parameters used by classifiers and normalisation. Finally, the test set was used to estimate the performance. The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II (LP1 and LP2). The most important thing to note here is that there are only 3 samples in LP1 and 2 samples in LP2 for client-dependent adaptation and fusion training. Instead of reimplementing base experts and applying them on this database, we used scores from [14]. The score files are made publicly available and are documented in [15][3]. There are altogether 7 face experts and 6 speech experts for LP1 and LP2, respectively. By combining 2 baseline experts at a time according multimodal or intramodal fusion problems, 32 fusion experiments are further identified. The 13 baseline experiments have $400 \times 13 = 5{,}200$ client accesses and $111{,}800 \times 13 = 1{,}453{,}400$ impostor accesses. The 32 fusion experiments have $400 \times 32 = 12{,}800$ client accesses and $111{,}800 \times 32 = 3{,}577{,}600$ impostor accesses.

5 Evaluation Using Pooled EPC Curves

Perhaps the most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [16]. It has been pointed out [17] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [17] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [17] was proposed. We will adopt this evaluation method, which is also in coherence with the original Lausanne Protocols defined for the XM2VTS database. The criterion to choose an optimal threshold is called weighted error rate (WER), defined as follows:

$$\mathrm{WER}(\alpha, \Delta) = \alpha \mathrm{FAR}(\Delta^*) + (1-\alpha)\,\mathrm{FRR}(\Delta^*), \tag{20}$$

where FAR and FRR are False Acceptance Rate and False Rejection Rate, respectively. Note that WER is optimised for a given $\alpha \in [0,1]$. Let Δ^*_α be the threshold that *minimises* WER on a *development set*. The performance measure tested on an *evaluation set* at a given Δ^*_α is called Half Total Error Rate (HTER), which is defined as:

$$\mathrm{HTER}(\alpha) = \frac{\mathrm{FAR}(\Delta^*_\alpha) + \mathrm{FRR}(\Delta^*_\alpha)}{2}. \tag{21}$$

[3] Accessible at http://www.idiap.ch/∼norman/fusion

The EPC curve simply plots HTER versus α, since different values of α give rise to different values of HTERs. The EPC curve can be interpreted in the same manner as the DET curve, i.e., the lower the curve is, the better the performance but for the EPC curve, the comparison is done at a given cost (controlled by α). Furthermore, one can plot a pooled EPC curve from several experiments. For instance, in order to compare two methods over M experiments, only one pooled curve is necessary. This is done by calculating HTER at a given α point by taking into account all the false acceptance and false rejection accesses over all M experiments. The pooled FAR and FRR across $j = 1, \ldots, M$ experiments for a given $\alpha \in [0, 1]$ is defined as follow:

$$\text{FAR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \text{FA}(\Delta_\alpha^*(j))}{NI \times M}, \tag{22}$$

and

$$\text{FRR}^{pooled}(\alpha) = \frac{\sum_{j=1}^{M} \text{FR}(\Delta_\alpha^*(j))}{NC \times M}, \tag{23}$$

where $\Delta_\alpha^*(j)$ is the optimised threshold at a given α, NI is the number of impostor accesses and NC is the number of client accesses. FA and FR count the number of false acceptance and the number of false rejection at a given threshold $\Delta_\alpha^*(j)$. The pooled HTER is defined similarly as in Eqn. (21).

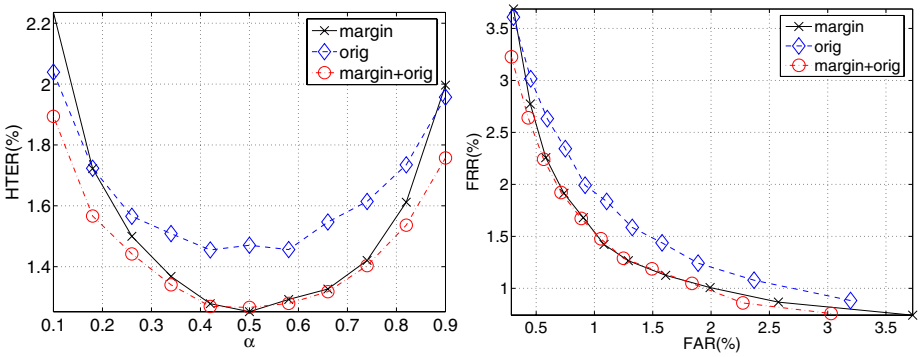
6 Experimental Results

Figure 2 shows both pooled EPC and ROC curves calculated from all 32×3 fusion experiments using original expert opinion ($y'_j \in \{y_i|\forall_i\}$), margin ($y'_j \in \{\mathcal{M}(y_i)y_i|\forall_i\}$) and both ($y'_j \in \{y_i, \mathcal{M}(y_i)y_i|\forall_i\}$). The ROC curves were plotted using FAR and FRR defined in Eqns. (22 and 23), whose *common* threshold was adjusted on a development (training) set. Note that for all these experiments, $\alpha_j|\forall_j$ were set to be equal. This reduces the fusion into the mean operator[4]. As can be seen, fusion with margin is better than the one using only the original expert opinions. Combining the two actually improves the performance even further. In fact, this improvement is significantly better than fusion using the original expert opinions across different α values according to the HTER significant test [18] with 95% of confidence. As a control experiment, we also performed fusion with $y'_j \in \{y_i, \mathcal{M}(y_i)|\forall_i\}$ using weighted sum. As expected, this approach does not improve the performance because $\mathcal{M}(y_i)$ does not contain any discriminative information. As a result, this control experiment is worse than using $y'_j \in \{y_i|\forall_i\}$ with EPC ranging between 1.5% and 3% of HTER (not shown here).

7 Conclusion

In this study, we proposed to use margin as a measure of confidence. When fusing two system opinions, their derived margins provide a relative information to which system

[4] In this database, weighted sum fusion with weights optimised using Fisher-ratio did not provide better performance than the mean operator



(a) EPC (b) ROC

Fig. 2. Pooled (a) EPC and (b) ROC curves of fusion experiments using original expert opinion (labeled as "orig"), product of expert opinion with margin (labeled as "margin"), and combination of both information (labeled as "margin+orig"), all using the mean operator. According to the HTER significant test, the "margin+orig" curve is always better than the "orig" curve, at different α, at 95% of confidence. These experiments were carried out on the XM2VTS database using 32 intramodal and multimodal fusion datasets, and each dataset contains the scores of two experts. Note that both (a) EPC and (b) ROC curves are *consistent* in that "margin+orig" is the lowest curve (for EPC) or closest to the origin (for ROC), implying the best generalisation performance among the three curves

is more important. This margin definition has the property that it is confined in the range $[0, 1]$, because it is derived from the distance between two cumulative density functions. Hence, margin can be used as a quality index. To the best of our knowledge, using margin to boost fusion has not been found in the literature yet. The second contribution of this work is the analysis of fusion function and how the quality information can be integrated with *a priori* weights of an existing fusion function. Suppose that y_i is the i-th opinion of an expert system and q_i is the associated quality. The fusion problem now can be treated as a fusion of $\{y_i, q_i y_i | \forall_i\}$. This has the same effect as modifying the *a priori* weight by adding q_i directly. 32×3 intramodal and multimodal fusion experiments were carried out on the XM2VTS multimodal database. Using pooled EPC curves (which summarise over each of the 32 experiments), we show that fusion using the confidence enhanced opinion $y_i q_i$ is better than using the original opinion y_i. Furthermore, combining the two, i.e., $\{y_i, y_i q_i\}$ improves the performance even further, and significantly, over different operating costs.

Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

1. K-A. Toh, W-Y. Yau, E. Lim, L. Chen, and C-H. Ng., "Fusion of Auxiliary Information for Multimodal Biometric Authentication," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 678–685.
2. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communnications," in *12th Int'l Conf. on Image Analysis and Processing*, Mantova, 2003, pp. 2–11.
3. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-Based Multimodal Biometric Verification Using Quality Signals," in *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, 2004, vol. 5404, pp. 544–554.
4. D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the Use of Quality Measures for Text Independent Speaker Recognition," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 105–110.
5. Y. Freund and R. Schapire, "A Short Introduction to Boosting," *J. Japan. Soc. for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.
6. V. N. Vapnik, *Statistical Learning Theory*, Springer, 1998.
7. N. Cristianini and J. Shawe-Taylor, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
8. K.-A. Toh, W.-Y. Yau, and X. Jiang, "A Reduced Multivariate Polynomial Model For Multimodal Biometrics And Classifiers Fusion," *IEEE Trans. Circuits and Systems for Video Technology (Special Issue on Image- and Video-Based Biometrics)*, vol. 14, no. 2, pp. 224–233, 2004.
9. Patrick Verlinde, Gerard Chollet, and Marc Acheroy, "Multimodal Identity Verification Using Expert Fusion," *Information Fusion*, vol. 1, no. 1, pp. 17–33, 2000.
10. T. Wark, S. Sridharan, and V. Chandran, "Robust Speaker Verification via Asynchronous Fusion of Speech and Lip Information," in *2nd Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'99)*, Washington, D.C., 1999, pp. 37–42.
11. C. Sanderson and K. K. Paliwal, "Noise Compensation in a Person Verification System Using Face and Multiple Speech Features," *Pattern Recognition*, vol. 36, no. 2, 2003.
12. N. Poh and S. Bengio, "How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks?," Research Report 04-18, IDIAP, Martigny, Switzerland, 2004, accepted for publication in *IEEE Trans. Signal Processing*, 2005.
13. J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.
14. N. Poh and S. Bengio, "Non-Linear Variance Reduction Techniques in Biometric Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 123–130.
15. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Research Report 04-44, IDIAP, Martigny, Switzerland, 2004, Accepted for publication in *AVBPA 2005*.
16. A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.
17. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
18. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.

A Classification Approach to Multi-biometric Score Fusion

Yan Ma[1], Bojan Cukic[2], and Harshinder Singh[1]

[1] Department of Statistics, West Virginia University, Morgantown, WV 26506
{yma,hsingh}@stat.wvu.edu
[2] Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506
cukic@csee.wvu.edu

Abstract. The use of biometrics for identity verification of an individual is increasing in many application areas such as border/port entry/exit, access control, civil identification and network security. Multi-biometric systems use more than one biometric of an individual. These systems are known to help in reducing false match and false non-match errors compared to a single biometric device. Several algorithms have been used in literature for combining results of more than one biometric device. In this paper we discuss a novel application of random forest algorithm in combining matching scores of several biometric devices for identity verification of an individual. Application of random forest algorithm is illustrated using matching scores data on three biometric devices: fingerprint, face and hand geometry. To investigate the performance of the random forest algorithm, we conducted experiments on different subsets of the original data set. The results of all the experiments are exceptionally encouraging.

1 Introduction

The use of biometrics for identity verification is becoming popular in many application areas such as border/port entry/exit, access control, civil identification and network security. It is well-known that no device capturing a single biometric trait works optimally in *every* application domain. Besides, unimodal biometric systems have limitations caused by noisy data, susceptibility to spoof attacks, instability of biometric characteristic due to environmental or physical factors [1]. By installing more than one biometric device and combining tests of several biometric traits, multi-biometric systems can overcome some of the limitations of single biometric devices and improve the small but significant failure rates of individual biometrics [2].

The critical issue in multimodal biometrics is to integrate the classification power of multiple devices, i.e., fuse information. Many fusion techniques have been proposed so far. These methods include: majority voting [3] [4] [5], Bayesian methods [3] [6] [8], logistic regression [3] [9], k-nearest neighbor [9], fuzzy integral [3] [10] [11], Dempster-Shafer theory [8], neural network [3] [12], classification tree [9] [13], linear discriminant function [13], sum rule [13] [14] [15], and

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 484–493, 2005.
© Springer-Verlag Berlin Heidelberg 2005

some simple combination techniques such as: min rule, max rule and product rule [14] [15]. Some of these schemes have been proved effective in improving the classification performance. However, there is no consensus on the *best* fusion technique.

Ross and Jain [13] experimentally explored two different approaches to information fusion: *combination* and *classification* [7] at matching scores levels. They indicated that the combination rule outperforms the classification approach. Classification trees and k-nearest neighbors are examples of the classification approach which finds the class label of an object on the basis of observed matching score vectors obtained from several modalities. The combination methods, such as the sum rule, derive a single scalar score from the matching score vector and the decision is based upon this single number.

The purpose of this paper is to contribute a novel application of random forests [18] in information fusion for user verification to investigate the performance potential of classification techniques. Information fusion from three biometric devices: face, hand geometry and fingerprint has been implemented. Random forests algorithm integrates information at the matching score level to build many tree classifiers, which subsequently form a "forest". The performance of random forest fusion algorithm is investigated using different training sets and independent testing sets.

The remainder of this paper is organized as follows. In the second section, the classification tree algorithm is described. Being an extension of the standard classification tree algorithm, the random forest algorithm is summarized in Section 3. Section 4 presents the experimental data. The system performance measures used in this paper are defined in Section 5. The methodology of the experiments is outlined in Section 6. Section 7 summarizes the experimental results.

2 Classification Trees

A classification tree is a tree-structured classifier built through a process known as recursive partitioning. The popular tree classifiers are CART, C4.5, QUEST, and FACT. Some of the classifiers create binary trees and some of them are able to generate multi-branch trees. All the classification tree algorithms focus on constructing a tree-like classification rule based on a given training data consisting of known class-labeled cases. But, different tree classifiers run different algorithms to handle issues in tree construction such as node splitting criterion, split stopping rule and pruning criterion.

Fig. 1 shows a tree generated by *rpart* program in *R* (http://www.r-project.org). The topmost node is called root which contains the entire sample of 10, 300 cases. The leaves of tree are called terminal nodes (represented by rectangles). They are tagged with class labels. Child nodes are formed by splitting their parent nodes. Each internal node "contains" a subset of the entire sample and also contains a rule which determines in which child node a particular case will fall. Following the rules specified by the tree, a case will finally reach one of the terminal nodes. The class label attached to such terminal node is assigned to each case that falls in it.

To choose the best split at a node, the algorithm searches through all the values at each variable. The main idea is to attain as homogeneous a set of labels as possible in each partition, so that the cases in each of the children nodes are more homogeneous than the cases in its parent node. Two well-known impurity based split selection methods are *gini index* and *entropy*. Both are measures of homogeneity of cases in a node. The smaller the measurement, the purer the cases; zero if all the cases belong to the same class. Suppose that a collection C consists of n cases from k classes. Gini index is defined as: $gini(C) = 1 - \sum_i p_i^2$, and the entropy is $entropy(C) = -\sum_i p_i log\ p_i$, $i = 1, 2, \ldots, k$, where p_i is the proportion of cases in C belonging to class i. In this paper, we rely on gini index as the split criterion to grow tree(s). The variable and cutoff for splitting a node are chosen so that the children nodes have as small a combined gini as possible. The same process is continued at the subsequent nodes and a full tree is generated.

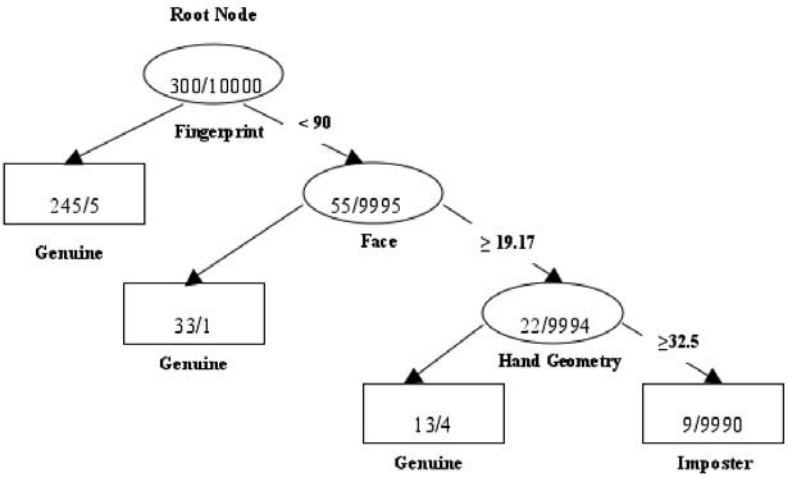


Fig. 1. A tree-structured classifier produced by R program for the data with binary predicted classes: genuine user and imposter. The numbers in each node are counts of genuine/imposter cases in that node. Underneeth each terminal node is the class label that dominates the cases in that node

The tree is grown to a point where the terminal nodes contain no more than a specified minimum number of cases or the significant majority of cases in the terminal nodes belong to the same class. Once a full tree is grown from the training data, the pruning starts. Pruning is a process of cutting back the tree branches to improve the predictive performance.

Classification tree algorithm has the advantage over other classification techniques because it provides an insight into the predictive structure of the data [21]. But it also has some disadvantages. One major problem associated with a classification tree is its instability, caused by the hierarchical structure of the tree [26]. Classification trees are very sensitive to small changes in the data set. A small

change or an error made in the root node split will be carried down through the whole tree, resulting in a different tree structure. Therefore, perturbations in the learning set can cause significant changes in the constructed tree [22]. One way to reduce the variability and improve accuracy is to grow an ensemble of heterogeneous trees [25], as explained below.

3 Random Forests

Random forests (RF) algorithm was developed by Leo Breiman (Breiman 2001). It is an ensemble method in the sense that instead of growing a single classification tree, we could build hundreds to thousands of trees. An improved classifier is obtained by integrating tree models in the forest. Each single tree is grown as follows [19]:

1. Take a bootstrap sample from the original data and the root node of the tree contains this sample instead of the original data.
2. At each node of the tree, except for the terminal nodes, randomly select a subset of the predictors, the locally optimal split is based on only this feature subset. Grow the tree as large as possible with no pruning.

Every tree in a forest of N trees represents a classification rule. Given a new case with matching score vector $\boldsymbol{x} = \{x_1, x_2, \ldots, x_P\}$, we begin with Tree 1 in the forest. The search starts from the root, the splitting rule is applied and the case is sent to one of the children nodes according to the rule. This is repeated until the terminal node is reached and the class label attached to the terminal node is assigned to this case. Thus, Tree 1 has made its decision. Then we go to Tree 2, follow the same procedure and find the class label for this case. Upon visiting N trees in the forest, we have N votes that the case belongs to either of classes. In a sense, each tree raises its own voice and "fights" with others to form the majority. Fig. 2 shows the construction of a random forest.

The procedure for growing a single tree outlined above randomize the selection of inputs in model building. Consequently, the trees will have different structures. Low correlation lowers the classification error rate of RF [20]. The first source of randomization is called bootstrap aggregating or, simply, bagging. A bootstrap sample is a random sample taken from the the original dataset with replacement. By taking bootstrap samples of the training data, multiple versions of a classifier are formed [22]. Different trees are built from different bootstrap samples. For each bootstrapped sample, about one-third of the cases are not used in the tree construction. These left-out cases are called out-of-bag (OOB) cases. They play an important role in algorithm's performance assessment. The random feature set selection at each node is another source of randomization in RF algorithm. In the standard classification tree algorithms, such as CART, the best split at a node is obtained by searching through all available predictors. In RF, the split at each node of a tree is only based on a random subset of the predictors. A new set of attributes is selected randomly for every split performed in RF nodes.

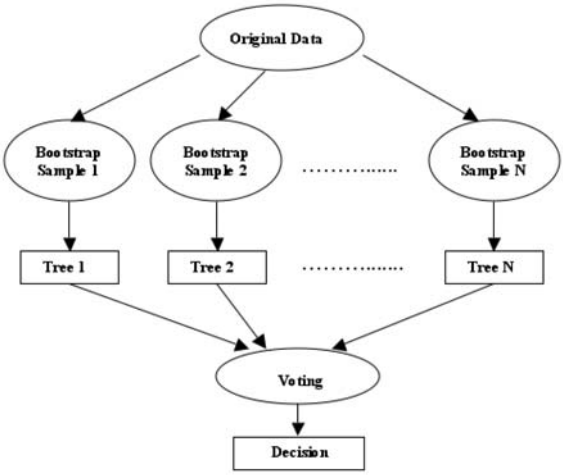


Fig. 2. Construction of Random Forest

Random forest algorithm overcomes instability, the main disadvantage of the classification tree algorithm. Voting among trees built from different bootstrap samples of the original data stabilizes the classifier and improves the performance of the forest over a single tree. The advantages of random forest include [17] [18] [19]: ease to use; robustness with respect to noise; faster tree construction due to the absence of tree pruning; built-in cross validation by using OOB cases to estimate test error rates; most importantly, high levels of predictive accuracy without overfitting.

From a single run of random forest, we can obtain a wealth of information such as classification error rate, variable importance ranking and proximity measures. A detailed description can be found in [19].

In RF, the OOB cases can be used to obtain an unbiased estimate of the test set error [19]. For each single tree in the forest, OOB cases are evaluated by the tree to test its classification capabilities. The test case is classified into the class having the most votes from individual trees. Comparison between this classification result and the known class label of the case provides an estimate of the error rate on the test set. An unbiased estimate of the misclassification rate is thus obtained automatically as a part of the run of the RF classification algorithm.

4 Experimental Database

The database of multimodal matching scores was collected at Michigan State University[1]. There are 500 genuine scores and 12,250 imposter scores obtained from each of three modalities: face, hand geometry and fingerprint. More information about this data set can be found in [13]. The data frame consists of

[1] Data used with permission

values from three attributes and a class label $Y = \{\text{Genuine, Imposter}\}$ for each case. In our process of building the random forest classifier, unlike in combination approaches, no data normalization or transformation is applied. We used original matching scores to build the model.

Most of the classification algorithms in use aim at minimizing overall error rate [16]. With an imbalanced data set, the algorithm tends to keep the misclassification error rate low on the large class while letting the smaller class have a relatively higher error rate. A balanced random forests algorithm can always be generated with proper sampling technique [23].

Four experiments are reported in this paper. The data sets used in each experiment are random samples from the original dataset (see Table 1). For each experiment, two types of tests were performed: *internal testing* and *external testing*. In internal testing, OOB cases were used to obtain an unbiased estimate of the misclassification rates internally during the run of the RF algorithm. In external testing, a data set disjoint from the training set is randomly selected and used for performance evaluation. Comparison between such classification rules and the true class labels produces an estimate of the error rates. We call these error rates external testing rates. In Table 1 and in the reminder of this paper, external testing is also called "separate" testing, to indicate separation between score samples in training subsets and evaluation subsets.

Table 1. Experimental Data Sets Description. Balanced sets have equal number of genuine and imposter cases

Experiment 1: *Internal testing set 1:* random sample of size 200 (balanced set) *Separate testing set 1:* random sample of size 800 (balanced set)
Experiment 2: *Internal testing set 2:* random sample of size 400 (balanced set) *Separate testing set 2:* random sample of size 600 (balanced set)
Experiment 3: *Internal testing set 3:* random sample of size 600 (balanced set) *Separate testing set 3:* random sample of size 1,000 (200 genuine and 800 imposter scores, disjoint from the internal test data)
Experiment 4: *Internal testing set 4*: random sample of size 1,250 (250 genuine and 1,000 imposter scores) *Separate testing set 4:* random sample of size 3,250 (250 genuine and 3,000 imposter scores, disjoint from the internal test data)

5 System Performance Measures

To evaluate the performance of random forest classifier in biometric score fusion, measurements such as Genuine Accept Rate (GAR) and False Accept Rate (FAR) are used. GAR is the probability that a genuine individual is (correctly)

accepted by the multi-biometric system. FAR measures the chance of an intruder being falsely accepted by the system as genuine. As with any biometric system, we prefer classification performance achieving a high GAR and a low FAR.

6 Methodology

Random forests algorithm integrates information at the matching score level as it builds each single tree in the forest. The construction of a forest is a random process. This results in the most significant drawback of random forests - its lack of reproducibility. We conducted four different experiments. To better investigate the performance of random forest classifier, we repeated each experiment 20 times. A 100 cutoffs were used to produce the $GARs$ and $FARs$ for each instance of an experiment.

We used R program which implements Breiman's random forest algorithm. Random forest algorithm avoids overfitting, meaning that the error rate stabilizes as more trees are added to the forest [24]. In our experiments, 500 trees generated sufficient overall classification accuracy in all experiments.

Prediction of a class that an individual case belongs to is determined either by majority voting between the trees (Default setting, cutoff = (0.50, 0.50)), or by user-defined thresholds. In random forest algorithm, *cutoff* is a vector of length equal to the number of classes. In the context of biometrics, *cutoff* is a two element vector. The "winning" class for a given case is the one with the maximum difference between the proportion of the votes and the cutoff. By majority voting rule, the class which wins at least 50% of the votes cast by the trees is the winner.

User-specified thresholds are more flexible. Suppose the *cutoff* is defined as a vector (c_1, c_2). Let the proportion of votes for two classes $\{1, 2\}$ be (p_1, p_2), where $p_1 + p_2 = 1$. If $p_1 - c_1 > p_2 - c_2$, or, in a different notation, if $p_1 > \frac{1+c_1-c_2}{2}$, then class 1 is the overall RF decision for the given case. User-specified cutoffs are very useful when the misclassification cost varies significantly among classes [17]. In many biometric applications, such as financial services, access control to sensitive spaces and criminal identification, the misclassification cost is heavily unbalanced. In such cases, falsely accepting an imposter might imply very high risks (or costs or damages). For example, if it is very expensive to misclassify a case as belonging to class c_1, then by assigning a low threshold to class c_1, we could reduce the overall misclassification cost.

7 Experiments and Results

Due to the space constrains, we do not report the experimental results at all thresholds within each trial. One of the best thresholds for each experiment and the majority voting scheme are reported. The thresholds are the *best* in a sense that they yield the largest sum of sensitivity (GAR) and specificity (1-FAR). As mentioned above, at each cutoff point, the experiment is conducted 20 times. Consequently, 20 internal and separate test results are generated at

Table 2. Random Forests Algorithm Results - Experiment 1

Cutoff	Quartiles	Internal Testing GAR (%)	Internal Testing FAR (%)	Separate Testing GAR (%)	Separate Testing FAR (%)
If $P_G > 50\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.00	99.75	0.25
	Median	100.00	0.00	99.75	0.50
	Minimum	100.00	0.00	99.50	1.00
If $P_G > 47\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.00	100.00	0.25
	Median	100.00	0.00	99.75	0.50
	Minimum	100.00	0.00	99.50	0.50

Table 3. Random Forests Algorithm Results - Experiment 2

Cutoff	Quartiles	Internal Testing GAR (%)	Internal Testing FAR (%)	Separate Testing GAR (%)	Separate Testing FAR (%)
If $P_G > 50\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.50	100.00	0.00
	Median	99.50	0.50	100.00	0.00
	Minimum	99.50	1.00	100.00	0.67
If $P_G > 53\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.00	100.00	0.00
	Median	99.50	0.50	100.00	0.00
	Minimum	99.50	1.00	100.00	0.33

Table 4. Random Forests Algorithm Results - Experiment 3

Cutoff	Quartiles	Internal Testing GAR (%)	Internal Testing FAR (%)	Separate Testing GAR (%)	Separate Testing FAR (%)
If $P_G > 50\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.00	100.00	0.75
	Median	100.00	0.33	100.00	0.88
	Minimum	99.67	0.33	100.00	0.88
If $P_G > 47\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	100.00	0.00	100.00	0.75
	Median	100.00	0.33	100.00	0.88
	Minimum	100.00	0.67	100.00	0.88

any given threshold. The sums of sensitivity and specificity were calculated by internal testing to rank the cutoff values. At each cutoff, 20 experiments result in 20 trials. Only the best rate (maximum), the most typical (median) rate and the worst (minimum) rate are reported herein. We similarly report the results obtained by separate (external) testing methodology. Tables 2 - 5 summarizes the results. Cutoffs are stated in terms of proportion of votes in favor of genuine, denoted by P_G.

The results we obtained are clearly very encouraging. In many cases, even the experiments resulting in minimum performance appear to be better than the ones obtained by the combination approaches on the same data set [13]. However, given that random forests do depend on training, it is possible that a somewhat different performance results could be obtained from other biometric matching score data sets. Certainly, we plan to perform additional studies and a more comprehensive evaluation in the near future.

Table 5. Random Forests Algorithm Results - Experiment 4

Cutoff	Quartiles	Internal Testing GAR (%)	Internal Testing FAR (%)	Separate Testing GAR (%)	Separate Testing FAR (%)
If $P_G > 50\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	99.60	0.10	98.80	0.20
	Median	99.60	0.10	98.80	0.20
	Minimum	99.20	0.10	98.80	0.23
If $P_G > 44\%$, vote for genuine; Otherwise, vote for imposter.	Maximum	99.60	0.10	98.80	0.20
	Median	99.60	0.10	98.80	0.20
	Minimum	99.60	0.30	98.80	0.23

References

1. Jain, A., Ross A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology. Special Issue on Image- and Video-Based Biometrics (2003)
2. http://www.cs.rit.edu/~jct9345
3. Lee, D., Srihari, S.N.: Handprinted Digit Recognition: A Comparison of Algorithms. In the Proceedings of the 3rd International Workship on Frontiers in Handwriting Recognition, 153-162, Buffalo, NY (1993)
4. Lam, L., Suen, C.Y.: Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans (1997) Vol. 27, No. 5
5. Zuev, Y., Ivanon, S.: The Voting as a Way to Increase the Decision Reliability. In: Foundations of Information/Decision Fusion with Applications to Engineering Problems, Washington, DC (1996) 206-210
6. Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles. Addison-Wesley Publishing Co., Reading MA (1981)
7. Nandakumar, K., Jain, A., Ross, A.: Score Normalization in Multimodal Biometric Systems. Available at: http://biometrics.cse.mse.edu
8. Xu, L., Krzyzak A., Suen, C.Y.: Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition. IEEE Transactions on Systems, Man, and Cybernetics (1992) Vol. 22, No.3
9. Verlinde P., Chollet, G.: Comparing Decision Fusion Paradigms Using k-NN Based Classifiers, Decision Trees and Logistic Regression in a Multimodal Identity Verification Application. In: Proceedings of the 2nd International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA), Washington DC (1999) 189-193
10. Tahani, H., Keller, J.M..: Information Fusion in Computer Vision Using the Fuzzy Integral. IEEE Transactions on Systems, Man and Cybernetics (1990) Vol. 20, No.3, 733-741
11. Lipnickas, A.: Classifiers Fusion with Data Dependent Aggregation Schemes. 7th International Conference on Information Networks. Systems and Technologies ICINASTe-2001
12. Ceccarelli, M., Petrosino, A.: Multi-feature Adaptive Classifiers for SAR Image Segmentation. Neurocomputing, Vol. 14 (1997) 345-363
13. Ross A., Jain, A.: Information Fusion in Biometrics. Pattern Recognition Letters 24 (2003) 2115-2125

14. Kittler, J., Hatef, M., Duin, R., Matas, J.: On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence (1998) Vol. 20, No. 3
15. Snelick, R., Indovina, M., Yen, J., Mink, A.: Multimodal Biometrics: Issues in Design and Testing. In Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, Canada (2003)
16. Chen, C., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data. Available at: http://stat-www.berkeley.edu/users/chenchao/666.pdf
17. Remlinger, K.S.: Introduction and Application of Random Forest on High Throughput Screening Data from Drug Discovery. Available at: http://www4.ncsu.edu/~ksremlin
18. Breiman, L.: Random Forests. Machine Learning (2001) Vol. 45, 5-32
19. Breiman L., Cutler, A.: Random Forests: Classification/Clustering. Available at: http://www.stat.berkeley.edu/users/breiman/RandomForests (2004)
20. Breiman, L.: Wald Lecture II, Looking Inside the Black Box. Available at: http://www.stat.berkeley.edu/users/breiman
21. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees, Wadsworth (1984)
22. Breiman, L.: Bagging Predictors. Machine Learning (1996) Vol. 24, 123-140
23. Liaw, A., Chen, C., Breiman, L: Learning From Extremely Imbalanced Data With Random Forests. Computational Biology and Bioinformatics, 36^{th} Symposium on the Interface, Baltimore, Maryland (2004)
24. Oh, J., Laubach, M., Luczak, A.: Estimating Neuronal Variable Importance with Random Forest. In: Proceedings of the 29th Annual Northeast Bioengineering Conference, NJIT, Newark, NJ (2003)
25. Speed, Terry (ed.): Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC (2003)
26. http://www.statsoftinc.com/textbook/stclatre.html

A Generic Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects

Sonia Garcia-Salicetti, Mohamed Anouar Mellakh,
Lorène Allano, and Bernadette Dorizzi

Département Electronique et Physique, Institut National des Télécommunications
9 Rue Charles Fourier, 91011 Evry France
{Sonia.Salicetti,Mohamed.Anouar_mellakh,Lorene.Allano,
Bernadette.Dorizzi}@int-evry.fr

Abstract. We propose in this paper a methodology for multibiometric systems evaluation on databases of virtual and real subjects of limited size (about 100 persons). Our study is limited to two biometric traits (modalities) that are a priori mutually independent, namely on-line signature and voice. Experiments are conducted on bimodal data of real subjects of the BIOMET database [9] and on several databases of virtual subjects constructed from BIOMET.

1 Introduction

The evaluation of a multibiometric system is not an easy task: indeed, there are very few available multimodal databases (M2VTS [1,2], XM2VTS [3,4], BANCA [5,6], DAVID [7], SMARTKOM [8]), most of which contain only two biometric modalities, usually face and voice. Also, multimodal databases available nowadays contain only about a hundred subjects, which makes difficult to extrapolate the success of a multimodal algorithm or method when being tested on a large population (thousands or millions of people). Moreover, multimodal databases more recently constructed as BIOMET [9], or under construction [10] have the tendency to contain more modalities (4 or 5) but not more subjects. In this precise matter, the order of magnitude of such databases remains indeed in about one hundred subjects. This can be explained by the fact that acquiring multimodal data is more time consuming and expensive than acquiring data from a single modality, and rises some other problems as higher acquisition failure and critical personal data protection. Indeed, acquisition failure is generated because the more modalities there are, the more it is likely that a data sample cannot be acquired in a given modality, thus generating the loss of a complete multimodal sample. This phenomenon is of course amplified whenever several sessions are recorded. Also, regarding personal data protection, the fact that a data collection may contain together fingerprints, signature, iris, and face, among others, of a given person, is obviously critical and not easily acceptable for donators which can be afraid of misuse or forgeries.

Many works in the multimodal fusion literature give results on about 100 real subjects, with no insight in the fact that such results may be in fact very biased. We address this problem in the present work and propose a new protocol for multibiometric systems evaluation on limited size databases of real subjects.

Moreover, it is also natural to study the possibility of using databases of virtual subjects, that is an individual generated by combining different biometric traits (mo-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 494–502, 2005.
© Springer-Verlag Berlin Heidelberg 2005

dalities) that belong to different persons. This procedure would simplify multimodal data construction because it would be sufficient to merge two or more databases of approximately the same number of subjects, containing each different modalities, to generate a multimodal data corpus containing more modalities. Although this question is crucial for the progress of research in multimodal fusion, few works have exploited the creation of virtual subjects for multimodal fusion [10,11]. The first question that arises is: which is the validity of this procedure? Then the next question is: if it is valid, which methodology should be used to evaluate multimodal systems on a given corpus of virtual subjects? Our aim in this work is also to answer to such crucial questions.

To that end, our methodology has been to create virtual subjects with data coming from a multimodal database of real subjects, that is the BIOMET database [9]. This permits us to do a comparative study of the behaviour of a bimodal fusion system (on-line signature and voice) on the real subjects and on several databases of virtual subjects generated from BIOMET. Indeed, the originality of this work is that we set the problem of using virtual subjects for systems evaluation relatively to the use of real subjects in multimodal databases. This gives more insight into what is in fact a real subjects database relatively to a virtual subjects one, and how evaluation should be performed in both cases.

As mentioned, our work is limited to two modalities, voice and on-line signature, combined in a previous work [12]. The choice of the modalities is a delicate question since it rises the problem of their mutual dependence/independence. We focus here in the combination of modalities that are a priori mutually independent, since it is only in this framework that we may consider building a virtual subject.

We combine such two modalities by a Support Vector Machine classifier with a linear kernel [13], a statistical technique that allows to learn the coefficients of a hyperplane and does not necessitate a priori scores normalisation. Actually, the objective of the present work is not to compare different classifiers. We show in this framework that a bimodal (voice, signature) database of real subjects of limited size (around 100 persons) introduces a bias when evaluating the fusion system, because the size of the database does not permit to represent all the possible data variability in the bimodal sense. Moreover, we show that using databases of virtual subjects is equivalent in certain conditions (with a given protocol) to the use of a database of real subjects of limited size. We provide here an evaluation protocol on both types of databases.

In the following, both experts (voice and signature) and the fusion method are first described (section 2), the experimental setup on BIOMET bimodal data is given in section 3, section 4 details the creation of virtual subjects from BIOMET bimodal data and section 5 focuses on comparative fusion experiences on real and on virtual subjects. Finally, conclusions and perspectives of this work are given in section 6.

2 Fusion of On-Line Signature and Voice

This study is carried out on a bimodal fusion system composed of two mono-modal biometric systems: a signature verification system described in [14] and a text-independent Speaker Verification system already described in [12]. We briefly describe in the following the two systems and the fusion method.

2.1 The Signature Verification System

As described in [14], each writer's signature is modelled by a continuous left-to-right HMM [15], characterised by a given number of states with an associated set of transition probabilities among them, and, in each of such states, a continuous density multivariate Gaussian mixture. The topology of the HMM only authorises transitions from each state to itself and to its immediate right-hand neighbours. An optimal number of states is computed for each writer and a personalised feature normalisation (of 25 features) is carried out to improve the quality of the modelling. The system exploits a fusion strategy of two complementary information provided by both the HMM likelihood and a "segmentation vector" obtained from the Viterbi path of the HMM modelling a given writer. As shown in [14], the combination of such two information permits to better separate the genuine and impostor distributions, thus improving significantly writer verification results.

2.2 The Text-Independent Speaker Verification System

This system is detailed in [12]. Considering a simple hypothesis test between two hypotheses H_λ (X has been uttered by λ) and H_{λ^*} (X has been uttered by another speaker), the system's output score is: $[\log (P_\lambda(X)) - \log (P_{\lambda^*}(X))]$ where $P_\lambda(X)$ and $P_{\lambda^*}(X)$ are the probability density functions associated to the densities of H_λ and H_{λ^*} given X. A single speaker-independent model is used to represent $P_{\lambda^*}(X)$. This model, also called Universal Background Model (UBM) [16], corresponds to a 256 components GMM with diagonal covariance matrices. Each client model is obtained by a mean-only Bayesian adaptation of the UBM using associated training speech data. The decision score for a test sequence corresponds to the mean log-likelihood ratio computed on the whole test utterance.

2.3 The Fusion Method

In this work, we have performed the fusion of two scores, respectively the outputs of the On-line Signature Verification System and the Text-independent Speaker Verification System, by means of a Support Vector Machine (SVM) [13]. In a few words, SVM's goal is to compute a hyperplane in a large dimension feature space which is considered because the input data are not linearly separable in the original space. The distance between the decision surface and the data is maximized, which leads to good generalization performance [13]. Let $X=(x_i)$ be the data with labels $Y=(y_i)$ where $y_i = 0$ *or* 1 represents the class of each person, and Φ is the function which sends the input data X in the feature space F. The distance between the hyperplane $H(w,b) = \{x \in F: \langle w, x \rangle + b = 0\}$ and X, is called the margin Δ. Following the Structural Risk Minimization (SRM) principle, Vapnik [13] has shown that maximizing the margin (or minimizing $||w||$) leads to an efficient generalization criterion. One defines in F the kernel K as: $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$, that avoids handling directly elements in F. The optimal hyperplane is found by solving a quadratic convex problem and, from the optimality conditions of Karush-Kuhn-Tucker [13], one can rewrite w in the following condensed manner:

$$w = \Sigma_{i \in SV} \alpha_i y_i \Phi(x_i)$$

where $SV = \{i: \alpha_i > 0\}$ denotes the set of support vectors.

We have chosen here, as in [12], $K(x,y)=<\Phi(x),\Phi(y)>^d$ with $d = 1$, that is a linear kernel. We fuse the scores of the two experts, each designed for the same person. We thus give as input to the SVM two scores, one per expert.

The optimization of the SVM was carried out on a database considered for training. During this training step, the optimal hyperplane $H(w^*,b^*)$ is computed. This optimal hyperplane generates a given False Rejection Rate (FRR) and a given False Acceptance Rate (FAR). In order to generate a DET (Detection Error Tradeoff) curve [17] during the test phase, the position of the optimal hyperplane is varied. This means that w^* remains constant but that b varies. This corresponds indeed to the variation of a decision threshold.

3 Experimental Setup on BIOMET Bimodal Data

3.1 BIOMET's Signature and Voice Data in Brief

BIOMET is a multimodal biometric database including face, fingerprint, on-line signature, hand shape and voice. We exploit signature and voice data from 77 people with time variability, captured in the two last BIOMET acquisition campaigns, which have a five months spacing between them. More details on the BIOMET database can be found in [9].

Signature data was captured on a digitizer at a rate of 100 samples per second. Each sample contains 5 information: the coordinates $(x(t),y(t))$ of each point sampled on the trajectory, the axial pen pressure $p(t)$ in such a point, and the position of the pen in space (the standard azimuth and altitude angles in the literature). The total number of signatures available per person is 15 genuine and 12 forgeries, made by four different impostors.

Speech data was recorded in quiet environment and using the same kind of microphone. Sampling rate is 16 kHz and sample size is 16 bits. In each session, each speaker uttered twice the 10 digits in ascending and descending order before reading sentences. The amount of available speech for each speaker is about 90 seconds per session.

3.2 Training Protocols per Modality

The Signature Verification expert is trained on 5 signatures randomly chosen among the 15 genuine signatures available.

As for the Text-independent Speaker Verification expert, each client model is adapted using the 10 digits utterance (about 15s of speech). Test data is composed of a segment of speech of approximately 15s, taken from read utterances. For more details, the reader should refer to [12].

3.3 Building the Bimodal Database of Real Subjects

To build the bimodal database, we associate the input data of the two experts (Signature and Voice). We consider for the voice expert two configurations: one without noise, and another with 0db noise.

This bimodal database is then split in 2 subsets: one of 39 persons devoted to training the Support Vector Classifier, named *FLB* (Fusion Learning Base), and the other of 38 persons for testing purposes, named *FTB* (Fusion Test Base). In order to reduce the bias related to the small number of persons in the database, we consider 50 different couples of training and test databases *(FLB,FTB)*, selected randomly, and compute average Errors Rates on the 50 generated *FTBs*. This choice corresponds to the "Trained-Boot" protocol reported in [18], that corresponds to a variant of the Bootstrap sampling principle [19].

For each person in *FLB* and *FTB*, we have at disposal 5 bimodal client accesses and in average 10 bimodal impostor accesses (this number varies across persons from 6 to 12 impostor accesses).

Figure 1 shows the bimodal scores distribution for the 77 persons of the database, in both voice expert's configurations: without noise ("database1") and with 0db noise ("database2"). We notice that discriminating clients from impostors will be more difficult in the case shown in Figure 1 (right), case in which the voice score is very noisy.

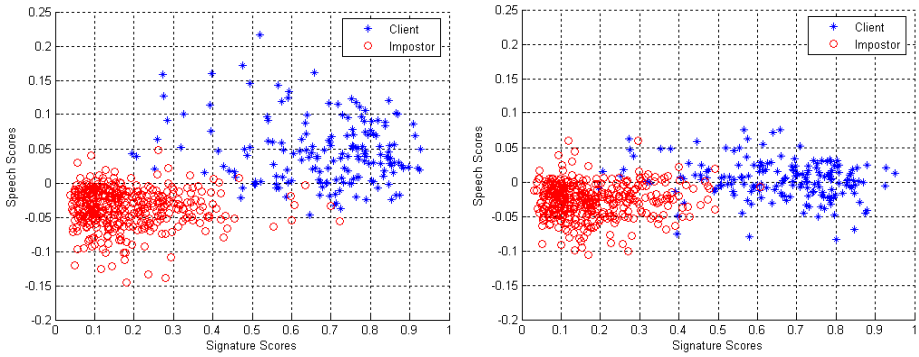


Fig. 1. Bimodal scores' distribution: Signature and Voice without noise (left), Signature and Voice with additive 0db noise (right)

4 Creating Virtual Subjects from BIOMET Bimodal Data

We create a virtual subject by pairing randomly signature data of a given subject to the speech data of another subject. In theory, for two modalities, we can create this way up to *(k-1)!* data sets of virtual subjects, where *k* is the total number of clients in the database. We chose to create 1000 data sets of virtual subjects as in [11].

Every database of virtual subjects is split as described in section 3.3 into a Fusion Learning Base (*FLB*) and a Fusion Test Base (*FTB*). For each possible value of *b* in the equation of the hyperplane *H(w*,b),* where *w** denotes the normal vector to the optimal hyperplane, we compute the mean False Acceptance Rate $\overline{FA}$ and the mean False Rejection Rate $\overline{FR}$ for the 1000 databases of virtual subjects, to obtain a "Virtual Mean DET Curve".

5 Comparative Fusion Experiences on Real and Virtual Subjects

As a first step, we compare the DET curve obtained on the BIOMET database to the 1000 DET curves corresponding to the 1000 databases of virtual subjects. Let's recall that the first curve represents average error rates over 50 different couples *(FLB,FTB)*. Figure 2 (left and right corresponding respectively to the fusion experience without and with additive noise on the voice expert score) shows that the average DET curve on the BIOMET database is inside the band generated by the 1000 DET curves corresponding to virtual subjects sets. This first result permits to conclude that the system behaves on the database of real subjects (when averaging error rates on 50 partitions of the Fusion Learning and Test databases) as on any of the databases of virtual subjects. This also supports the mutual independence assumption between the two modalities that we consider, on-line signature and voice. Moreover, the use of virtual subjects data sets permits to have an estimation of performance variability, providing in fact a "confidence interval" for performance obtained on a real subjects data set of limited size (100 persons). In other words, the database of real subjects is a data set with an inherent bias because of the small number of clients it contains. This bias is greatly increased if a single partition in a Fusion Learning and Testing Databases *(FLB,FTB)* is considered like widely done in the literature. Indeed, the statistics of bimodal data found in the test set (represented by the real subjects present in such set) may be very different of that present in the training set, leading this way to a unreliable and misleading evaluation of the fusion system. It is thus necessary to generate different couples *(FLB,FTB)* that correspond to different distributions of individuals in *FLB* and *FTB* respectively, and to average error rates over those trials. Next experience supports our assumption that considering different partitions or couples *(FLB,FTB)* reduce the bias related to the small number of real subjects in the database. If a large database would be available, this procedure would not be necessary.

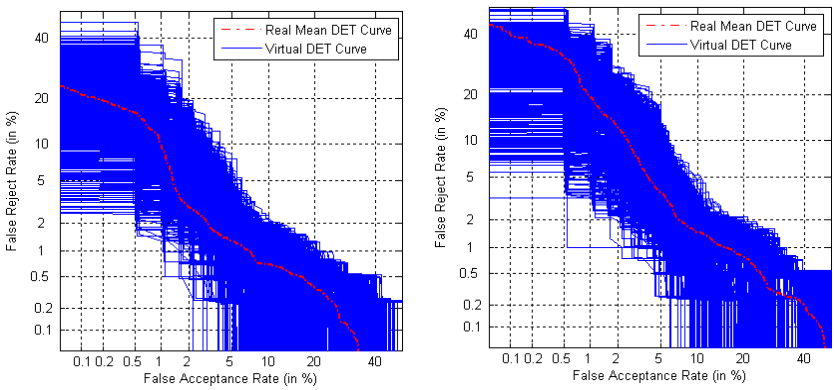


Fig. 2. DET curve for database 1 (left), and for database 2 (right)

Indeed, we now compare, in a second step, the Virtual Mean DET Curve of the 1000 databases of virtual subjects with the mean DET curve on the BIOMET database. In both cases, on database1 (voice without noise) and database2 (voice with 100% noise) shown in Figure 3, we notice that the curves have the same behaviour.

This shows that evaluating this way the fusion system on 1000 virtual data sets is equivalent to evaluating the fusion system on the database of real subjects by averaging results over 50 partitions *(FLB,FTB)* of such database.

The difference between results in Figure 3 left and right can be explained by the fact that the 50 partitions *(FLB, FTB)* are randomly chosen, and therefore are not the same in both cases. This shows that even if 50 partitions or couples *(FLB,FTB)* reduce the bias related to the small number of real subjects in the database, it is still not enough to lead to stable results. Indeed there are C_{77}^{39} possible couples *(FLB, FTB)* and the number of couples that should be considered to "cancel" the bias related to the small number of real subjects in the database is to be studied.

For more insight, we represent in Figure 4 the standard deviation of the errors (False Acceptance Rate and False Rejection Rate) obtained for each value of the decision threshold, on BIOMET data and on 1000 virtual subjects data sets.

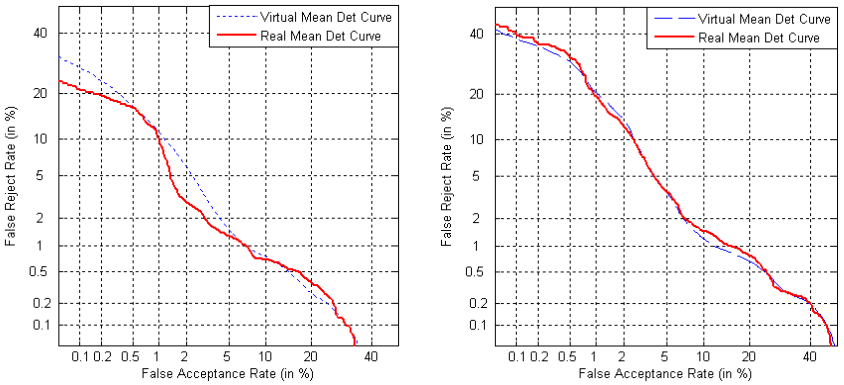


Fig. 3. Virtual Mean DET Curve vs. average Error Rates on the real database for database 1 (left), and for database 2 (right)

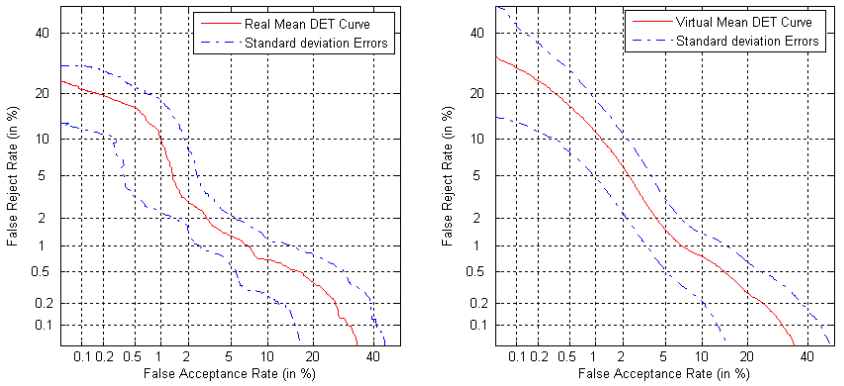


Fig. 4. Mean DET Curve on 50 couples *(FLB, FTB)* on BIOMET data (left) and Virtual Mean DET Curve on 1000 virtual data sets (right), both with associated standard deviation

For this experiment, we chose data without noise. When comparing the standard deviation of errors on real data and on 1000 virtual subjects data sets, we observed that both are comparable when only 50 couples *(FLB, FTB)* are considered. This means that the difference between two samples *(FLB,FTB)* on real data is of the same order of magnitude than the one between two virtual data sets chosen among 1000.

6 Conclusions

We have studied in the present work the possibility of valid multibiometric systems evaluation on limited size databases (about 100 subjects) of real subjects, and also on databases of virtual subjects. Our study focuses on two modalities which are a priori mutually independent, on-line signature and voice, and exploits bimodal data from 77 subjects of the BIOMET database. Several databases of virtual subjects were constructed from BIOMET bimodal data. Our first conclusion is that a limited size database (about 100 subjects) of real subjects behaves exactly as a virtual subjects set of the same size when evaluating the multibiometric system. This of course supports the mutual independence assumption of the two biometric traits that we consider. In other words, this confirms a natural intuition that a database of real subjects has an inherent bias, since each subject represents a specific combination of the modalities considered, and about 100 instances are not enough to cover all the possible variance of such combination, not even for two modalities. This bias is of course stronger if more than two modalities are considered. To cope with this fact, we propose a protocol for multibiometric systems evaluation on limited size databases (about 100 subjects) of real subjects, consisting in creating several partitions (we have shown that 50 partitions is an acceptable compromise) of the data set in a Fusion Learning Base and a Fusion Test Base *(FLB,FTB)* and in averaging error rates over such 50 trials for each value of the threshold. Indeed, evaluating a fusion system on only one partition *(FLB,FTB)* like usually done in the literature, gives biased and thus unreliable results, even if the subjects that are in the database are real!

Moreover, we have shown that it is equivalent to evaluate a fusion system on the database of real subjects by averaging error rates over 50 partitions *(FLB,FTB)*, and on 1000 virtual subjects data sets if a mean False Acceptance Rate and a mean False Rejection Rate are computed on the 1000 data sets for each value of the decision threshold. As a conclusion, we have also proposed a protocol for evaluating a multibiometric system on virtual subjects data sets.

Finally, we can conclude that, in the case of mutual independence of the modalities that are considered, the use of virtual subjects with the protocol above given is a powerful tool to estimate the performance variability, providing a "confidence interval" for performance obtained on a real subjects data set of limited size (100 persons). It is thus recommended for a complete and reliable evaluation of multibiometric systems.

Acknowledgements

This work has been carried out in the framework of GET's (Groupe des Ecoles des Télécommunications) research project Bio-Identity, to which GET-ENST (Ecole Nationale Supérieure des Télécommunications) participates. The authors thank particularly Gérard Chollet and his team for results of speaker verification.

References

1. S. Pigeon & L. Vandendorpe, "The M2VTS Multimodal Face Database", In Proceedings of AVBPA 97, Springer LNCS, Bigün et al. Eds, 1997.
2. http://www.tele.ucl.ac.be/PROJECTS/M2VTS/
3. K Messer J Matas J Kittler, J Luettin and G Maître, "XM2VTSDB: The Extended M2VTS Database", Second International Conference on Audio and Video-based Biometric Person Authentication, 1999.
4. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/results/
5. E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, Jo. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J.-P. Thiran. The BANCA Database and Evaluation Protocol. Audio- and Video-Based Biometric Person Authentication (AVBPA), Guilford, 2003.
6. http://www.ee.surrey.ac.uk/Research/VSSP/banca/
7. J.S.D Mason, F. Deravi, C. Chibelushi & S. Gandon BT DAVID – Final Report, Speech and Image PDETessing Research Group, Dept. of Electrical and Electronic Engineering, University of Wales Swansea, UK.
8. http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html
9. S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux-Les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", Proc. of 4th International Conference on Audio and Video-Based Biometric Person Authentication, pp. 845-853, Guildford, UK, July 2003.
10. A. Ross, A. Jain, "Information Fusion in Biometrics", Pattern Recognition Letters 24, pp. 2115-2125, 2003.
11. M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain, « Multimodal Biometric Authentication Methods: A COTS Approach », Workshop on Multimodal User Authentification (MMUA), pp. 99-106, Santa Barbara, California, USA, Dec. 2003.
12. B. Ly Van, R. Blouet, S. Renouard, S. Garcia-Salicetti, B. Dorizzi, G. Chollet: « Signature with text-dependent and text-independent speech for robust identity verification », Workshop on Multimodal User Authentification (MMUA), pp. 13-18, Santa Barbara, California, USA, Dec. 2003.
13. V. Vapnik, "The Nature of Statistical Learning Theory", *Statistics for Engineering and Information Science,* Second Edition, Springer, 1999.
14. B. Ly Van, S. Garcia-Salicetti, B. Dorizzi, "Fusion of HMM's Likelihood and Viterbi Path for On-line Signature Verification", Biometric Authentication Workshop (BioAW), Lecture Notes in Computer Science (LNCS) 3087, pp. 318-331, Prague, Czech Republic, May 2004.
15. L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall Signal Processing Series*, 1993.
16. D.A Reynolds, T.F. Quatieri, R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing 10, Special Issue on the NIST'99 evaluations, pp. 19-41, 2000.
17. A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET curve in Assessment of Detection Task Performance", Proc. of Eurospeech 1997, 4, pp. 1895-1898, 1997.
18. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communications", in Proc. Of the 12th International Conference on Image Analysis and Processing (ICIAP), 2003.
19. S. Shamsunder, "Signal Processing Applications of the Bootstrap", IEEE Signal Processing Magazine, pp. 38-55, January 1998.

Multi-biometrics 2D and 3D Ear Recognition

Ping Yan and Kevin W. Bowyer

Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame IN 46545, USA
{pyan,kwb}@cse.nd.edu

Abstract. Multi-biometric 2D and 3D ear recognition are explored. The data set used represents over 300 persons, each with images acquired on at least two different dates. Among them, 169 persons have images taken on at least four different dates. Based on the results of three algorithms applied on 2D and 3D ear data, various multi-biometric combinations were considered, and all result in improvement over a single biometric. A new fusion rule using the interval distribution between rank 1 and rank 2 outperforms other simple fusion rules. In general, all the approaches perform better with multiple representations of a person.

1 Introduction

Fingerprints, face and iris have received wide attention both in academic research and in the biometrics industry. Fingerprint and iris are considered as generally more accurate than face, but face is more flexible for use in surveillance scenarios. However, face by itself is not yet as accurate and flexible as desired. Ear images can be acquired in a similar manner to face images, and at least one previous study suggests they are comparable in recognition power [1], so additional work on ear biometrics has promise to lead to increased recognition flexibility and power.

Three algorithms have been explored on 2D and 3D ear images, and based on that, three kinds of multi-biometrics are considered: multi-modal, multi-algorithm and multi-instance. Various multi-biometric combinations all result in improvement over a single biometric. Multi-modal 2D PCA together with 3D ICP gives the highest performance. To combine 2D PCA-based and 3D ICP-based ear recognition, a new fusion rule using the interval distribution between rank 1 and rank 2 outperforms other simple combinations. The rank one recognition rate achieves 91.7% with 302 subjects in the gallery. In general, all the approaches perform much better with multiple images used to represent one subject. In our dataset, 169 subjects had 2D and 3D images of the ear acquired on at least four different dates, which allows us to perform multi-instance experiments. The highest rank one recognition rate reaches 97% with the ICP approach used to match a two-image-per-person probe against a two-image-per-person gallery. In addition, we found that different fusion rules perform differently on different combinations. The min rule works well when combing the multiple presentations of one subject, while the sum rule works well when combining multiple modalities.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 503–512, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Data Acquisition

All the images used in this paper were acquired at the University of Notre Dame in 2003-2004. In each acquisition session, the subject sat approximately 1.5 meters away from the sensor, with the left side of the face facing the camera. Data was acquired with a Minolta Vivid 910 range scanner. One 640x480 3D scan and one 640 x 480 color image are obtained near simultaneously.

The earliest good image for each of 302 persons was enrolled in the gallery. The gallery is the set of images that a "probe" image is matched against for identification. The latest good image of each person was used as the probe for that person. This results in an average of 4.3 weeks time lapse between the gallery and probe. Including the images for multi-instance experiments, there are a total of 942 pairs of 3D and 2D images used in this work (302+302+169+169). A subset of 202 persons of data was used in initial experiments to explore algorithm options.

3 Algorithms

Three different algorithms have been examined. The PCA (Principle Component Analysis) based approach has been widely used in face recognition [2–4]. In our experiments, a standard PCA based algorithm [5] is used on both 2D and 3D ear data. Based on the observation that edge images of the range image are much cleaner than for the 2D edge images, we develop an edge-based Hausdorff distance method for 3D ear recognition using the range image. Also, Besl and McKay's classic ICP algorithm [6] has been applied on 3D ear data. Approaches considered include a PCA ("eigen-ear") approach with 2D intensity images, achieving 63.8% rank-one recognition; a PCA approach with range images, achieving 55.3%; Hausdorff matching of edge images from range images, achieving 67.5%, and ICP matching of the 3D data, achieving 84.1%. Results of these four single-biometric experiments are represented as CMC curves in Figure 1 [7].

4 Multi-biometrics

Recently, multi-biometrics have been investigated by several researchers [8–11]. Multi-biometrics can be divided into three simple classes, according to the method of combination. These are multi-modal, multi-algorithm and multi-instance. In general, multi-modal uses different modalities of biometrics, like face, voice, fingerprint, iris and ear of a same subject. Also we consider that for a given biometric, the data from different sensors are one kind of multi-modal, like 2D intensity data and 3D range data. Multi-algorithm uses different algorithms on the same data. For example, we can use both PCA and ICP on 3D ear data. Multi-instance has more than one representation for a given subject. For example, if we took three 2D ear images of the same person on different dates, then the three images together can be treated as a representation of this person.

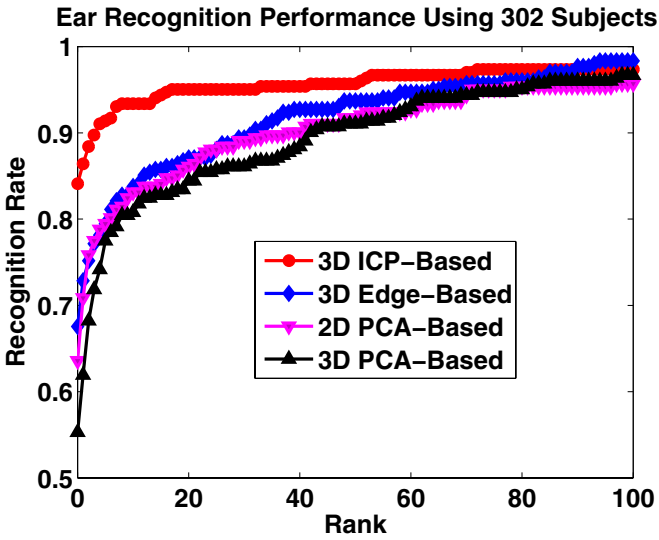


Fig. 1. Performance of Different Approaches

A complex combination can involve more than one kind of multi-biometrics. For example, we can combine 2D PCA and 3D ICP, which includes both multi-modal and multi-algorithm biometrics.

4.1 Fusion Levels and Score Normalization

Each simple biometric has four steps: (1) obtain the data from the sensor, (2) extract the interesting area or features from the raw data, (3) compare the data to a group of enrolled data to obtain the matching score and (4) determine the correct or incorrect matching based on the matching score [12]. Based on these different steps, there are several possible fusion levels. Sensor level fusion combines the raw sensor outputs. Feature extraction level fusion combines multiple extracted features from each biometric. Matching score level fusion combines the matching scores from each biometric. Decision level fusion uses the results from different biometrics and makes the final decision based on all of them.

In our study, the fusion rules work at the matching score level. Since each simple biometric has different meaning, range and distribution of matching scores, score normalization is required in order to combine them. In our experiments, min-max score normalization has been applied on all the results before we do the fusion: $s' = (s - min)/(max - min)$.

4.2 Multi-modal Biometrics

Multi-modal biometrics in this paper refers to the combination of 2D intensity data and the 3D range data. There are three algorithms based on 3D range data,

and one on 2D intensity data. Therefore, the combinations include 2D PCA with 3D ICP, 2D PCA with 3D PCA, and 2D PCA with 3D edge-based approach.

First, two simple fusion rules are tried on all three combinations. As shown in Table 1, the sum rule performs much better than the min rule. This result is similar to the conclusion in [4, 11]. Also an advanced sum rule is tested. The rank one matching in each modality is given an additional weight, which measures the distance between itself and the rank two match. The advanced sum rule yields better results than the simple sum rule.

Table 1. Fusion on Multiple Modalities (302 subjects)

Multi-modals	MIN	Simple SUM	Advanced Sum
2D PCA + 3D ICP	76.4%	81.1%	82.5%
2D PCA + 3D PCA	72.2%	78.8%	79.1%
2D PCA + 3D Edge	73.5%	80.5%	82.5%

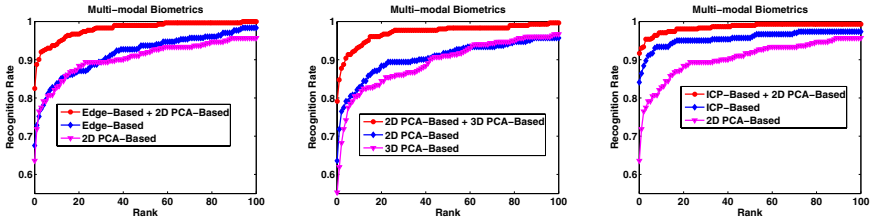
The sum rule adds individual matching scores from different matches. Equal weights are assigned to each modality without any bias. However, in general, some modalities have better performance than others. In order to show the bias of several modalities, different weights are assigned to individual modalities. We test the weight assignment by using 202 subjects on 2D PCA combining with 3D ICP. As shown in Table 2, the highest performance is 93.1%, obtained when the weight of ICP is 0.8, and the weight of PCA is 0.2.

Table 2. Different Weights for Fusing the ICP and PCA results

Weight 2D PCA	Weight 3D ICP	Performance (202 Subjects)	Performance (302 subjects)
1	0	71.4%	63.6%
0	1	85.1%	84.1%
0.9	0.1	73.3%	66.9%
0.8	0.2	76.7%	68.9%
0.7	0.3	78.2%	73.8%
0.6	0.4	81.7%	78.5%
0.5	0.5	84.2%	82.5%
0.4	0.6	86.6%	88.7%
0.3	0.7	89.1%	90.7%
0.2	0.8	93.1%	90.4%
0.1	0.9	91.6%	86.4%

Applying the same weighted sum rule to the other two combinations, the best performance is obtained when there is equal weight for each modality. This is

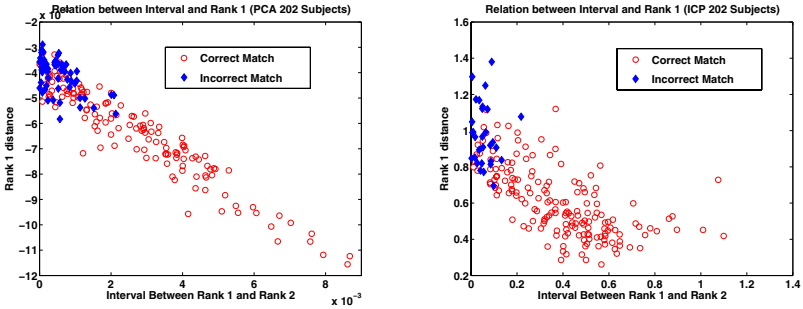
because 2D PCA, 3D PCA and edge-based approaches have similar performance. The rank one recognition is 79.1% when combining 2D PCA and 3D PCA, and it is 82.5% when combining 2D PCA and 3D edge-based algorithm. CMC curve are shown in Figure 2.



(a) (b) (c)

Fig. 2. Multi-modal Biometrics Performance

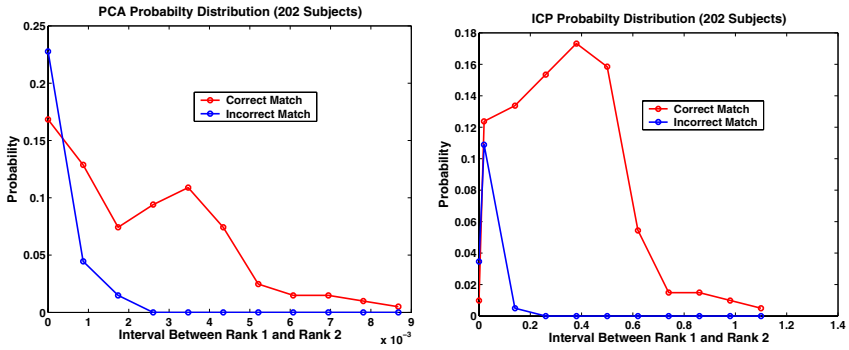
Our third combination rule is based on the analysis of the interval between rank 1 and rank 2 in both PCA and ICP results. Figure 3 shows that the overlap area between the correct matches and incorrect matches is much less in ICP than in PCA, which means that it is easier to use a threshold to separate the correct and incorrect matches in the ICP than in the PCA results.



(a) (b)

Fig. 3. Relationship Between Correct Matches and Incorrect Matches

Figure 4 shows the probability distribution of the different intervals between the correct matches and incorrect matches. In general, the greater the gap between the rank 1 and rank 2, the higher the possibility that it is a correct match. When the interval in ICP is greater than 0.2, they are all correct matches. The corresponding value in PCA is 0.002. For both ICP and PCA, we split the interval range into 10 steps. All the interval values are placed into these 10 steps. The percentage of the correct over incorrect matches in each interval step is shown in Table 3.



(a) (b)

Fig. 4. Interval Distribution Between Correct Matches and Incorrect Matches

Table 3. Fraction of the Correct Match in the Different Interval Level

	1	2	3	4	5	6	7	8	9	10
PCA	0.4250	0.7429	0.8333	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ICP	0.2222	0.5319	0.9999	1.000	1.000	1.000	1.000	1.000	1.000	1.000

When an interval falls into a certain range, we can determine the possibility that it is a correct or incorrect match from Table 3. Using this information we can combine the PCA and ICP in a smarter way. Before the combination, the interval between the rank 1 and rank 2 is computed first for each comparison in the ICP and PCA. Then the corresponding percentage of the correct match and incorrect match is obtained according to Table 3. Using this strategy to combine the PCA and ICP results on 202 subjects, the rank one recognition rate is 93.1%, which is the same as the best results obtained from the simple weight scheme shown in table 2.

Till now, all the results are calculated from 202 subjects. Since the small dataset has a distribution similar to the larger dataset (302 subjects), we predict the distribution of the larger dataset by using the value in Table 3. The rank one recognition rate is 91.7%, which is even better than the results (90.1%) using simple weighted sum scheme. Thus it seems that performance may be increased by using a smart fusion step. However the increase is not statistically significant and this issue deserves further exploration.

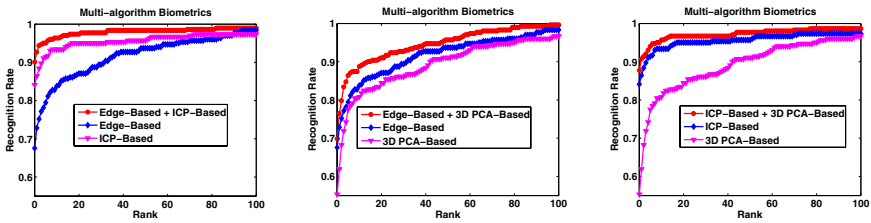
4.3 Multi-algorithm Biometrics

Three different algorithms have been developed to use on the 3D data. These are the ICP-based algorithm, PCA-based algorithm and edge-based algorithm. After score normalization, the weighted sum rule is used for combinations. Rank one recognition rates are demonstrated in Table 4. The best performance is achieved when combining ICP and edge-based algorithm on the 3D data.

Table 4. Multi-algorithm Biometrics Using Weighted Sum Rule

	3DICP	3DPCA	3DEdge	Performance
ICP + PCA	0.90	0.10		87.70%
ICP + Edge	0.80		0.20	90.2%
PCA + Edge		0.40	0.60	69.9%

From Table 1, if we only consider with those not so good performance, like 2D PCA, 3D PCA and 3D edge-based approach, the multi-modal biometrics has better performance than the multi-algorithm.



(a) (b) (c)

Fig. 5. Multi-algorithm Biometrics Performance

4.4 Using Multiple Images to Represent a Person

In general, approaches perform better with a multiple-sample representation of a person, and scale better to larger datasets. We have 169 subjects that have at least 4 good images in both 2D and 3D data. Each pair of 2D and 3D images were taken on a different date. In this section, we will concentrate on the multi-galleries or multi-probes using 2D PCA and 3D ICP algorithms.

For each subject, there are four 2D and 3D images available. We consider three possible multiple-instance representations based on these images. These are (a) 1 in the gallery and the other 3 images in the probe, (b) 2 in the gallery and the other 2 in the probe, and (c) 3 in the gallery and the other 1 in the probe. Two fusion rules, min and sum, are attempted to combine the results, shown in Table 5. It is interesting here that we have multi-instance better than multi-modal. Using one gallery and one probe for these 169 subjects, the rank one rate is 73.4% for 2D PCA, and 81.7% for 3D ICP. Combining the results of 2D PCA and 3D ICP, the best performance obtained is 88.2%.

In the multi-galleries and multi-probes experiments, the best performance is achieved when 2 images are put into the gallery and the other 2 put into the probe. This is true in both 2D PCA and 3D ICP algorithms. This combination gives us 4 matches, whereas the other combinations give 3 matches. Also we noticed that the min rule is much more powerful than the sum rule in the 3D ICP performance, while it has similar performance to the sum rule in the 2D PCA performance. We attribute the performance of the min rule to the possibility of minimizing "outliers"in the 3D matching.

Table 5. Fusion on Multiple Galleries and Probes (169 Subjects)

	2D PCA		**3D ICP**	
1G1P	73.4%		81.7%	
	MIN	**SUM**	**MIN**	*SUM*
1G3P	82.2%	83.4%	95.3%	81.1%
2G2P	84.0%	87.5%	97.0%	81.7%
3G1P	81.7%	80.5%	91.1%	81.7%

Matching of 3D ear images has many sources of "outliers". There can be outlier noise in a given 3D image, such as a "spike" from 3D sensing. Also in matching one 3D image to another, incorrect point correspondences may arise, possibly due to points existing in one scan but not the other. Increasing the number of representations for a certain person in both the gallery and probe gives a better chance to find the correct correspondence between the points. Thus, the performance increases significantly in the ICP experiment.

5 Summary and Discussion

We find that multi-modal, multi-algorithm or multi-instance improve performance over a single biometric. The combination of the 2D PCA and 3D ICP gives the highest performance of any pairs of biometrics considered. Three different multi-biometric combinations were considered. All result in improvement over a single biometric. Among the four single modal ear biometrics, the ICP-based recognition outperforms the other three methods. And it is expected that the best combination includes the ICP as one of the components. Multi-modal with 2D PCA and 3D ICP gives the highest performance. As to the other three not as good methods, multi-modal biometrics turns out to have better performance than the multi-algorithm biometrics.

The fusion experiments on multi-modal, multi-algorithm and multi-instance biometrics yield different results. The sum rule outperforms the min rule on multi-modal and multi-algorithm biometrics, while the min rule performs well on the multi-instance biometrics, especially when using the ICP algorithm. Min rule has the power to reduce the noise from the original data, which is suitable for the application to multi-instance biometrics. The new fusion rule we introduced in combining 2D PCA and 3D ICP is based on analyzing the interval between rank one and rank two. And the performance result is the best of the fusion rules we used.

The multi-modal 3D ICP plus 2D PCA recognition was 87.7% on the 302 person dataset, as listed in Table 4. It is useful to ask how a multi-modal result compares to the multi-instance results for the individual modals. The multi-modal approach represents a person by two images, in both the gallery and as a probe. If we look at the two-image representation in each of the individual imaging modes, we get 87.5% for 2D PCA and 97% for 3D ICP on the subset of 169

of the 302 persons, Table 5. The multi-modal result for this same subset of 169 persons is 88.2%. Thus we find that the multi-modal result barely improves over the two-image 3D ICP result, and that the four-image 3D ICP result for multi-instance is substantially better than the multi-modal result. This is a different relative performance than found by Chang [13] in a study of multi-modal face recognition, where multi-modal 2D + 3D performance was greater than multi-image 2D or multi-image 3D. However, our work differs in several potentially important respects. One is of course that we study ear recognition rather than face recognition. But also, Chang used the same PCA-based approach for both the 2D face and the 3D face recognition, whereas we use an ICP approach for our 3D recognition. This is important because it appears in our results that the ICP-based approach is substantially more powerful than the PCA-based approach for 3D. Another potentially important difference is that in our multi-image results, the two images used to represent a person are taken at different times, at least a week apart. Chang used images from the same acquisition session in his multi-image results. It is quite possible that images taken on different days give a more independent sample, and so better performance.

6 Improved ICP Algorithm

The multi-biometric results presented in previous sections indicate that 3D shape matching with an ICP-based approach has strong potential for ear biometrics. Therefore, after the results in previous sections were completed, considered various refinements to this approach, several of which were incorporated into an improved algorithm. The amount of the ear shape used in the gallery and probe representations was adjusted to reduce interference from the background. An step to remove outlier point matches was added to reduce the effects of incorrect correspondences. Our improved algorithm produces substantially better results. Using the 302-person dataset, with a single 3D ear scan as the gallery enrollment for a person, and a single 3D ear scan as the probe for a person, the new algorithm achieves 98.7% rank-one recognition. This performance from a single modality and algorithm is high enough that a larger and more challenging data set is needed in order to experimentally evaluate its use in possible multi-biometric scenarios. We are currently developing such a dataset for future experiments.

Acknowledgements

This work was supported in part by National Science Foundation EIA 01-30839 and Department of Justice grant 2004-DD-BX-1224.

References

1. Chang, K., Bowyer, K., Barnabas, V.: Comparison and combination of ear and face images in appearance-based biometrics. In: IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 25. (2003) 1160–1165

2. Turk, M., Pentland, A.: Eigenfaces for recognition. In: Journal of Cognitive Neuroscience. Volume 3(1). (1991) 71–86
3. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face recognition algorithms. In: IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 22(10). (2000) 1090–1104
4. Chang, K., Bowyer, K., Flynn, P.: Face recognition using 2D and 3D facial data. In: Workshop on Multimodal User Authentication. (2003) 25–32
5. Beveridge, R., She, K., Draper, B., Givens, G.: Evaluation of face recognition algorithm (release version 4.0). (In: www.cs.colostate.edu/evalfacerec/index.html)
6. Besl, P., McKay, N.: A method for registration of 3-D shapes. In: IEEE Trans. Pattern Analysis and Machine Intelligence. (1992) 239–256
7. Yan, P., Bowyer, K.W.: Ear biometrics using 2d and 3d images. In: Technical Report TR 2004-31, CSE Department, University of Notre Dame. (2004.)
8. Jain, A., Ross, A.: Multibiometric system. In: Communications of the ACM. Volume 47(1). (2004) 34–40
9. Bigűn, E., J.Bigűn, Fischer, S.S.: Expert conciliation for multi modal person authentication systems using Bayesian statistics. In: Proceedings of the International Conference on Audio and Video-Based Biometric Person Authentication. (Mar. 1997) 291–300
10. Brunelli, R., Falavigna, D.: Person identification using multiple cues. In: IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 12(10). (1995) 955–966
11. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. In: IEEE Trans. Pattern Analysis and Machine Intelligence. Volume 20(3). (1998) 226–239
12. Ross, A., Jain, A.: Information fusion in biometrics. In: Pattern Recognition Letters. Volume 24. (2003) 2115–2125
13. Chang, K., Bowyer, K., Flynn, P.: An evaluation of multi-modal 2d+3d face biometrics. Accepted to appear. In: IEEE Trans. Pattern Analysis and Machine Intelligence. (2005)

Biometric Authentication System Using Reduced Joint Feature Vector of Iris and Face

Byungjun Son and Yillbyung Lee

Division of Computer and Information Engineering, Yonsei University
134 Shinchon-dong, Seodaemoon-gu, Seoul 120-749, Korea
{sonjun,yblee}@csai.yonsei.ac.kr

Abstract. In this paper, we present the biometric authentication system based on the fusion of two user-friendly biometric modalities: Iris and Face. Using one biometric feature can lead to good results, but there is no reliable way to verify the classification. In order to reach robust identification and verification we are combining two different biometric features. we specifically apply 2-D discrete wavelet transform to extract the feature sets of low dimensionality from iris and face. And then to obtain Reduced Joint Feature Vector(RJFV) from these feature sets, Direct Linear Discriminant Analysis (DLDA) is used in our multimodal system. This system can operate in two modes: to identify a particular person or to verify a person's claimed identity. Our results for both cases show that the proposed method leads to a reliable person authentication system.

1 Introduction

Biometric authentication, which identifies an individual person using physiological and/or behavioral characteristics, such as iris, face, fingerprints, hand geometry, handwriting, retinal, vein, and speech, is one of the most reliable and capable than knowledge-based(e.g., password) or token-based(e.g., a key) techniques, since biometric features are hardly stolen or forgotten. However, recognition based on any one of these modalities may not be sufficiently robust or else may not be acceptable to a particular user group or in a particular situation or instance.

Current approaches to the use of single biometrics in personal identity authentication are therefore limited, principally because no single biometric is generally considered both sufficiently accurate and user-acceptable for universal application. Multimodal biometrics can provide a more balanced solution to the security and convenience requirements of many applications [1], [2], [3]. However, such an approach can also lead to additional complexity in the design and management of authentication systems. Additionally, complex hierarchies of security levels and interacting user/provider requirements demand that a system is adaptive and flexible in configuration.

There are three main strategies to build multimodal biometric systems. The first method is to apply decision fusion which means combining accept or reject

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 513–522, 2005.
© Springer-Verlag Berlin Heidelberg 2005

decisions of unimodal systems [4]. The other method to construct a multimodal system is using the feature fusion. This means that features extracted using multiple sensors are concatenated. Finally there is the confidence level fusion which means combining matching scores reported by multiple matchers [5].

Our methodology to build multimodal biometric system focuses on the Feature level fusion using face information in combination with iris. Iris and face can be used efficiently in multimodal system because face recognition is friendly and non-invasive whereas iris recognition is one of the most accurate biometrics [1]. When we construct the multimodal system using the feature fusion, one of the most important things we have to consider is a dimensionality of the biometric feature set. It has a disadvantage that the size of the combined feature set is normally large. In recognition systems using the biometric features, one may try to use large feature set to enhance the recognition performance. However, the increase in the number of the biometric features has caused other problems. For example, the recognizer using higher dimension feature set requires more parameters to characterize the classifier and requires more storage. Thus, it will increase the complexity of computation and make its real-time implementation more difficult and costly. Furthermore, a larger amount of data is needed for training. The system we propose is given in Fig. 1 and the dimensionality of the biometric feature set is reduced efficiently in each step.

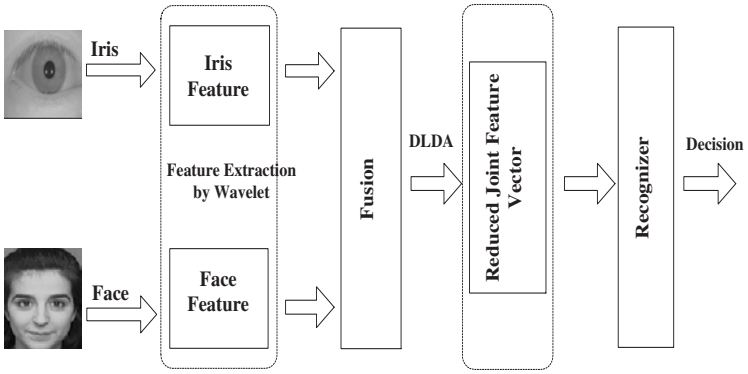


Fig. 1. Bimodal biometric system using Iris and Face

This paper is organized as follows. Section 2 briefly describes image preprocessing to obtain Iris and face images. In section 3, we overview a multilevel two-dimensional Discrete Wavelet Transform (DWT) to extract feature vectors from the iris and face images. we will form a Joint Feature Vector(JFV) from the two biometric feature vectors. Also, we describe the Direct Linear Discriminant Analysis(DLDA) scheme [6] to linearly transform the joint feature vector to new feature space with higher separability and lower dimensionality. The same operations of DWT and DLDA are performed in training as well as testing phases. Experimental results and analysis will be stated in section 4, and finally the conclusions are given in section 5.

2 Image Preprocessing

The images acquired from an image acquisition device always contain not only the appropriate images but also some inappropriate ones. Therefore, we need to check the quality of eye image to determine whether the given images are appropriate for the subsequent processing or not and then to select the proper ones among them in real time. Some images ascertained as inappropriate ones are excluded from the next processing [7].

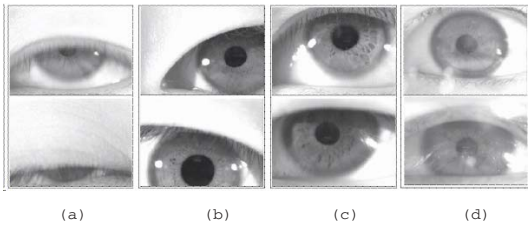


Fig. 2. Examples of images with bad quality: (a)the images with the blink (b)the images whose the pupil part is not located in the middle (c)the images obscured by eyelids or the shadow of the eyelids (d)the images with severe noises

The images excluded from the subsequent processing include as follows; the images with the blink (Fig. 2(a)), the images whose the pupil part is not located in the middle and some parts of the iris area disappear (Fig. 2(b)), the images obscured by eyelids or the shadow of the eyelids (Fig. 2(c)), and the images with severe noises like Fig. 2(d). Fig. 2 shows the examples of images with bad quality.

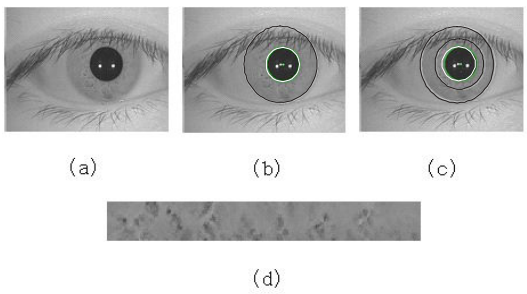


Fig. 3. (a)Original eye image (b)Image of the inner boundary and outer boundary (c)Image of the collarette boundary (d)localized iris image

An iris area can be localized from the eye image passed in the quality check step by separating the part of an image between the inner boundary and outer boundary. Fig. 3 shows the results of finding the inner boundary, the outer boundary and the collarette boundary in the eye image and the image of iris area, where is used in feature extraction, localized by using these boundaries [7], [8].

In this experiment, we just use the localized face images from Olivetti-Oracle Research Lab(ORL)without any preprocessing [9]. The IIS face database accessible at http://smart.iis.sinica. edu.tw/ is also used without any preprocessing [10].

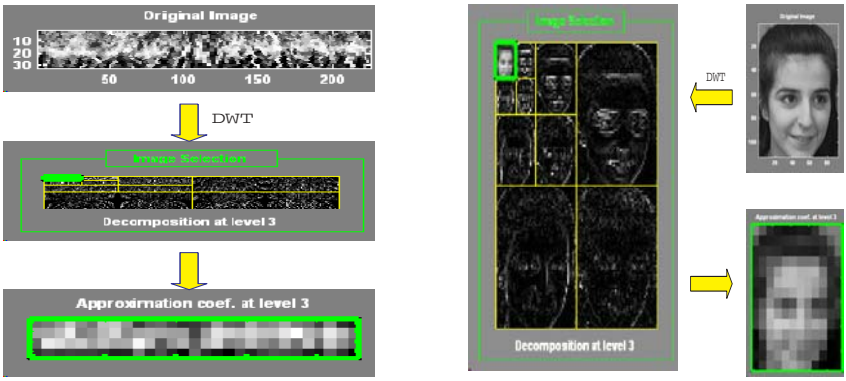
3 Feature Extraction

Most applications emphasize finding a feature set that produces efficient and implementable results. If the dimension of features defining a problem is too high, we must select a robust set of features from an initial set to provide appropriate representation. We have chosen the DWT and DLDA approach to obtain a robust and lower dimensional set of features with high discriminating power. Our previous works have already shown that DWT+DLDA approach can be successfully used on unimodal biometric data [11].

3.1 Wavelet Transform

The hierarchical wavelet functions and its associated scaling functions are to decompose the original signal or image into different subbands. The decomposition process is recursively applied to the subbands to generate the next level of the hierarchy. The traditional pyramid-structured wavelet transform decomposes a signal into a set of frequency channels that have narrower bandwidths in the lower frequency region [12]. The DWT was applied for texture classification and image compression because of its powerful capability for multiresolution decomposition analysis. The wavelet decomposition technique can be used to extract the intrinsic features for the recognition of persons by their biometric data. We employ the multilevel 2D Daubechies wavelet transform to extract the iris and face features. Using the wavelet transform, we decompose the image data into four subimages via the high-pass and low-pass filtering with respect to the column vectors and the row vectors of array pixels. Fig. 4 shows the process of pyramid-structured wavelet decomposition.

In this paper, we use the statistical features and the two or three-level lowest frequency subimage to represent unimodal biometric feature vectors, thus statistical features were computed from each subband image. First, we divide the subimages into local windows in order to get robust feature sets against shift and noisy environment. Next, we extract first-order statistics features, that is, mean and standard deviation from local windows on the corresponding subimages to represent feature vectors. Generally, the mean extracts low spatial-frequency features and the standard deviation can be used to measure local activity in the amplitudes from the local windows [13]. Also, low frequency components represent the basic figure of an image, which is less sensitive to varying images. The feature vectors composed of these features include both local and global information . The level of low frequency subimage chosen to extract the feature vector depends on size of the image. If the size is smaller then our localized iris image and ORL face, the one or two-level lowest frequency subimage might be have higher discriminating power. That is the reason why we choose three-level decomposition on the iris image and ORL face and two-level on the IIS face.



(a) Iris (b) Face

Fig. 4. Example of a three-level wavelet transform of the iris and face images

After the iris and face feature vector are extracted by wavelet transform, the original iris image vector of 7,200 dimensions is transformed to the feature vector of 116 dimensions. Also, the ORL face vector of 10,304 dimensions and IIS face vector of 3,600 dimensions are reduced to 168 dimensions and 241 dimensions respectively. For fusion we use concatenation between the iris and face feature vectors. So we can form a Joint Feature Vector(JFV) y and construct biomodal model using JFV. However, the dimensionality of JFV is too high to reduce the recognition time and save memory.

3.2 Direct Linear Discriminant Analysis

To further reduce the feature dimensionality and enhance the class discrimination, we apply the Direct Linear Discriminant Analysis(DLDA). By using DLDA, we can extract a Reduced Joint Feature Vector(RJFV) z with higher discriminating power and lower dimensionality then the Joint Feature Vector(JFV) y.

Existing LDA methods first use PCA to project the data into lower dimensions, and then use LDA to project the data into an even lower dimension [14]. The PCA step, however, can remove those components that are useful for discrimination. The key idea of DLDA method is to discard the null space of between-class scatter S_b – which contains no useful information – rather than discarding the null space of S_w , which contains the most discriminative information [6]. Each scatter is given as follows:

$$S_b = \sum_{i=1}^{J} n_i(\mu_i - \mu)(\mu_i - \mu)^T, \qquad S_w = \sum_{i=1}^{J} \sum_{x \in C_i} (x - \mu_i)(x - \mu_i)^T$$

where n_i is the number of JFVs in class i, μ_i is the mean of class i, μ is the global mean, and J is the number of classes.

The DLDA method is outlined below. We do not need to worry about the computational difficulty that both scatter matrices are too big to be held in

memory because the dimensionality of input data is properly reduced by wavelet transform.

First, we diagonalize the S_b matrix by finding a matrix V such that

$$V^T S_b V = D$$

where the columns of V are the eigenvectors of S_b and D is a diagonal matrix that contains the eigenvalues of S_b in decreasing order. It is necessary to discard eigenvalues with 0 value and their eigenvectors, as projection directions with a total scatter of 0 do not carry any discriminative power at all [6].

Let Y be the first m columns of V (an $n \times m$ matrix, n being the feature space dimensionality),

$$Y^T S_b Y = D_b \qquad (m \times m)$$

where D_b contains the m non-zero eigenvalues of S_b in decreasing order and the columns of Y contain the corresponding eigenvectors.

The next step is to let $Z = YD^{1/2}$ such that $Z^T S_b Z = I$. Then we diagonalize the matrix $Z^T S_w Z$ such that

$$U^T (Z^T S_w Z) U = D_w \tag{1}$$

where $U^T U = I$. D_w may contain zeros in its diagonal. We can sort the diagonal elements of D_w and discard some eigenvalues in the high end, together with the corresponding eigenvectors.

We compute the LDA matrix as

$$A = U^T Z^T \tag{2}$$

Note that A diagonalizes the numerator and denominator in Fisher's criterion.

Finally, we compute the transformation matrix(3) that takes an $n \times 1$ feature vector and transforms it to an $m \times 1$ feature vector.

$$z_{reduced} = D_b^{-1/2} A y \tag{3}$$

where z is a Reduced Joint Feature Vector and y is a Joint Feature Vector.

4 Experimental Results

4.1 Biometric Database

▷ **Face Database**

- **ORLFace** We used face images from Olivetti-Oracle Research Lab(ORL) [9]. The ORL data set consists of 400 frontal faces: 10 tightly cropped images of 40 subjects with variations in poses, illuminations, facial expressions and accessories. The size of each image is 92 × 112 pixels, with 256 grey levels per pixel.

- **IISFace** The IIS face database is accessible at http://smart.iis.sinica. edu.tw/ [10]. We sampled frontal face images of 100 subjects from the IIS face database, each subject having 10 images with varying expressions. The size of each image is 92×104 pixels, with 256 grey levels per pixel.

▷ **Iris Database.** Eye images were acquired through CCD camera with LED (Light-Emitting Diode) lamp around lens under indoor light. The size of eye images is 320×240 pixels with 256 grey intensity values, and the size of normalized iris images is 225×32 pixels.

- **Iris1.** This data set consists of 1000 iris images acquired from 100 individuals. They are good quality images which pass Image Quality Checking Step(IQCS) of high level.
- **Iris2.** Iris2 consists of 1000 iris images containing some bad quality ones acquired from 100 individuals. They are iris images which pass Image Quality Checking Step(IQCS) of low level.
- **Iris3.** Iris3 is composed of 400 good quality images sampled from Iris1 to combine with ORLFace.
- **Iris4.** Iris4 is composed of 400 iris images containing some bad quality ones sampled from Iris2 to combine with ORLFace.

4.2 Identification and Verification Results on Each Database

In this work, we randomly choose five images per person for training from face and iris, the other five for testing. To reduce variation, each experiment is repeated at least 20 times. We used the following recognition method because it well suit with DWT+DLDA method and is very fast and simple. The training data and test data are transformed by transformation matrix (3), and assign the test data x to the class of its nearest mean, where we say that $\mu_i \in \{\mu_1, \mu_2, \cdots, \mu_J\}$ is a nearest mean to x if

$$D(\mu_i, x) = min_k D(\mu_k, x), \qquad k = 1, 2, \cdots, J \tag{4}$$

where D is Euclidean distance measure.

Table 1. shows the identification rates of unimidal and multimodal systems vs. dimension of biometric feature for IISFace and Iris data.

Table 2. shows the identification rates of unimidal and multimodal systems vs. dimension of biometric feature for IISFace and Iris data.

As can be seen from Table 1. and 2., the multimodal systems using RJFV of face and iris outperform the unimodal systems. In addition, the identification rates for IIS+Iris1 and IIS+Iris2 are 99.12% and 98.4% over 30 and 35 feature dimension, respectively. For ORL+Iris3 and ORL+Iris4, they are 99.7% and 98.7% over 20 and 10 feature dimension, respectively. It shows that the multimodal system using RJFV can achieve much better identification rate over much lower feature dimension than unimodal system. From the results of IIS+Iris2 and ORL+Iris4, we can also see the proposed system can achieve better performance than unimodal system even though one biometric feature set is poor.

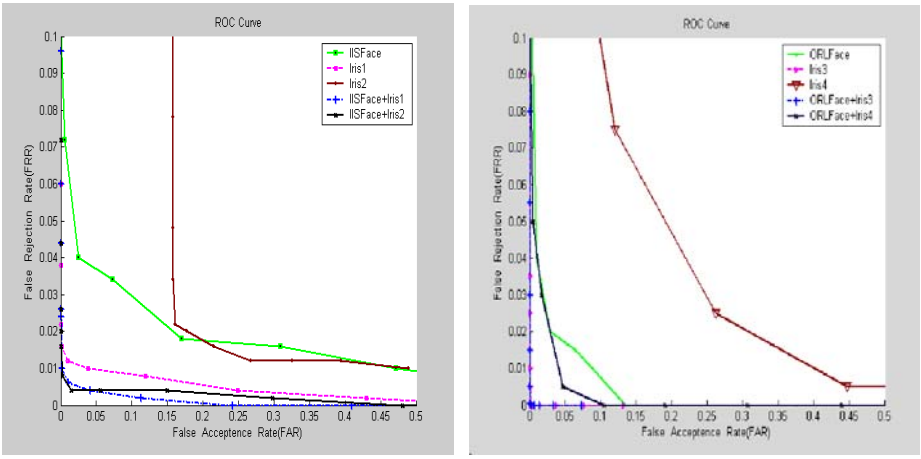
Table 1. Person identification rate for IISFace and Iris data(%)

Feature Dimension	IIS	Iris1	Iris2	IIS+Iris1	IIS+Iris2
30	97.6	98.4	85.92	99.12	97.82
35	97.73	98.47	85.1	99.43	98.4
40	97.58	98.44	85.74	99.58	98.43
45	97.81	98.34	86.16	99.54	98.8
50	97.84	98.69	86.2	99.67	99.04
55	97.66	98.76	**86.38**	99.56	99.23
60	97.87	98.51	85.64	99.75	99.32
65	97.93	98.48	85.86	99.64	99.37
70	97.78	98.76	86.16	99.62	99.21
75	97.67	98.63	85.52	99.79	99.45
80	97.84	98.66	85.52	99.77	**99.56**
85	97.92	**98.91**	85.54	**99.87**	99.43
90	**97.98**	98.85	85.06	99.83	99.48

Table 2. Person identification rate for ORLFace and Iris data(%)

Feature Dimension	ORL	Iris3	Iris4	ORL+Iris3	ORL+Iris4
5	84.9	92.95	68.5	93.18	91.05
10	93.6	97.25	77.7	98.45	98.7
15	94.72	98.65	80.55	99.28	98.95
20	95.63	99.1	82.5	99.7	99.1
25	96.38	99.18	**83.85**	99.77	99.35
30	96.03	99.23	81.8	99.88	99.45
35	**96.43**	99.47	82.15	99.95	99.48
39	96.4	**99.63**	83.45	**99.98**	**99.55**

Two commonly used error measures for a verification system are False Acceptance Rate(FAR) – an imposter is accepted – and False Rejection Rate(FRR) – a client is rejected. If one wants to compare different biometric systems, it is problematic that value "similarities" or, inversely, "distance" are defined very differently, and therefore threshold values often have incomparable meanings. This difficulty is avoided by Receiver Operating Characteristic(ROC) Curve, in which the similarity threshold parameter is eliminated and FRR is seen as a function of FAR. The results of the person verification experiments are shown in Fig. 5. As can be clearly seen, the proposed system using RJFV of face and iris performs considerably better than the unimodal system even though one biometric feature set is poor. It is important to point out that the considerable performance improvement was achieved, although only low dimensional features and poor features were used. Our proposed system can provide users with strong authentication and enhanced convenience for security and reduce verification time.



(a) IISFace and Iris (b) ORLFace and Iris

Fig. 5. Receiver Operating Characteristic(ROC) Curves

5 Conclusion

In this paper, we have shown that the use of data fusion allows to improve significantly the performance of multimodal identification systems. We have also shown that Iris and face can be used efficiently in multimodal system. The grey-level images of iris and face can be simultaneously acquired and used to achieve the performance that may not be possible by single biometric alone. In addition, the DWT+DLDA method has been used to obtain the Reduced Joint Feature Vectors(RJFV) with higher discriminating power and lower dimensionality. These methods of feature extraction well suit with multimodal system as well as unimodal system while allowing the algorithm to be translation and rotation invariant.For future works, it is necessary to conduct experiments on a large number of data so as to verify the efficiency and robustness of our approach. Other techniques for feature extraction and pattern matching can be handled from this point of view so as to propose the efficient methods for a reliable human recognition system.

Acknowledgements

This work was supported in part by the Brain Neuroinformatics Program sponsored by KMST.

References

1. Y. Wang, T. Tan, and A.K. Jain: " Combining Face and Iris Biometrics for Identity Verification", *Proceeding of Int. Conf. on Audio and Video based Biometric Person Authentication*, Guildford, UK, 2003.
2. A.K. Jain, R. Bolle, and S. Pankanti: Biometrics- Personal Identification in Newroked Society, *Kluwer Academic Publishers*, 1999.

3. A. Ross and A.K. Jain: " Information Fusion in Biometrics", *Proceeding of Int. Conf. on Audio and Video based Biometric Person Authentication*, Halmstad, Sweden, 2001.
4. Y. Zuev and S. Ivanon: " The voting as a way to increase the decision reliability", *Faundations of Information/Decision Fusion with Applications to Engineering Problems(Washington D.C., USA)*, 1996.
5. A.K. Jain, S. Prabhakar, and S. Chen: " Combining multiple matchers for a high security fingerprint verification system", *Pattern Recognition Letters*, vol. 20, pp. 1371-1379, 1999.
6. Jie Yang, Hua Yu: "A Direct LDA Algorithm for High-Dimensional Data – with Application to Face Recognition," *Pattern Recognition*, 34(10):2067–2070, 2001.
7. B. Son, G. Kee, Y. Byun, and Y. Lee: " Iris Recognition System Using Wavelet Packet and Support Vector Machines", *Proceeding of International Workshop on Information Security Applications*, Jeju, Korea, 2003.
8. John G. Daugman:"High confidence visual recognition of persons by a test of statistical indenpendence", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, January 1993.
9. AT&T Laboratories Cambridge. The ORL Database of Faces. `http://www.cam-orl.co.uk/facedatabase.html`.
10. Laboratories of Intelligent Systems, Institute of Information Science. The IIS Face Database. `http://smart.iis.sinica.edu.tw/index.html`.
11. B. Son, J. Ahn, J. Park, and Y. Lee: " Identification of Humans using Robust Biometrics Features", *Proceeding of Joint International Workshops on Structural, Syntactic, and Statistical Pattern Recognition*, Lisbon, Portugal , 2004.
12. G. Strang and T. Q. Nguyen, Wavelets and Filter Banks, *Wellesley-Cambridge Press*, 1998.
13. Mallat, S.G., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Trans. Pattern Recognition and Machine Intelligence*, 11(4), pp.674-693, 1989.
14. D. Swets and J. Weng: "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 8, pp.831–836, 1996.

An On-Line Signature Verification System Based on Fusion of Local and Global Information

Julian Fierrez-Aguilar[1,⋆], Loris Nanni[2], Jaime Lopez-Peñalba[1], Javier Ortega-Garcia[1], and Davide Maltoni[2]

[1] Biometrics Research Lab./ATVS, EPS, Universidad Autonoma de Madrid, Campus de Cantoblanco, C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{julian.fierrez,jaime.lopez,javier.ortega}@uam.es
[2] Biometric Systems Lab., DEIS, Università di Bologna, viale Risorgimento 2, 40136 Bologna, Italy
{lnanni,dmaltoni}@deis.unibo.it

Abstract. An on-line signature verification system exploiting both local and global information through decision-level fusion is presented. Global information is extracted with a feature-based representation and recognized by using Parzen Windows Classifiers. Local information is extracted as time functions of various dynamic properties and recognized by using Hidden Markov Models. Experimental results are given on the large MCYT signature database (330 signers, 16500 signatures) for random and skilled forgeries. Feature selection experiments based on feature ranking are carried out. It is shown experimentally that the machine expert based on local information outperforms the system based on global analysis when enough training data is available. Conversely, it is found that global analysis is more appropriate in the case of small training set size. The two proposed systems are also shown to give complementary recognition information which is successfully exploited using decision-level score fusion.

1 Introduction

Automatic extraction of identity cues from personal traits (e.g., signature, fingerprint, voice, and face image) has given raise to a particular branch of pattern recognition, biometrics, where the goal is to infer identity of people from biometric data [1]. The increasing interest on biometrics is related to the number of important applications where an automatic assessment of identity is a crucial point. Within biometrics, automatic signature verification has been an intense research area because of the social and legal acceptance and widespread use of the written signature as a personal authentication method [2]. This work is focused on on-line signature verification, i.e., the time functions of the dynamic signing process (e.g., position trajectories, or pressure versus time) are available for recognition.

⋆ Part of this work has been carried out while J.F.-A. was guest scientist at DEIS, Università di Bologna

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 523–532, 2005.
© Springer-Verlag Berlin Heidelberg 2005

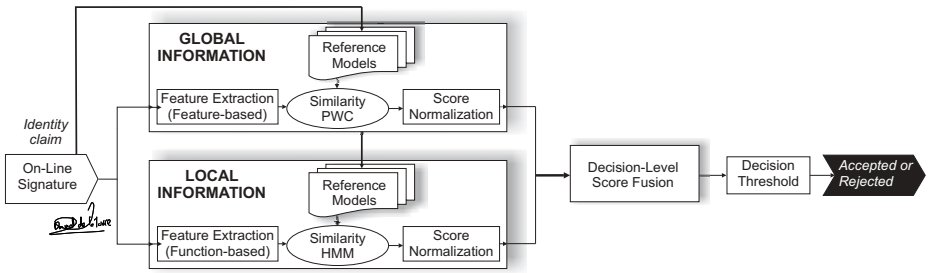


Fig. 1. System model of multilevel signature recognition

Different approaches are considered in the literature in order to extract relevant information from on-line signature data [2]; they can coarsely be divided into: *i*) feature-based approaches, in which a holistic vector representation consisting of global features is derived from the acquired signature trajectories [3–6], and *ii*) function-based approaches, in which time sequences describing local properties of the signature are used for recognition (e.g., position trajectory, velocity, acceleration, force, or pressure) [6–9].

Worth noting, one major research trend in biometric verification is the successful exploitation of the different information levels embodied in the biometric signal at hand. This is usually done by combining the different confidences provided by a number of different machine experts [10, 11] each one working at a different information level. Examples can be found regarding other biometrics like fingerprint verification, where the combined use of local (e.g., minutiae-based) and global (e.g, ridge-based) approaches has been shown to improve verification performance [12]. Regarding on-line signature verification, some works on multi-level approaches are [6, 13].

In the present contribution, we extend our previous work on local function-based recognition [8, 9, 14] by developing a new global feature-based approach. We then combine both systems at the decision level. Results using all the 16500 signatures from the 330 subjects of the publicly available MCYT Bimodal Biometric Database [15] are presented, yielding remarkable performance both with random and skilled forgeries.

The global machine expert is described in Sect. 2 with emphasis on the feature-based representation. The local expert is briefly sketched in Sect. 3. The decision-level combination strategies compared in this work (see Fig. 1 for the system model) are introduced in Sect. 4. Experimental procedure and results are given in Sect. 5. Conclusions are finally drawn in Sect. 6.

2 Machine Expert Based on Global Information

The subsystem exploiting global information is based on the precedent works [3–5]. Our contributions in this regard are as follows: *i*) the set of features described in these works (approximately 70 considering the three works) is extended, leading to a 100-dimensional feature vector representation, *ii*) feature selection ex-

periments on the complete set are carried out, obtaining experimental evidence on the individual relative discriminative power of the proposed and the existing features, and *iii*) a non-parametric statistical recognition strategy based on Parzen windows is used, obtaining remarkable performance in the common case of small training set size.

Feature extraction. The complete set of global features is given in Table 1. Note that an on-line signature acquisition process capturing position trajectories and pressure signals both at pen-down and pen-up intervals is supposed. Otherwise, the feature set should be reduced discarding features based on trajectory signals during pen-ups (e.g., features 32 and 41). Even though the given set has demonstrated to be robust to the common distortions encountered in the handwritten scenario, note that not all the parameters are fully rotation/scale invariant, so either a controlled signature acquisition is assumed (as in MCYT database, where users where asked to sign within grid guidelines) or translation/rotation registration should be performed before computing them. Although pen inclination has shown discriminative power in some works [16], and pen inclination signals are available in MCYT [15], no features based on pen inclination are introduced in the proposed set (as pen inclination turned out to be highly unstable in previous experiments [9]). The features in Table 1 are sorted by individual inter-user discriminative power as described in Sect. 5.2.

Similarity computation. Given the feature vectors of the training set of signatures of a client $\mathcal{C}$, a non-parametric estimation $\lambda_{\mathcal{C}}^{\mathrm{PWC}}$ of their multivariate probability density function is obtained by using Parzen Gaussian Windows [17]. On the other hand, given the feature vector $\mathbf{o}_T$ of an input signature and a claimed identity $\mathcal{C}$ modelled as $\lambda_{\mathcal{C}}^{\mathrm{PWC}}$, the following similarity matching score is used

$$s_{\mathrm{PWC}} = p\left(\mathbf{o}_T | \lambda_{\mathcal{C}}^{\mathrm{PWC}}\right) \tag{1}$$

which is consistent with Bayes estimate in case of equal prior probabilities [17].

3 Machine Expert Based on Local Information

A brief description of the local function-based approach is given in this section, for more details we refer to [8, 9].

Feature extraction. Signature trajectories are first preprocessed by subtracting the center of mass followed by a rotation alignment based on the average path tangent angle. The signature is parameterized then as the following set of 7 discrete-time functions $\{x[n], y[n], p[n], \theta[n], v[n], \rho[n], a[n]\}$, $n = 1, \ldots, N_s$, sampling frequency = 100 Hz., and first order time derivatives of all of them, totaling 14 discrete functions; N_s, p, θ, v, ρ and a stand respectively for signature time duration in time samples, pressure, path tangent angle, path velocity magnitude, log curvature radius and total

Table 1. Set of global features considered in this work sorted by individual discriminative power as described in Sect. 5.2 (T denotes time interval, t denotes time instant, N denotes number of events, θ denotes angle, **bold** denotes novel feature, *italic* denotes adapted from [3–5], roman denotes extracted from [3–5]). Note that all notations are either defined or referenced somewhere in the table (e.g., j is defined and referenced in **4**, Δ is defined in *15*, histograms in $51, 61, 70, 93, \ldots$ are referenced in 34, etc.)

Ranking	Feature Description	Ranking	Feature Description
1	signature total duration T_s	*2*	N(pen-ups)
3	N(sign changes of dx/dt and dy/dt)	**4**	average jerk $\bar{j}$ [3]
5	standard deviation of a_y	**6**	standard deviation of v_y
7	(standard deviation of y)$/\Delta_y$	**8**	N(local maxima in x)
9	standard deviation of a_x	**10**	standard deviation of v_x
11	j_{rms}	**12**	N(local maxima in y)
13	t(2nd pen-down)$/T_s$	14	(average velocity $\bar{v}$)$/v_{x,\max}$
15	$\frac{A_{\min}=(y_{\max}-y_{\min})(x_{\max}-x_{\min})}{(\Delta_x=\sum_{i=1}^{\text{pen-downs}}(x_{\max\,\vert i}-x_{\min\,\vert i}))\Delta_y}$	*16*	$(x_{\text{last pen-up}}-x_{\max})/\Delta_x$
17	$(x_{\text{1st pen-down}}-x_{\min})/\Delta_x$	*18*	$(y_{\text{last pen-up}}-y_{\min})/\Delta_y$
19	$(y_{\text{1st pen-down}}-y_{\min})/\Delta_y$	**20**	$(T_w\bar{v})/(y_{\max}-y_{\min})$
21	$(T_w\bar{v})/(x_{\max}-x_{\min})$	*22*	(pen-down duration T_w)$/T_s$
23	$\bar{v}/v_{y,\max}$	*24*	$(y_{\text{last pen-up}}-y_{\max})/\Delta_y$
25	$\frac{T((dy/dt)/(dx/dt)>0)}{T((dy/dt)/(dx/dt)<0)}$	26	$\bar{v}/v_{\max}$
27	$(y_{\text{1st pen-down}}-y_{\max})/\Delta_y$	*28*	$(x_{\text{last pen-up}}-x_{\min})/\Delta_x$
29	(velocity rms v)$/v_{\max}$	*30*	$\frac{(x_{\max}-x_{\min})\Delta_y}{(y_{\max}-y_{\min})\Delta_x}$
31	(velocity correlation $v_{x,y}$)$/v_{\max}^2$ [4]	*32*	$T(v_y>0\vert\text{pen-up})/T_w$
33	$N(v_x=0)$	34	direction histogram s_1 [4]
35	$(y_{\text{2nd local max}}-y_{\text{1st pen-down}})/\Delta_y$	36	$(x_{\max}-x_{\min})/x_{\text{acquisition range}}$
37	$(x_{\text{1st pen-down}}-x_{\max})/\Delta_x$	*38*	$T(\text{curvature}>\text{Threshold}_{\text{curv}})/T_w$
39	(integrated abs. centr. acc. a_{Ic})$/a_{\max}$ [4]	40	$T(v_x>0)/T_w$
41	$T(v_x<0\vert\text{pen-up})/T_w$	*42*	$T(v_x>0\vert\text{pen-up})/T_w$
43	$(x_{\text{3rd local max}}-x_{\text{1st pen-down}})/\Delta_x$	*44*	$N(v_y=0)$
45	(acceleration rms a)$/a_{\max}$	*46*	(standard deviation of x)$/\Delta_x$
47	$\frac{T((dx/dt)(dy/dt)>0)}{T((dx/dt)(dy/dt)<0)}$	48	(tangential acceleration rms a_t)$/a_{\max}$
49	$(x_{\text{2nd local max}}-x_{\text{1st pen-down}})/\Delta_x$	*50*	$T(v_y<0\vert\text{pen-up})/T_w$
51	direction histogram s_2	**52**	t(3rd pen-down)$/T_s$
53	(max distance between points)$/A_{\min}$	**54**	$(y_{\text{3rd local max}}-y_{\text{1st pen-down}})/\Delta_y$
55	$(\bar{x}-x_{\min})/\bar{x}$	56	direction histogram s_5
57	direction histogram s_3	58	$T(v_x<0)/T_w$
59	$T(v_y>0)/T_w$	60	$T(v_y<0)/T_w$
61	direction histogram s_8	62	(1st $t(v_{x,\min})$)$/T_w$
63	direction histogram s_6	**64**	T(1st pen-up)$/T_w$
65	spatial histogram t_4	66	direction histogram s_4
67	$(y_{\max}-y_{\min})/y_{\text{acquisition range}}$	68	(1st $t(v_{x,\max})$)$/T_w$
69	(centripetal acceleration rms a_c)$/a_{\max}$	**70**	spatial histogram t_1
71	θ(1st to 2nd pen-down)	*72*	θ(1st pen-down to 2nd pen-up)
73	direction histogram s_7	**74**	$t(j_{x,\max})/T_w$
75	spatial histogram t_2	**76**	$j_{x,\max}$
77	θ(1st pen-down to last pen-up)	**78**	θ(1st pen-down to 1st pen-up)
79	(1st $t(x_{\max})$)$/T_w$	**80**	$\bar{j}_x$
81	T(2nd pen-up)$/T_w$	82	(1st $t(v_{\max})$)$/T_w$
83	$j_{y,\max}$	**84**	θ(2nd pen-down to 2nd pen-up)
85	$j_{\max}$	**86**	spatial histogram t_3
87	(1st $t(v_{y,\min})$)$/T_w$	**88**	(2nd $t(x_{\max})$)$/T_w$
89	(3rd $t(x_{\max})$)$/T_w$	90	(1st $t(v_{y,\max})$)$/T_w$
91	$t(j_{\max})/T_w$	**92**	$t(j_{y,\max})/T_w$
93	direction change histogram c_2	**94**	(3rd $t(y_{\max})$)$/T_w$
95	direction change histogram c_4	**96**	$\bar{j}_y$
97	direction change histogram c_3	*98*	θ(initial direction)
99	θ(before last pen-up)	**100**	(2nd $t(y_{\max})$)$/T_w$

acceleration magnitude. A claim-dependent linear transformation is finally applied to each discrete-time function so as to obtain zero mean and unit standard deviation function values.

Similarity computation. Given the parameterized enrollment set of signatures of a client $\mathcal{C}$, a left-to-right Hidden Markov Model $\lambda_{\mathcal{C}}^{\text{HMM}}$ is estimated [17]. No transition skips between states are allowed and multivariate Gaussian Mixture density observations are used. On the other hand, given the function-based representation $\mathbf{O}_T$ of a test signature (with a duration of N_s time samples) and a claimed identity $\mathcal{C}$ modelled as $\lambda_{\mathcal{C}}^{\text{HMM}}$, the following similarity matching score is used

$$s_{\text{HMM}} = \frac{1}{N_s} \log p\left(\mathbf{O}_T | \lambda_{\mathcal{C}}^{\text{HMM}}\right) \tag{2}$$

4 Fusion of Global and Local Information

Two sound theoretical frameworks for combining classifiers with application to biometric verification are described in [10] and [11]. More recent works are reviewed in [1]. These works conclude that the weighted average is a good way of combining the similarity scores provided by the different experts (under some mild assumptions that may not hold in practice).

In this work, fusion strategies based on the max and sum rules [11] are compared. Similarity scores given by the global and local experts are normalized to zero mean and unit standard deviation before fusion.

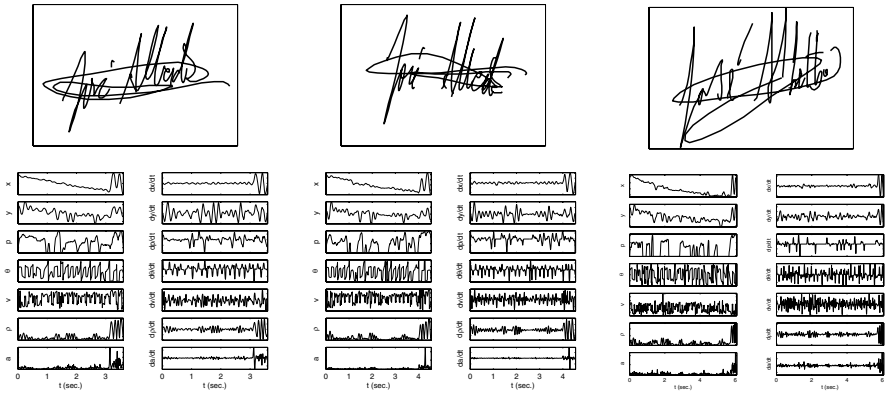
5 Experiments

5.1 Database and Experimental Protocol

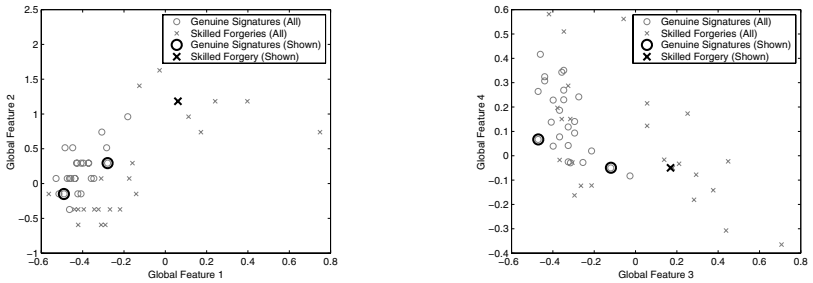
All the signatures of the MCYT database [15] are used for the experiments (330 signers with 25 genuine signatures and 25 skilled forgeries per signer –forgers are provided the signature images of the clients to be forged and, after training with them several times, they are asked to imitate the shape with natural dynamics, i.e., without breaks or slowdowns). Two examples of genuine signatures (left and central columns) and one forgery (right column) are given in Fig. 2.

Signature corpus is divided into training and test sets. In case of considering skilled forgeries, training set comprises either 5 or 20 genuine signatures and test set consist of the remaining samples (i.e., 330×20 or 330×5 client, respectively, and 330×25 impostor similarity test scores). In case of considering random forgeries (i.e., impostors are claiming others' identities using their own signatures), client similarity scores are as above and we use one signature of every other user as impostor data so the number of impostor similarity scores is 330×329.

Overall system performances using *a posteriori* user-independent decision thresholds are reported by means of DET plots [19]. Average EER tables for *a posteriori* user-dependent thresholds are also given following the operational



Two genuine signatures (left and central columns) and one skilled forgery (right column) for a client using name and complex flourish [18]. The function-based description used for local recognition is depicted below each signature.



Best individually performing global features, i.e., 1st versus 2nd (left), and 3rd versus 4th (right), are depicted for all the signatures of the user above. Features from the genuine signatures and forgery above are highlighted.

Fig. 2. Signature examples from MCYT corpus together with their extracted features

procedure proposed in [20] for computing the individual EER of each user. For more details on *a priori* and *a posteriori* decision thresholding techniques and their application to signature verification, we refer the reader to [14].

5.2 Feature Selection

Due to the high number of proposed features (100), and the large number of signatures considered (16500), features have been ranked according to scalar inter-user class separability. Feature selection is then based on selecting an increasing number of ranked features.

For each feature F_k, $k = 1, \ldots, 100$, we compute the scalar Mahalanobis distance [17] d^{M}_{i,F_k} between the mean of the F_k-parameterized training signatures of client i, $i = 1, \ldots, 330$, and the F_k-parameterized set of all training signatures

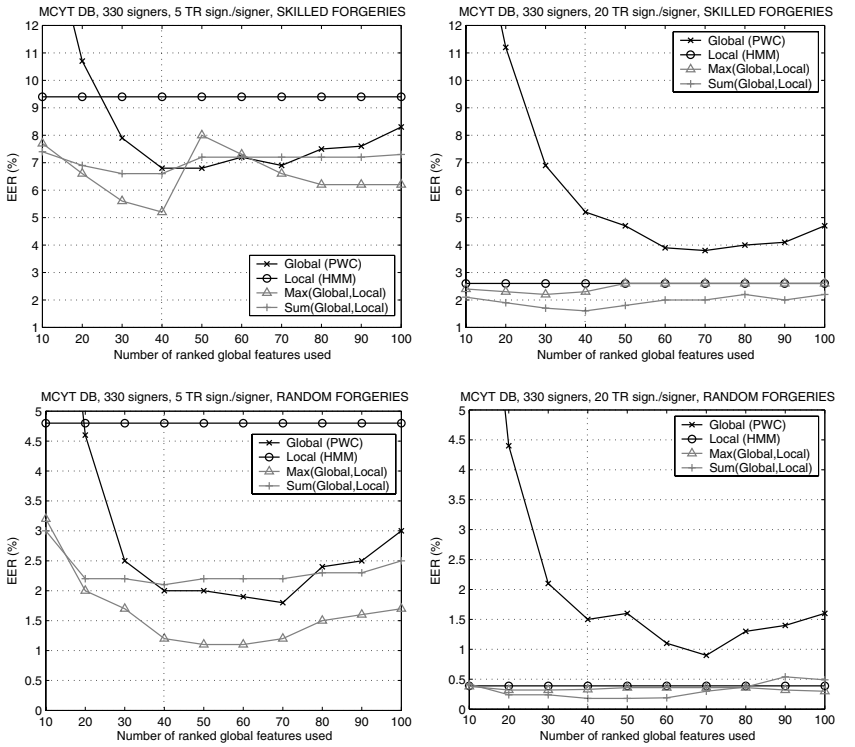


Fig. 3. Verification performance using *a posteriori* user-independent decision thresholding for an increasing number of ranked global features

from all users. Features are then ranked according to the following inter-user class separability measure $S(F_k)$

$$S(F_k) = \sum_{i=1}^{330} \sum_{j=1}^{330} |d^{\mathrm{M}}_{i,F_k} - d^{\mathrm{M}}_{j,F_k}| \tag{3}$$

5.3 Results

In Fig. 3, verification performance results in four common conditions (few/many training signatures and skilled/random forgeries) are given for *i*) the global expert with an increasing number of ranked global features, *ii*) the local expert, and *iii*) their combination through max and sum rules.

Worth noting, the system based on global analysis outperforms the local approach when training with 5 signatures, and the opposite occurs when training with 20 signatures. The two systems are also shown to provide complementary information for the verification task, which is well exploited in the cases of small and large training set sizes using the max and sum rules respectively. Also interestingly, we have found a good working point of the combined system in the

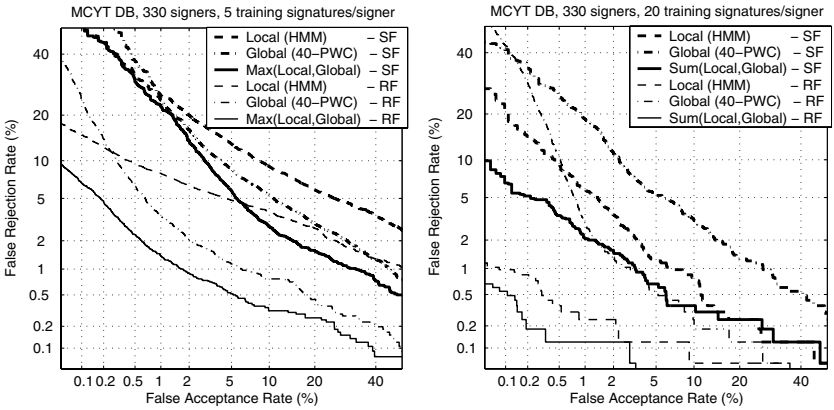


Fig. 4. Verification performance for *a posteriori* user-independent decision thresholding

four conditions depicted in Fig. 3 when using the first 40 ranked features for the global approach. This is highlighted with a vertical dashed line. Detection trade-off curves for this working point are given in Fig. 4.

Verification performances of individual and combined systems for *a posteriori* user-independent and user-dependent decision thresholds are given in Tables 2 and 3. User-dependent decision thresholding leads to error rates significantly lower than user-independent decision thresholding. This effect has also been noticed in previous works [7, 14]. When using user-dependent thresholds and for the four conditions considered, the local approach is found to outperform the global one and the sum rule performs better than the max rule. Also remarkably, the global approach is found to be robust to the score misalignment produced by the strong user-dependencies found in signature recognition, as performance difference between using user-dependent and user-independent thresholds is not as high as the one found for the local approach.

6 Conclusions

An on-line signature recognition system based on fusion of local and global analysis of input signatures has been described. Global analysis is based on a novel feature-based description of signatures and non-parametric statistical modeling based on Parzen windows. Local analysis relies on a function-based approach and parametric statistical modeling through Hidden Markov Models.

Feature selection and performance experiments are conducted on the large MCYT database comprising 16500 different signatures from 330 contributors. Verification performance on random and skilled forgeries has been given for user-specific and global decision thresholds. The machine expert based on global information is shown to outperform the system based on local analysis in the case of small training set size and user-independent thresholds. It has been also found to be quite robust to the severe user-dependencies encountered in signature

Table 2. Verification performance with **5 training signatures** for *a posteriori* user-independent and user-dependent decision thresholding. Average EERs in %

	skilled forgeries		**random forgeries**	
	user-indep.	user-dep.	user-indep.	user-dep.
Local (HMM)	9.39	2.51	4.86	0.59
Global (40 Feat. + PWC)	6.89	5.61	2.02	1.27
Combined (MAX)	**5.29**	2.39	**1.23**	0.41
Combined (SUM)	6.67	**2.12**	2.14	**0.24**

Table 3. Verification performance with **20 training signatures** for *a posteriori* user-independent and user-dependent decision thresholding. Average EERs in %

	skilled forgeries		**random forgeries**	
	user-indep.	user-dep.	user-indep.	user-dep.
Local (HMM)	2.60	**0.51**	0.39	0.0041
Global (40 Feat. + PWC)	5.21	2.38	1.58	0.3180
Combined (MAX)	2.30	0.53	0.33	0.0064
Combined (SUM)	**1.70**	0.55	**0.18**	**0.0005**

recognition. The two proposed systems are also shown to give complementary recognition information which has been exploited with simple rules. Relative improvements in the verification performance as high as 44% (for skilled forgeries) and 75% (for random forgeries) have been obtained as compared to state-of-the-art works[1].

Future work includes applying feature subset selection methods to the proposed set of global features, and exploiting the user-dependencies found in the global and and local approaches through target-dependent score normalization procedures [14] and user-dependent fusion approaches [22].

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Technology under project TIC2003-08382-C05-01. J. F.-A. is also supported by a FPI Fellowship from Comunidad de Madrid.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. on Circuits and Systems for Video Technology **14** (2004) 4–20
2. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification - the state of the art. Pattern Recognition **22** (1989) 107–131

[1] The local system has participated in SVC 2004 (with minor modifications regarding score normalization [14]), where was ranked as the system in first and second place, for random and skilled forgeries, respectively [21].

3. Nelson, W., Kishon, E.: Use of dynamic features for signature verification. In: Proc. of the IEEE Intl. Conf. on Systems, Man, and Cyber. Volume 1. (1991) 201–205
4. Nelson, W., Turin, W., Hastie, T.: Statistical methods for on-line signature verification. Intl. Journal of Pattern Recognition and Artificial Intell. **8** (1994) 749–770
5. Lee, L.L., Berger, T., Aviczer, E.: Reliable on-line human signature verification systems. IEEE Trans. on Pattern Anal. and Machine Intell. **18** (1996) 643–647
6. Kashi, R.S., Hu, J., Nelson, W.L., Turin, W.: On-line handwritten signature verification using hidden markov model features. In: Proc. of ICDAR. (1997) 253–257
7. Jain, A.K., Griess, F., Connell, S.: On-line signature verification. Pattern Recognition **35** (2002) 2963–2972
8. Ortega-Garcia, J., Fierrez-Aguilar, J., Martin-Rello, J., Gonzalez-Rodriguez, J.: Complete signal modeling and score normalization for function-based dynamic signature verification. In: Proc. of AVBPA, Springer LNCS-2688 (2003) 658–667
9. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: A function-based on-line signature verification system exploiting statistical signal modeling. Intl. Journal of Pattern Recognition and Artificial Intelligence (2004) (submitted).
10. Bigun, E.S., Bigun, J., Duc, B., Fischer, S.: Expert conciliation for multi modal person authentication systems by bayesian statistics. In: Proc. of AVBPA, Springer LNCS-1206 (1997) 291–300
11. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. on Pattern Anal. and Machine Intell. **20** (1998) 226–239
12. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer (2003)
13. Zhang, K., Nyssen, E., Sahli, H.: A multi-stage on-line signature verification system. Pattern Analysis and Applications **5** (2002) 288–295
14. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target dependent score normalization techniques and their application to signature verification. IEEE Trans. on Systems, Man and Cybernetics, part C **35** (2005) (to appear).
15. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., et al.: MCYT baseline corpus: A bimodal biometric database. IEE Proc. VISP **150** (2003) 395–401
16. Sakamoto, D., et al.: On-line signature verification incorporating pen position, pen pressure and pen inclination trajectories. In: Proc. of ICASSP. (2001) 993–996
17. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (2003)
18. Fierrez-Aguilar, J., Alonso-Hermira, N., Moreno-Marquez, G., Ortega-Garcia, J.: An off-line signature verification system based on fusion of local and global information. In: Proc. of BIOAW, Springer LNCS-3087 (2004) 295–306
19. Martin, A., et al.: The DET curve in assessment of decision task performance. In: Proc. of EuroSpeech. (1997) 1895–1898
20. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: FVC2000: Fingerprint Verification Competition. IEEE Trans. on PAMI **24** (2002) 402–412
21. Yeung, D.Y., et al.: SVC2004: First International Signature Verification Competition. In: Proc. of ICBA, Springer LNCS-3072 (2004) 16–22
22. Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Bayesian adaptation for user-dependent multimodal biometric authentication. Pattern Recognition **38** (2005) (to appear)

Human Recognition at a Distance in Video by Integrating Face Profile and Gait

Xiaoli Zhou, Bir Bhanu, and Ju Han

Center for Research in Intelligent Systems
University of California, Riverside CA 92521, USA
{xzhou,bhanu,jhan} @vislab.ucr.edu

Abstract. Human recognition from arbitrary views is an important task for many applications, such as visual surveillance, covert security and access control. It has been found to be very difficult in reality, especially when a person is walking at a distance in read-world outdoor conditions. For optimal performance, the system should use as much information as possible from the observations. In this paper, we propose an innovative system, which combines cues of face profile and gait silhouette from the single camera video sequences. For optimal face profile recognition, we first reconstruct a high-resolution face profile image from several adjacent low-resolution video frames. Then we use a curvature-based matching method for recognition. For gait, we use Gait Energy Image (GEI) to characterize human walking properties. Recognition is carried out based on the direct GEI matching. Several schemes are considered for fusion of face profile and gait. A number of dynamic video sequences are tested to evaluate the performance of our system. Experiment results are compared and discussed.

1 Introduction

It has been found to be very difficult to recognize a person from arbitrary views in reality, especially when one is walking at a distance in real-world outdoor conditions. For optimal performance, the system should use as much information as possible from the observations. A fusion system, which combines face and gait cues from low-resolution video sequences, is a practical approach to accomplish the task of human recognition.

The most general solution to analyze face and gait information from arbitrary views is to estimate 3-D models. However, the problem of building reliable 3-D models for articulating objects like the human body remains a hard problem. In recent years, the way to perform integrated face and gait recognition without resorting to 3-D models has made some progress. In [1], Kale et al. present a view invariant gait recognition algorithm and a face recognition algorithm based on sequential importance sampling. The fusion of frontal face and gait cues is in the single camera scenario. In [2], Shakhnarovich et al. compute an image-based visual hull from a set of monocular views of multiple cameras. It is then used to render virtual canonical views for tracking and recognition. A gait recognition scheme is based on silhouette extent analysis. Eigenfaces are used for recognizing frontal face rendered by the visual hull. In a later work [3], Shakhnarovich et al. discuss the issues of cross-modal correlation and score transformations for different modalities and present the probabilistic settings for the cross-modal fusion.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 533–543, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Most gait recognition algorithms rely on the availability of the side view of the subject since human gait or the style of walking is best exposed when one presents a side view to the camera. For face recognition, on the other hand it is preferred to have frontal views analyzed. The requirement of different views is easily satisfied by an individual classifier, while it brings some difficulties to the fusion system. In Kale's and Shakhnarovich's fusion system, both of them use the side view of gait and the frontal view of face. So in Kale's work [1], only the final segment of the NIST database can present a nearly frontal view of face, while in Shakhnarovich's work [2][3], multiple cameras must be used to get both the side view of gait and the frontal view of face simultaneously.

In this paper, an innovative system is proposed, which combines cues of face profile and gait silhouette from the single camera video sequences. We use face profile instead of frontal face in the system since a side view of face is more probable to get than a frontal view of a face when one exposes the best side view of gait to the cameral. It is very natural to integrate information of the side face view and the side gait view. However, it is hard to get enough information of a face profile directly from a low-resolution video frame for recognition tasks. To overcome this limitation, we use super-resolution algorithms for face profile analysis. We first reconstruct a high-resolution face profile image from several adjacent low-resolution video frames. The approach relies on the fact that the temporally adjacent frames in a video sequence, in which one is walking with a side view to the camera, contain slightly different, but unique, information for face profile. Then we extract good features from the high-resolution face profile images. Finally, a curvature-based matching method is applied [4]. For gait, we use Gait Energy Image (GEI) to characterize human walking properties [5]. Recognition is carried out based on the direct GEI matching.

Face profile cues and gait cues are considered being integrated by several schemes. The first two are SUM rule and PRODUCT rule [6]. We assume features of face profile and features of gait we use statistically independent, so matching scores reported by the individual classifier can be combined based on Bayesian Theory. The last one is an indexing-verification scheme, which consolidates the accept/reject decisions of multiple classifiers [7]. The overall technical approach is shown in Fig. 1.

2 Technical Approach

2.1 High-Resolution Image Construction for Face Profile

Multiframe resolution enhancement, or super-resolution, seeks to construct a single high-resolution image from several low-resolution images. These images must be of the same object, taken from slightly different angles, but not so much as to change the overall appearance of the object in the image. The idea of super-resolution was first introduced in 1984 by Tsai and Huang [8] for multiframe image restoration of band-limited signals. In the last two decades, different mathematical approaches have been developed. All of them seek to address the question of how to combine irredundant image information in multiple frames. A good overview of existing algorithms is given by Borman and Stevenson [9] and Park et al. [10]. In this paper, we use an iterative method proposed by Irani and Peleg [11][12].

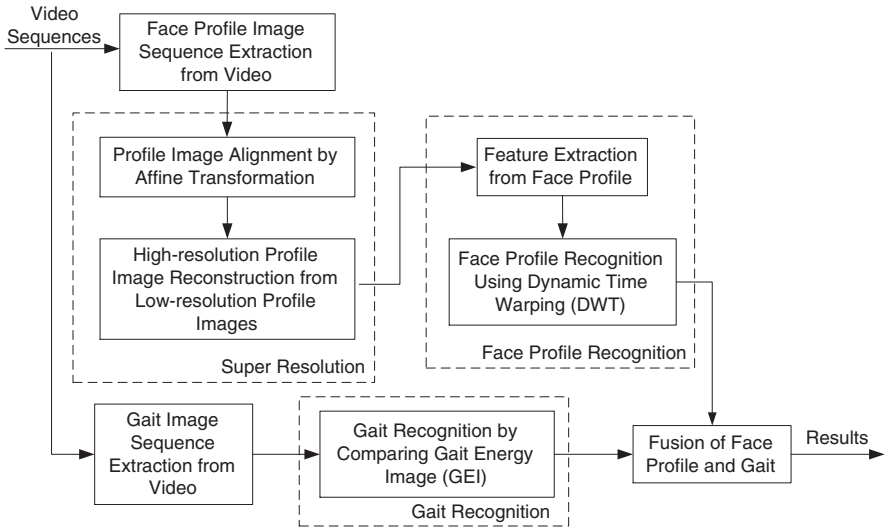


Fig. 1. Technical approach for integrating face profile and gait in video

The Imaging Model. The imaging process, yielding the observed image sequence g_k, is modeled by:

$$g_k(m,n) = \sigma_k(h(T_k(f(x,y))) + \eta_k(x,y)) \tag{1}$$

where

1. g_k is the sensed image of the tracked object in the k_{th} frame.
2. f is a high resolution image of the tracked object in a desired reconstruction view. Finding f is the objective of the super-resolution algorithm.
3. T_k is the 2-D geometric transformation form f to g_k, determined by the computed 2-D motion parameters of the tracked object in the image plane (not including the decrease in sampling rate between f and g_k). T_k is assumed to be invertible.
4. h is a blurring operator, determined by the Point Spread Function of the sensor (PSF). When lacking knowledge of the sensor's properties, it is assumed to be a Gaussian.
5. η_k is an additive noise term.
6. σ_k is a downsampling operator which digitizes and decimates the image into pixels and quantizes the resulting pixels values.

The receptive field (in f) of a detector whose output is the pixel $g_k(m,n)$ is uniquely defined by its center (x,y) and its shape. The shape is determined by the region of the blurring operator h, and by the inverse geometric transformation T_k^{-1}. Similarly, the center (x,y) is obtained by $T_k^{-1}((m,n))$. An attempt is made to construct a higher resolution image $\hat{f}$, which approximates f as accurately as possible, and surpasses the visual quality of the observed images in $\{g_k\}$.

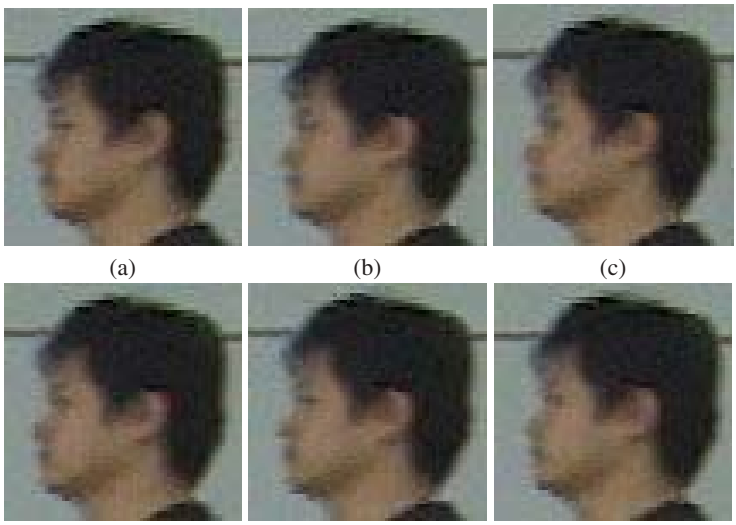


Fig. 2. The six low-resolution face profile images resized by using bilinear interpolation (a-f)

The Super Resolution Algorithm. The algorithm for creating higher resolution images is iterative. Starting with an initial guess $f^{(0)}$ for the high resolution image, the imaging process is simulated to obtain a set of low resolution images $\{g_k^{(0)}\}_{k=1}^K$ corresponding to the observed input images $\{g_k\}_{k=1}^K$. If $f^{(0)}$ were the correct high resolution image, then the simulated images $\{g_k^{(0)}\}_{k=1}^K$ should be identical to the observed image $\{g_k\}_{k=1}^K$. The difference images $\{g_k - g_k^{(0)}\}_{k=1}^K$ are used to improve the initial guess by "backprojecting" each value in the difference images onto its receptive field in $f^{(0)}$, yielding an improved high resolution image $f^{(1)}$. This process is repeated iteratively to minimize the error function:

$$e^{(n)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|g_k - g_k^{(0)}\|_2^2} \tag{2}$$

The imaging process of g_k at the n_{th}iteration is simulated by:

$$g_k^{(n)} = (T_k(f^{(n)}) * h) \downarrow s \tag{3}$$

where $\downarrow s$ denotes a downsampling operator by a factor s, and * is the convolution operator. The iterative update scheme of the high resolution image is expressed by:

$$f^{(n+1)} = f^{(n)} + \frac{1}{K}\sum_{k=1}^{K} T_k^{-1}(((g_k - g_k^{(n)}) \uparrow s) * p) \tag{4}$$

where K is the number of low resolution images. $\uparrow s$ is an upsampling operator by a factor s, and p is a "backprojection" kernel, determined by h and T_k as explained below.

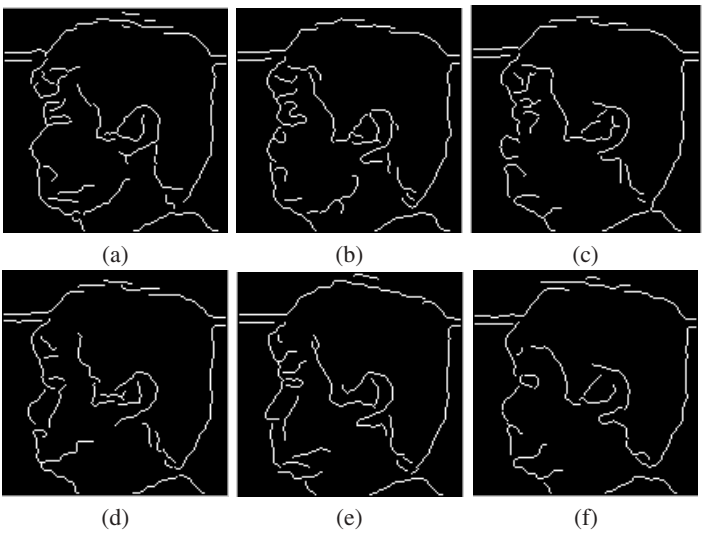


Fig. 3. The edge images of six low-resolution face profiles



Fig. 4. The reconstructed high-resolution face profile and its edge image

The averaging process reduces additive noise. The algorithm is numerically similar to common iterative methods for solving sets of linear equations, and therefore has similar properties, such as rapid convergence.

In our system, we reconstruct a high-resolution face profile image from six adjacent video frames. It relies on the fact that the temporally adjacent frames in a video sequence, in which one is walking with a side view to the camera, contain slightly different, but unique, information for face profile. We assume that six low-resolution face profile images have been localized and extracted from six adjacent video frames. We then align these six low-resolution face profile images using affine transformation. Finally, we apply the super resolution algorithm above to construct a high-resolution face profile image from the six aligned low-resolution face profile images. The resolution of the original low-resolution face profile images is 70x70 and the resolution of the reconstructed high-resolution face profile image is 140x140. Figure 2 shows the six low-resolution face profile images from six adjacent video frames. For comparison, we resize the six low-resolution face profile images by using bilinear interpolation. Figure 3 shows the corresponding edge images of six low-resolution face profiles. Figure 4 shows the reconstructed high-resolution face profile image and its edge image. From

these figures, we can see that the reconstructed high-resolution image is much better than any of the six low-resolution images. It is clearly shown in the edge images that the edges of the high-resolution image are much smoother and more reliable than that of the low-resolution images. This explains why we need to apply super resolution algorithm to our problem. Using the reconstructed high-resolution image, we can extract good features for face profile matching.

2.2 Face Profile Recognition

Face profile is an important aspect for the recognition of faces, which provides a complementary structure of the face that is not seen in the frontal view. For face profile recognition, we use a curvature-based matching approach [4], which does not focus on all fiducial point extraction and the determination of relationship among these fiducial points like most of current algorithms do, but attempt to use as much information as a profile possesses. The scale space filtering is used to smooth the profile and then the curvature of the filtered profile is computed. Using the curvature value, the fiducial points, including the nasion and throat can be reliably extracted using a fast and simple method after pronasale is decided. Then a dynamic time warping method is applied to compare the face profile portion from nasion to throat based on the curvature value. Figure 5 shows the extracted face profile and the absolute values of curvature. Figure 6 gives an example of dynamic time warping of two face profiles from the same person.

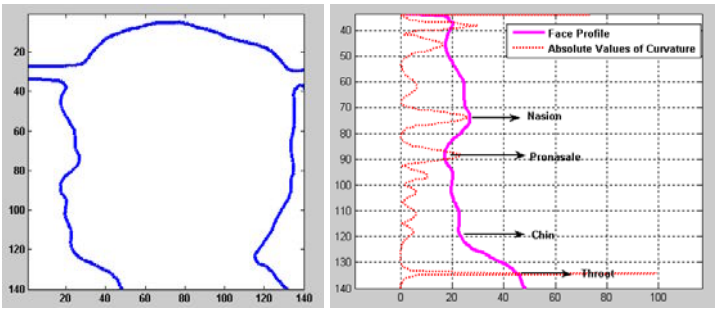


Fig. 5. The extracted face profile and the absolute values of curvature

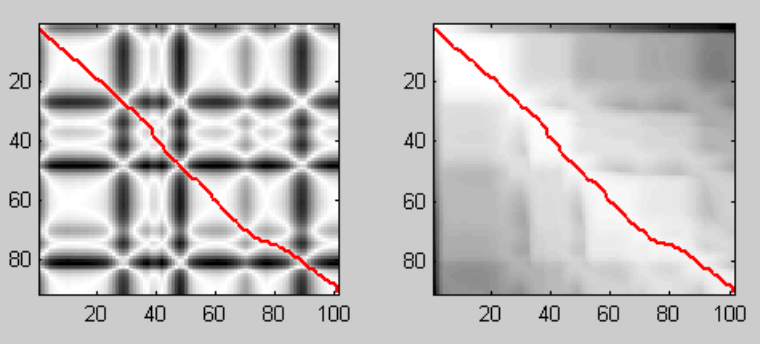


Fig. 6. The similarity matrix (left) and the dynamic programming matrix (right)

From the similarity matrix in Fig. 6, we can see a light stripe (high similarity values) approximately down the leading diagonal. From the dynamic programming matrix in Fig. 6, we can see the lowest-cost path between the opposite corners visibly follows the light stripe, which overlay the path on the similarity matrix. The least cost is the value in the bottom-right corner of the dynamic programming matrix. This is the value we would compare between different templates when we are doing classification.

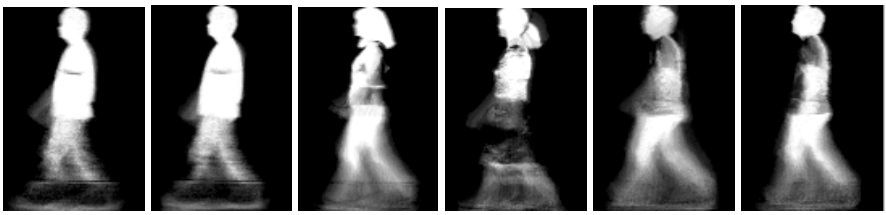


Fig. 7. The Gait Energy Images

2.3 Gait Recognition

Gait Frequency and Phase Estimation. Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency. Therefore, it is possible to divide the whole gait sequence into cycles and study them separately. We assume that silhouette extraction has been performed on original human walking sequences, and begin with the extracted binary silhouette image sequences. The silhouette preprocessing includes size normalization (proportionally resizing each silhouette image so that all silhouettes have the same height) and horizontal alignment (centering the upper half silhouette part with respect to its horizontal centroid). In a preprocessed silhouette sequence, the time series signal of lower half silhouette part size from each frame indicates the gait frequency and phase information. The obtained time series signal consists of few cycles and lots of noise, which lead to sidelobe effect in the Fourier spectrum. To avoid this problem, we estimate the gait frequency and phase by maximum entropy spectrum estimation.

Gait Representation. Given a preprocessed binary gait silhouette sequence $B(x, y, t)$, the grey-level gait energy image (GEI) is defined as follows [5]:

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} B(x, y, t) \tag{5}$$

where N is the number of frames in the complete cycle(s) of a silhouette sequence, t is the frame number of the sequence (moment of time), x and y are values in the 2D image coordinate. Figure 7 is some examples of the Gait Energy Images pairs. As expected, it reflects major shapes of silhouettes and their changes over the gait cycle. We refer to it as gait energy image because: (a) each silhouette image is the normalized gait (human walking) area; (b) a pixel within the silhouette in a image means that human walking occurs at this position and this moment; (c) a pixel with higher intensity value in GEI means that human walking occurs more frequently at this position (i.e., with higher energy).

GEI has several advantages over the gait representation of binary silhouette sequence. GEI is not sensitive to incidental silhouette errors in individual frames. The robustness could be further improved if we discard those pixels with the energy values lower than a threshold. Moreover, with such a 2D template, we do not need to consider the normalized time moment of each frame, and the incurred errors can be therefore avoided.

Direct GEI Matching. One possible approach is recognizing individuals by measuring the similarity between the gallery (training) and probe (testing) templates. Given GEIs of two gait sequences, $G_g(x, y)$ and $G_p(x, y)$, their distance can be measured by calculating their normalized matching error:

$$D(G_g, G_p) = \frac{\sum_{x,y} |G_g(x, y) - G_p(x, y)|}{\sqrt{\sum_{x,y} G_g(x, y) \sum_{x,y} G_p(x, y)}}, \tag{6}$$

where $\sum_{x,y} |G_g(x, y) - G_p(x, y)|$ is the matching error between two GEIs, $\sum_{x,y} G_g(x, y)$ and $\sum_{x,y} G_p(x, y)$ are total energy in two GEIs, respectively.

2.4 Integrating Face Profile and Gait for Recognition at a Distance

Face profile cues and gait cues are considered being integrated by several schemes. Commonly used classifier combination schemes are obtained based on Bayesian Theory, where the representations are assumed conditionally statistically independent. Under different assumptions, there are PRODUCT rule, SUM rule, MAX rule, MIN rule, MEDIAN rule and MAJORITY VOTE rule [6]. We employ SUM rule and PRODUCT rule in our fusion system, with which the similarity scores obtained individually from face profile classifier and gait classifier are combined. Before the similarity scores are combined, it is necessary to map the scores obtained from the different classifiers to the same range of values. Some of the commonly used transformations include linear, logarithmic, exponential and logistic. We use exponential transformation here. The combined similarity score is ranked, which is the result of the fusion system.

The last one is an indexing-verification scheme. In a biometric fusion system, a less accurate, but fast and simple classifier can pass on a smaller set of candidates to a more accurate, but time-consuming and complicated classifier. In our system, the face profile classifier passes on a smaller set of candidates to the gait classifier. Then the result of the gait classifier is the result of the fusion system.

3 Experimental Results

The data is obtained by Sony DCR-VX1000 digital video camera recorder. We collect 28 video sequences of 14 people walking outside and exposing a side view to the camera, at about 30 frames per second. The shutter speed is 1/60 and the resolution of each frame is 720x480. The distance between people and the instrument is about 7 feet. Each of the persons has two sequences. For 4 of the subjects, the data was collected on two separate days and about 1 months apart. Figure 8 shows the six adjacent video frames of one person.

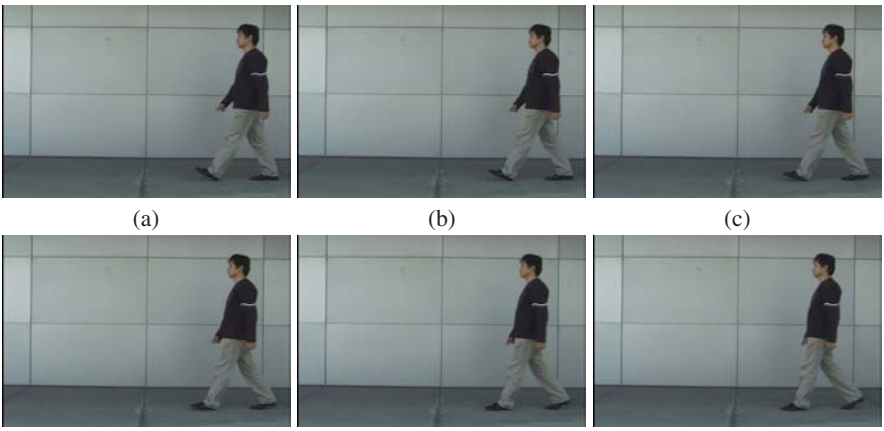


Fig. 8. Six adjacent video frames (a-f)

Table 1. Experimental Results

Combination scheme	Recognition rate		
	Gait	Face profile	Integration
No combination	85.7%	64.3%	
SUM rule			100%
PRODUCT rule			92.9%
Indexing-verification			92.9%

From each sequence, we construct one high-resolution face profile image and one GEI. Since there are two sequences per person, we totally obtain 28 high-resolution face profile images and 28 GEIs for 14 people. Recognition performance is used to evaluate the significance of our method, the quality of extracted features and their impact on identification. The results for our database are shown in Table 1. We can see that 64.3% people are correctly recognized (5 errors out of 14 persons) by face profile and 85.7% people are correctly recognized by gait (2 errors out of 14 persons), respectively. For the fusion schemes, the best performance is achieved by the SUM rule at 100% accuracy. The PRODUCT rule and the indexing-verification scheme obtain the same recognition rate at 92.9%. When we use the indexing-verification scheme, we choose the first three matching results of the face profile classifier as candidates. Then the gait classifier measures the similarity between the corresponding GEI of the testing people and the corresponding GEI of the training people in the candidate list.

There are two people who are not correctly recognized by gait, but when the face profile classifier is integrated, the recognition rate is improved. It is because the clothes of these two people are very different in the training and the testing video sequence, the GEI method can not recognize them correctly. However, the face profiles of these two people don't change so much in the training and the testing sequences. It shows that face profile is a useful cue for the fusion system. On the other hand, since the face profile classifier is comparatively sensitive to the variation of facial expression and

noise, the face profile classifier can not get a good recognition rate by itself. When the gait classifier is combined, the better performance is achieved.

Through the experiments, we can see that our fusion system using face profile and gait is very promising. The fusion system has better performance than either of the individual classifier. It shows that our fusion system is relatively robust in reality under different conditions. Although the experiments are only done on a small database, our system has potential since it integrates cues of face profile and cues of gait reasonably, which are independent biometrics.

4 Conclusions

This paper introduces a practical system combining face profile and gait for human recognition from video. For optimal face profile recognition, we first reconstruct a high-resolution face profile images, using both the spatial and temporal information present in a video sequence. For gait recognition, we use Gait Energy Image (GEI) to characterize human walking properties. Serval schemes are considered for fusion. The experiments show that our system is very promising. Moreover, it is very natural to integrate information of the side face view and the side gait view. However, several important issues that will concern some real-world applications are not addressed in this paper. For example, one problem is how to extract face profile images from video camera automatically and precisely in crowded surveillance applications. Another problem is how to pick up the different frames for the super-resolution algorithm so that the optimal face profile can be reconstructed. These topics will be considered in the future.

References

1. Kale, A., Roychowdhury, A.K., Chellappa, R.: Fusion of gait and face for human identification. Acoustics, Speech, and Signal Processing, 2004. Proceedings. **5** (2004) 901-904
2. Shakhnarovich, G., Lee, L., Darrell, T.: Integrated face and gait recognition from multiple views. Computer Vision and Pattern Recognition, 2001. Proceedings. **1** (2001) 439–446
3. Shakhnarovich, G., Darrell, T.: On probabilistic combination of face and gait cues for identification. Automatic Face and Gesture Recognition, 2002. Proceedings. **5** (2002) 169–174
4. Bhanu, B., Zhou, X.L.: Face recognition from face profile using dynamic time warping. 17th International Conference on Pattern Recogntion. **4** (2004) 499–502
5. Han, Ju, Bhanu, B.:Statistical feature fusion for gait-based human recognition. Computer Vision and Pattern Recognition, 2004. Proceedings. **2** (2004) 842–847
6. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence. **20** (1998) 226–239
7. Zuev, Y., Ivanon, S.: The voting as a way to increase the decision reliability. Foundations of Information/Decision Fusion with Applications to Engineering Problems. **20** (1996) 206–210
8. Tsai, R.Y., Huang, T.S.: Multiframe image resoration and registration. Advances in Computer Vision and Image Processing(T.S. Huang, ed.), JAI Press Inc.. (1984)
9. Borman, S., Stevenson, R.: Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. University of Notre Dame, Tech. Rep.. (1998)

10. Park, S. C., Park, M. K., Kang, M. G.: Super-resolution image reconstruction: A technical overview. IEEE Signal Processing Magazine. **20** (2003) 21–36
11. Irani, M., Peleg, S.: Motion analysis for image enhancement: Resolution, occlusion and transparency. Journal of Visual Communication and Image Representation. **4** (1993) 324–335
12. Irani, M., Peleg, S.: Improving resolution by image registration. CVGIP: Graphical Models and Image Processing. **53** (1991) 231–239

Identity Verification Utilizing Finger Surface Features

Damon L. Woodard and Patrick J. Flynn

Dept. of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556, USA
{dwoodard,flynn}@nd.edu

Abstract. In this paper we present a unique approach to personal identification which utilized finger surface features as a biometric identifier. Finger surface features are extracted from dense range images of an individual's hand. The shape index (a curvature-based surface representation) is used to represent the surfaces of the index, middle, and ring fingers of an individual. This representation is used along with a correlation coefficient based matcher to determine similarity. Our experiments make use of data from 223 subjects possessing a 16 week time lapse between collections. We examine the performance of individual finger surfaces in a verification context as well as the performance when using the three finger surfaces in conjunction. We present the results of our experiments, which indicate that this approach performs well for a first-of-its-kind biometric technique.

1 Introduction

Usually a biometric system performs one or both of the following tasks: *enrollment* (registration) and *verification* (authentication). The enrollment task involves entering users into the system. A biometric sensor is used to obtain one or more samples of the selected biometric identifiers. These identifiers are digitized for feature extraction. The extracted features are stored within the system as a template. The collection of templates of enrolled subjects is sometimes referred to as the *gallery*. Verification is concerned with the confirmation or denial of an individual's identity. The individual asserts his or her identity, and a decision is made based upon the chosen biometric identifier. During verification, the user is scanned by the biometric sensor, and the same features captured at enrollment are extracted from the digitized identifier and converted into a template. Each template collected along with an associated identity claim is called a *probe*. This template is then compared to the stored templates for a match. The quality of a match is quantified by a matching score S. We assume that larger values of S indicate better match quality. The decision of whether a match exists is made by comparing the matching score S to a decision threshold value T. If $S \geq T$, the identity claim is assumed true and so reported. Verification is considered a 1:1 matching problem because the feature template is compared to only one reference template. Based on whether the identity claim originates

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 544–554, 2005.
© Springer-Verlag Berlin Heidelberg 2005

from an enrollee or impostor, the system either correctly or incorrectly accepts or rejects the identity claim.

Other research efforts have investigated the use of finger characteristics as biometric features. Commonly used finger characteristics include finger length and width. A number of research efforts have examined the effectiveness of using these as biometric features. Jain *et al.* [1] developed a system that used measurements of the fingers and hand to establish identity. Sanchez-Reillo *et al.* [2] used a similar approach. Another characteristic used in prior research is that of finger shape. Jain and Duta [3] investigated the use of hand and finger shape extracted from the hand's silhouette as a biometric identifier. Very little work has been performed in 3D hand biometrics. Lay [4] used a grating pattern projected on the back surface of the hand and its distortion by the hand's shape as a biometric identifier. Our work represents the first to use the fine finger surface features such as skin folds and crease patterns extracted from dense range data as a biometric identifier. A curvature based representation is extracted from the registered finger images and used to generate a feature template. This template is matched against stored templates using correlation.

This paper, which is adapted from [5], is organized as follows. We begin by providing details of the data collection and preprocessing procedures. We continue with a discussion of techniques used for matching score calculation as well as the biometric fusion rules implemented. Verification experiments are presented, which demonstrate the performance of our technique. We concluded with a summary of our results and suggestions for future research.

2 Data Collection

Our hand data collections were part of a large multimodal database assembly effort which has been underway since early 2002. For hand data collection the Minolta 900/910 sensor was used [6]. This sensor captures both a 640×480 range image and a registered 640×480 24-bit color intensity image nearly simultaneously. During data collection, the sensor is positioned approximately 1.3m from a flat wall which has been covered with a black piece of cloth. Black cloth was chosen as the background to simplify the hand data segmentation task discussed later. Prior to data collection, the subject was instructed to remove all jewelry. The presence of jewelry during range image capture causes the emitted light from the sensor to scatter when contact is made with the reflective surface of the jewelry. The result is missing or inaccurate range image data near and at that location. The subject was instructed to place his or her right hand flat against the wall with the fingers naturally spread as the image is captured. Between image captures, the subject is instructed to remove his or her hand from the wall and then return it to approximately the same position.

Figure 1(a) shows a sample 640x480 color image of a hand. Figure 1(b) is a pseudo intensity of the same hand rendered using the 640x480 range image as a polygonal mesh. Figure 1(c) depicts the surface detail detected near a knuckle. Our only requirement for hand placement is that the fingers are placed such



(a) Intensity Image Sample (b) Range Image Sample (c) Range Image Detail at Finger Knuckle

Fig. 1. Sample Intensity/Range Images and Range Image Detail

that there is space between two adjacent fingers. By contrast, finger guide pegs were required for precise hand placement is past efforts [1–3]. Our database of collected data was obtained from male and female subjects between the ages of 18 and 70 from various races. The majority of our data was collected from adults between the ages of 18 and 24. Data collection was performed on three separate weeks. During the first week, two images from 132 subjects were collected. Three images were collected a week later from the same 132 subjects. The third week of data collection took place approximately 16 weeks later. During the third week, three images were collected from 177 subjects of which 86 had participated in data collections during the prior two weeks. Hence, our efforts yielded a total of 1,191 hand range images.

3 Preprocessing

A number of preprocessing tasks are required prior to performing our experiments. All of the source code required for preprocessing was written in the MATLAB 6.5 programming language for easy prototyping. The four required tasks were data re-sampling, hand segmentation, finger extraction, and feature template generation.

3.1 Data Re-sampling

Due to slight variations in sensor position from week to week, the pixel spacing between adjacent range image pixels varied. The sampling interval values tended to cluster around 0.425mm. We re-sampled the range images on a 0.4mm grid (in both the x and y directions) to obtain a consistent sampling interval for all of the range images.

3.2 Hand Segmentation

In order to work with only the range image pixels lying on the surface of the hand, the task of hand segmentation was required. To simplify this task, the intensity image of the hand was used. There is a pixel to pixel correspondence between intensity and range images. Therefore, we employed a combination of

edge and skin detection techniques to the intensity image to reliably segment the hand from the image, thereby allowing for segmentation in the range image. The RGB color space skin detection rules specified in Kovač et al. [7] along with an implementation of a Sobel edge detector comprise the segmentation module.

The results of the segmentation module is a binary image of the hand. The binary image is traversed to extract the pixels lying on the hand silhouette's contour. Afterwards the contour of the hand's silhouette is computed. The surfaces of the index, middle, and ring fingers are denoted as α, β, γ respectively and are used in our experiments.

3.3 Finger Extraction

The convex hull of the contour of the hand's silhouette is used to locate the valleys between the fingers represented as circles in Figure 2(a). The valley positions are used as segment boundaries allowing for α, β, and γ to be extracted and processed individually. Once α, β, and γ are extracted, we connected the pixels determined to be segment boundaries. We fill in this closed curve, producing a binary finger mask, as depicted in Figure 2(c). The shaded areas of Figure 2(b) represents the extracted finger pixels. We lessen the effects of noisy range data at the edges of the fingers by removing a two pixel wide portion of the finger mask perimeter. To address finger pose variations, finger masks along with their corresponding range pixels are rotated so that the major axis of the finger mask is coincident with the horizontal axis in an output finger range image. Following rotation, the finger mask pixels along with the corresponding range data is placed in a 80×240 finger image in which the finger mask is centered vertically and positioned five pixels from the right in the output image. The output finger images are used to generate the feature templates, which are used for comparisons.



(a) Finger Valley Locations (b) Extracted Finger Pixels (c) Finger Mask Example

Fig. 2. Finger Valley Locations, Extracted Finger Pixels, and Finger Mask Example

3.4 Feature Template Generation

For each valid pixel of the finger mask in the output image, a surface curvature estimate is computed with the corresponding range data. Valid pixels of the finger mask are those in which the data at the corresponding pixel location in the range image lies on a finger surface and has been marked as valid by the

sensor in the original range image. The linear regression technique, summarized in Flynn and Jain [8], is employed for this task. At each finger surface point of interest $\boldsymbol{p}$, we obtain a set of points $S_{\boldsymbol{p}}$ which neighbor it. We estimate the surface normal and two orthogonal vectors which span the tangent plane centered at $\boldsymbol{p}$. A bi-cubic Monge surface

$$z = f(x,y) = ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + qxy + gy^2 + hx + iy + j \quad (1)$$

is then fit to $S_{\boldsymbol{p}}$ using linear regression. From the result, we calculate analytically the partial derivatives f_x, f_y, f_{xy}, f_{xx}, and f_{yy} to obtain the principal curvature values, κ_{min} and κ_{max} using the formula

$$\kappa_{min,max} = \frac{f_{xx} + f_{yy} + f_{xx}f_y^2 + f_{xx}f_x^2 - 2f_xf_yf_{xy}}{2(1+f_x^2+f_y^2)^{\frac{3}{2}}} \pm \sqrt{\left(\frac{f_{xx} + f_{yy} + f_{xx}f_y^2 + f_{yy}f_x^2 - 2f_xf_yf_{xy}}{2(1+f_x^2+f_y^2)^{\frac{3}{2}}}\right)^2 - \frac{f_{xx}f_{yy} - f_{xy}^2}{(1+f_x^2+f_y^2)^2}}\,. \quad (2)$$

These estimates of curvature contain noise. It has been suggested that the range data be smoothed prior to curvature estimation in order to limit the effects of noise. It was determined that if this approach is used in our application, many of the fine finger surface features are smoothed from the data. This problem is addressed by choosing a relatively large number of points for the Monge surface fit. The window used for determining neighboring points was varied from 3×3 to 15×15 pixels corresponding to a 2D extent of 1.2mm×1.2mm and 6mm×6mm. The optimal window size was chosen as 9×9 or 81 points which corresponds to a 2D extent of 3.6mm×3.6mm. By using a larger number of points during surface fitting, the range data is implicitly smoothed. We found that if more than 81 points are used to fit the surface, many fine surface features are smoothed out from the range data.

The computed principal curvature values were then used to compute the *Shape Index SI* value at each pixel, given by the formula:

$$SI = \frac{1}{2} - \frac{1}{\pi}\arctan\frac{(\kappa_{max} + \kappa_{min})}{(\kappa_{max} - \kappa_{min})} \quad \kappa_{max} \geq \kappa_{min}, \quad (3)$$

SI is a scalar in [0,1] with values that allow shape classification. Shape index was first proposed by Keonderink [9] and has been used successfully by Dorai and Jain [10, 11] for free-form surface representation as well as global object recognition.

In the rare case in which the computed principal curvature values are equal, thereby forcing the shape index formula to be undefined at a particular pixel, the shape index value at that pixel is assigned the value of zero. A zero value indicates that the surface at this pixel location is planar which is consistent with the case of equal principal curvature values.

4 Matching Technique

The match score is the sample correlation coefficient given by the formula:

$$CC(SI_P, SI_G) = \frac{\sum_{\substack{(i,j)\\valid}} \left(SI_P(i,j) - \overline{SI}_P\right) * \left(SI_G(i,j) - \overline{SI}_G\right)}{\sqrt{\left(\sum_{\substack{(i,j)\\valid}} (SI_P(i,j) - \overline{SI}_P)^2\right) * \left(\sum_{\substack{(i,j)\\valid}} (SI_G(i,j) - \overline{SI}_G)^2\right)}}, \tag{4}$$

where $SI_P(i,j)$, $SI_G(i,j)$ are valid shape index values and $\overline{SI}_P$, $\overline{SI}_G$ are the sample mean shape index values in the probe and gallery images, respectively. The resulting match score lies on the interval of [-1,1] where a larger value indicates a better match.

Accurate match score calculation for each technique is dependent on proper alignment of the finger images. During preprocessing, care was taken to automatically align and center the finger mask in each output image. During each matching attempt, the number of overlapping pixels in the gallery and probe image is computed for three vertical offsets (+1 pixel, no offset, -1 pixel) and the offset that maximizes the number of pixels in the overlap is employed during match score computation. On average, the set of overlapping pixels consists of approximately 18,500 pixels. We also experimented with horizontal shifting of images but found that it did not improve matching performance and hence was not required.

5 Score-Level Fusion Rules

In addition to examining each individual finger's performance as a biometric, biometric fusion at the score level is implemented as described in Ross and Jain[12] and Hong *et al.* [13]. The matching score for each finger is treated as an output from a separate biometric system. The multiple scores are then fused into one overall match score using fusion rules proposed by Kittler *et al.*[14]. The first of three score fusion rules implemented is the *average fusion rule* defined as

$$FS_{avg} = \frac{1}{N}\left(\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\beta_i + \sum_{i=1}^{n}\gamma_i\right), \tag{5}$$

where $\alpha_{1,..n}$, $\beta_{1,..n}$, and $\gamma_{1,..n}$ are the match scores calculated for each finger and $N = 3n$, which is the total number of match scores calculated during a single verification attempt. As stated earlier, we perform experiments involving the use of multiple probe and gallery samples. Therefore, n is equal to the product of the number of probe and gallery samples used during a single verification attempt. The second fusion rule implemented is the *median fusion rule* defined as

$$FS_{Median} = Median\left\{\alpha_1, \alpha_2, ..\alpha_n, \beta_1, \beta_2, ..\beta_n, \gamma_1, \gamma_2, ..\gamma_n\right\}, \tag{6}$$

where $\alpha_{1,..n}$, $\beta_{1,..n}$, and $\gamma_{1,..n}$ are the match scores calculated for each finger and FS_{Median} is the median valued match score of all match score calculated during a single verification attempt. The final fusion rule implemented is the *maximum fusion rule* defined as

$$FS_{Max} = Max\{\alpha_1, \alpha_2, ..\alpha_n, \beta_1, \beta_2, ..\beta_n, \gamma_1, \gamma_2, ..\gamma_n\}, \tag{7}$$

where $\alpha_{1,..n}$, $\beta_{1,..n}$, and $\gamma_{1,..n}$ are the match scores calculated for each finger and FS_{Max} is the maximum valued match score calculated during a single verification attempt.

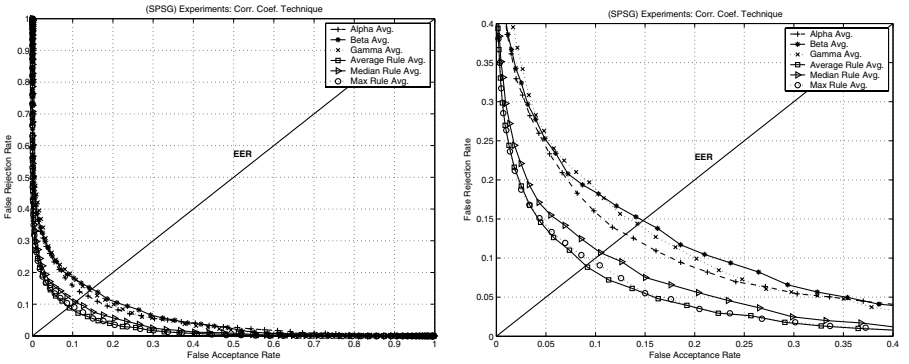
6 Experiments

Verification experiments involved the use of an open universe model, as described by Phillips *et al.*[15]. In this model, a subject in the probe set may or may not be present in the gallery set. The experiments used a probe set of 177 subjects and a gallery of 132 subjects. Of the subjects used, 86 are present in both the probe and gallery sets. Therefore, a total of 223 unique subjects were used for these experiments, many more than in previous related work. Eight images were collected for each subject yielding a total of 1,784 hand images each producing three finger images for a total of 5,352 finger images available for experimentation. A total of 168 verification experiments were performed.

The probe and gallery images of our experiments possessed a relatively long time lapse between their collection in order to evaluate the stability of the extracted features over time. During a verification attempt, only comparisons between images of the same finger type are performed. The following sections present the results of using each of the matching techniques for verification as Receiver Operating Characteristic (ROC). The match threshold was varied from 0 to 1 in increments of 0.01 for curve generation. At each threshold value the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are calculated. For each experiment configuration we use the Equal Error Rate (EER) to quantify verification performance.

6.1 Group A Experiments

For this configuration, a single probe image is compared to a single gallery image during each verification attempt. For each experiment, templates from 132 probe subjects are compared to templates of 177 gallery subjects which result in a total of 23,364 performed verification attempts. Of these attempts, 86 are genuine and 23,278 are impostor. The false acceptance rates (FAR) and false rejection rates (FRR) are computed for each threshold value and plotted as a ROC curve in Figure 3. Each curve represents the average of fifteen experiments. As expected, the fusion rule experiments exhibit better performance than the single finger experiments. Figure 3(b) provides a close view of the graph region where the equal error rates lie. The lowest equal error rate obtained is approximately 9% and achieved by the average fusion rule. The set of fusion rules experiments exhibited better performance than any of the single finger surface experiments.

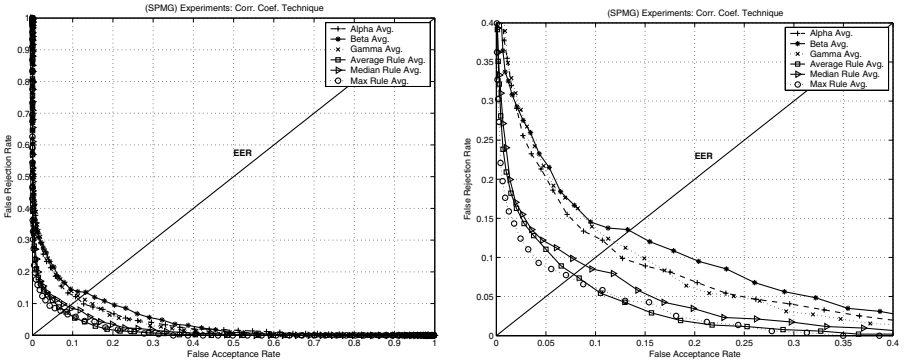


(a) ROC Curves (b) Equal Error Rates

Fig. 3. Verification Performance: Group A Experiments

6.2 Group B Experiments

During this set of experiments, the number of gallery samples used during match score calculation was increased in an effort to reduce the equal error rate. During the single finger surface experiments, the overall match score is computed as the average of the matching scores computed during that verification attempt. In contrast, during the fusion rule experiments the overall match score is computed according to fusion rule. This experiment configuration involved comparing a single probe image to either two or three gallery images during each verification attempt. The curves of the ROC plot in Figure 4 represent the average of six experiments. There was not a significant difference in performance when using either two or three gallery samples which may indicate that there was not a significant amount of variation present between the gallery sample images. The Group B experiments obtained a lowest equal error rate of about 7% as compared to 9% obtained during Group A experiments, as shown in Figure 4(b). The fusion rules continued to outperform the single finger type experiments.



(a) ROC Curves (b) Equal Error Rates

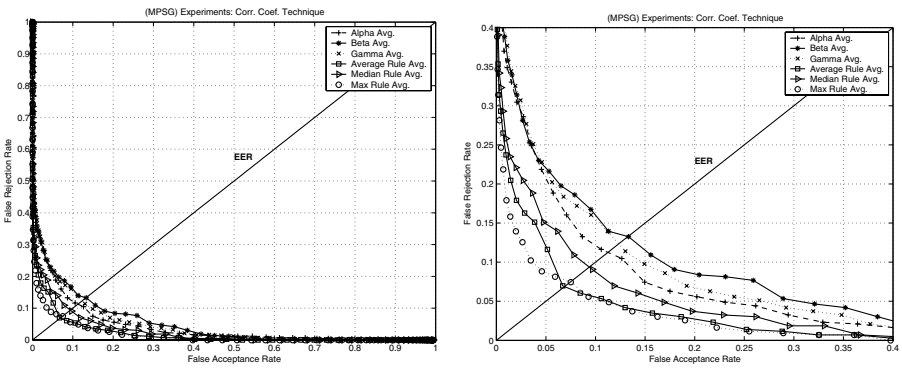
Fig. 4. Verification Performance: Group B Experiments

6.3 Group C Experiments

Three probe image samples were compared to a single gallery image sample during this set of experiments. The curves of the ROC plot in Figure 4 represent the average of five experiments. A slight decrease in equal error rate occurred using this configuration as compared to the previous experiment group. The lowest equal error rate was attained using the average fusion rule and was approximately 6%. This performance improvement over the Group A experiments would suggest the existence of variation between probe image samples collected from the same subject.

6.4 Group D Experiments

The final experiment configuration involved utilizing all images collected during a single session in the probe and gallery sets. Three probe image samples are compared to all five of the gallery image samples collected 16 weeks prior. The result of a single experiment is presented in Figure 6. Using this experiment configuration, we achieved an equal error rate of 5.5% using both the maximum and average fusion rule, as shown in Figure 6(b). This slight performance improvement in performance over the experiments performed in the pervious section suggests that there is exists a significant amount of variation between images in the gallery set.

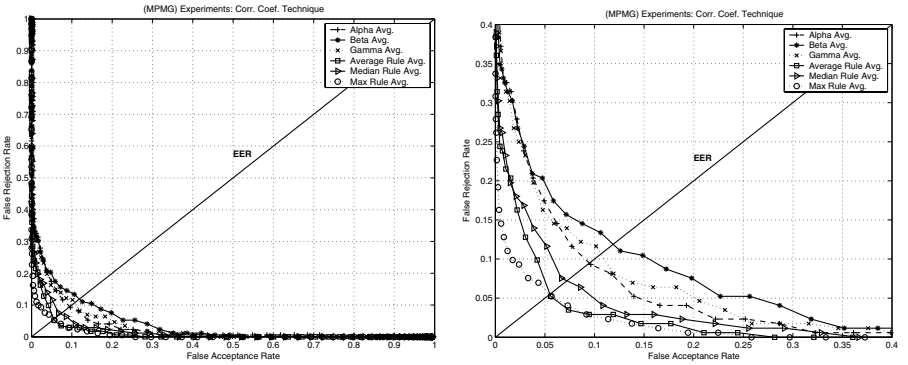


(a) ROC Curves (b) Equal Error Rates

Fig. 5. Verification Performance: Group C Experiments

7 Conclusions and Future Work

The results of an initial research effort in the use of 3D finger surface as a biometric identifier were presented. Shape index was shown to be a suitable surface representation for this application. The relatively low equal error rates obtained during our experiments utilizing time lapsed data suggest that finger surface features displayed little change over time. To address variations in collected sample



(a) ROC Curves (b) Equal Error Rates

Fig. 6. Verification Performance: Group D Experiments

images, multiple probe and gallery image samples were used during matching score calculations. This approach yielded an increase in verification performance. Other research has shown that biometric systems which use multiple modalities can achieve better performance than single modality based systems. These results indicate that further research is required. Future work would involve the fusion of features detectable in 2D images with those extracted from the range data. These features would include finger shape, skin color and texture, as well as skin crease patterns. Because our data collection efforts involve obtaining both range and intensity images using a single sensor, it would be logical that perhaps the information obtained from these two modalities should be combined for identity verification purposes. In addition, the combination of finger surface and other biometric identifiers such as 2D/3D face data and palm print could also result in performance increases.

Acknowledgements

The research presented in this paper was supported by the Defense Advanced Research Projects Agency and the Office of Naval Research under grant N00014-02-1-0410, and by the National Science Foundation under grant EIA-0130839.

References

1. Jain, A. K., Ross, A., and Pankanti, S.: A Prototype Hand Geometry-Based Verification System. In: Proc. of 2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication(AVBPA), Washington D.C. (1999) 166–171
2. Sanchez-Reillo, R., Sanchez-Avila, C., and Gonzalez-Marcos, A.: Biometric Identification through Hand Geometry Measurements. IEEE Trans. on Pattern Analysis and Machine Intelligence **22** (2000) 1168–1171
3. Jain, A. K. and Duta, N.: Deformable Matching Of Hand Shapes For Verification. In: Proceedings of International Conference on Image Processing. (1999) 857–861

4. Lay, Y. L.: Hand Shape Recognition. Optics and Laser Technology **32** (2000) 1–5
5. Woodard, D. L.: Exploiting Finger Surface as a Biometric Identifier. PhD thesis, The University of Notre Dame, Notre Dame, IN 46556, USA (2004)
6. Konica Minolta Website: `http://konicaminolta.com`. Accessed (2004)
7. Kovač, J, Peer, P. and Solina, F.: Human Skin Colour Clustering for Face Detection. In Zajc, Baldomir, ed.: EUROCON 2003 - International Conference on Computer as a Tool, Ljubljana, Slovenia (2003)
8. Flynn, P. J. and Jain, A. K.: On reliable curvature estimation. Proc. IEEE Conf. Computer Vision and Pattern Recognition **89** (1989) 110–116
9. Koenderink, J. J. and van Doorn, A. J.: Surface shape and curvature scales. Image and Vision Computing **10** (1992) 557–564
10. Dorai, C. and Jain, A. K.: COSMOS-A Representation Scheme for Free-Form Surfaces. In: International Conference on Computer Vision. (1995) 1024–1029
11. Dorai, C. and Jain, A. K.: COSMOS-A Representation Scheme for 3D Free-Form Objects. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 1115–1130
12. Ross, A. and Jain, A. K.: Information Fusion in Biometrics. Pattern Recognition Letters **24** (2003) 2115–2125
13. Hong, L., Jain, A. K., and Pankanti, S.: Can Multibiometrics Improve Performance. Technical Report MSU-CSE-99-39, Department of Computer Science, Michigan State University, East Lansing, Michigan (1999)
14. Kittler, J., Hatel, M., Duin, R.P.W., and Matasm, J.: On Combining Classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 226–239
15. Phillips, P. J. and Moon, H., Rizvi, S. A. and Rauss, P. J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1090–1104

Palmprint Authentication Based on Orientation Code Matching

Xiangqian Wu[1], Kuanquan Wang[1], and David Zhang[2]

[1] School of Computer Science and Technology,
Harbin Institute of Technology (HIT), Harbin 150001, China
{xqwu,wangkq}@hit.edu.cn
http://biometrics.hit.edu.cn
[2] Biometric Research Centre, Department of Computing,
Hong Kong Polytechnic University, Kowloon, Hong Kong
csdzhang@comp.polyu.edu.hk

Abstract. This paper presents a novel approach of palmprint authentication by matching the orientation code. In this approach, each point on a palmprint is assigned a orientation. And all point orientations of a palmprint constitute a palmprint orientation code (POC). Four directional templates with different directions are devised to extract the POC. The similarity of two POC is measured using their Hamming distance. This approach is tested on the public PolyU Palmprint Database and the experimental results demonstrate its effectiveness.

1 Introduction

Computer-aided personal recognition is becoming increasingly important in our information society. Biometrics is one of the most important and reliable methods in this field [1, 2]. Within biometrics, the most widely used biometric feature is the fingerprint [3, 4] and the most reliable feature is the iris [1, 5]. However, it is very difficult to extract small unique features (known as minutiae) from unclear fingerprints [3, 4] and the iris input devices are expensive. Other biometric features, such as the face and the voice, are as yet not sufficiently accurate. The palmprint is a relatively new biometric feature. Compared with other currently available features, palmprint has several advantages [6]. Palmprints contain more information than fingerprints, so they are more distinctive. Palmprint capture devices are much cheaper than iris devices. Palmprints contain additional distinctive features such as principal lines and wrinkles, which can be extracted from low-resolution images. By combining all features of palms, such as palm geometry, ridge and valley features, and principal lines and wrinkles, it is possible to build a highly accurate biometrics system.

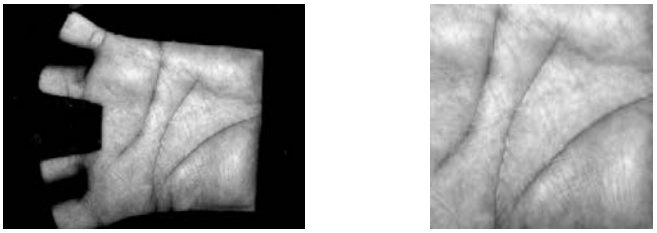
Many algorithms have been developed for palmprint recognition in the last several years. Han [7] used Sobel and morphological operations to extract line-like features from palmprints. Similarly, for verification, Kumar [8] used other directional masks to extract line-like features. Wu [9] used Fisher's linear discriminant to extract the algebraic feature (called Fisherpalms). The performance of these methods are heavily affected by the illuminance. Zhang [10, 11] used

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 555–562, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2-D Gabor filters to extract the texture features (called PalmCode) from low-resolution palmprint images and employed these features to implement a highly accurate online palmprint recognition system. In this paper, we encoded a palmprint using the orientation information of its points. This code is called palmprint orientation code (POC). In the matching stage, the POCs are matched according to their Hamming Distance.

When palmprints are captured, the position, direction and amount of stretching of a palm may vary so that even palmprints from the same palm may have a little rotation and translation. Furthermore, palms differ in size. Hence palmprint images should be orientated and normalized before feature extraction and matching. The palmprints used in this paper are from the Polyu Palmprint Database [12]. The samples in this database are captured by a CCD based palmprint capture device [10]. In this device, there are some pegs between fingers to limit the palm's stretching, translation and rotation. These pegs separate the fingers, forming holes between the forefinger and the middle finger, and between the ring finger and the little finger. In this paper, we use the preprocessing technique described in [10] to align the palmprints. In this technique, the tangent of these two holes are computed and used to align the palmprint. The central part of the image, which is 128×128, is then cropped to represent the whole palmprint. Such preprocessing greatly reduces the translation and rotation of the palmprints captured from the same palms. Figure 1 shows a palmprint and its cropped image.

The rest of this paper is organized as follows. Section 2 discusses POC extraction. Section 3 presents the similarity measurement of POCs. Section 4 contains some experimental results. And in Section 5, we provide some conclusions.



(a) Original Palmprint (b) Cropped Image

Fig. 1. An example of the palmprint and its cropped image

2 Orientation Code Extraction

In this section, we devise several directional templates to define the orientation of each pixel. The 0°-directional template is devised as below:

$$T_{0^\circ} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{1}$$

And the α-directional template (T_α) is obtained by rotate T_{0° with Angle α.

Denote I as an image. The magnitude in the direction α of I is defined as

$$M_\alpha = I * T_\alpha \tag{2}$$

where "$*$" is the convolution operation. M_α is called the α-directional magnitude (α-DM).

Since the gray-scale of a pixel on the palm lines is smaller than that of the surrounding pixels which are not on the palm lines, we take the direction in which the magnitude is minimum as the orientation of the pixel. That is, the orientation of Pixel (i, j) in Image I is computed as below:

$$O(i, j) = \arg\min_{\forall \alpha} M_\alpha(i, j) \tag{3}$$

O is called the Palmprint Orientation Code (POC). Four directional templates (0°, 45°, 90° and 135°) are used to extract the POC in this paper.

The size of the preprocessed palmprint is 128×128. Extra experiments shows that the image with 32×32 is enough for the POC extraction and matching. Therefore, before compute the POC, we resize the image from 128×128 to 32×32. Hence the size of the POC is 32×32.

Figure 2 shows an example of the POCs, in which (a) is the original palmprint, (b) is POC (the different orientations are represented by the different grayscales) and (c)–(f) are the pixels with the orientation 0°, 45°, 90° and 135°, respectively. This figure shows that the POC keeps the most information of the palm lines.

Figure 3 shows some examples of the POCs, in which (a) and (b) are from a palm while (c) and (d) are from another palm, and (e)-(h) are their POCs. According to this figure, POCs from the same palms are very similar while the ones from different palms are quite different.

3 Similarity Measurement of POC

Because all POCs have the same length, we can use Hamming distance to define their similarity. Let C_1, C_2 be two POCs, their Hamming distance ($H(C_1, C_2)$) is defined as the number of the places where the corresponding values of C_1 and C_2 are different.

The matching score of two POCs C_1 and C_2 is then defined as below:

$$S(C_1, C_2) = 1 - \frac{H(C_1, C_2)}{N} \tag{4}$$

where $H(C_1, C_2)$ is the Hamming distance of C_1 and C_2; N is the size of a POC. Actually, $S(C_1, C_2)$ is the percentage of the places where C_1 and C_2 have the same orientation.

Obviously, $S(C_1, C_2)$ is between 0 and 1 and the larger the matching score, the greater the similarity between C_1 and C_2. The matching score of a perfect

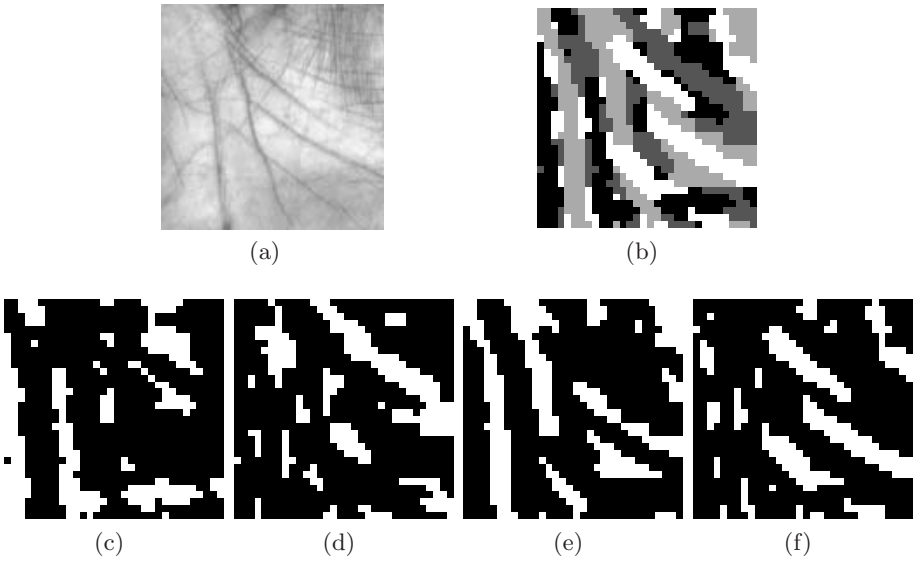


Fig. 2. An example of POC. (a) the original palmprint. (b) POC; (c)–(f) the pixels with the orientation 0°, 45°, 90° and 135°, respectively

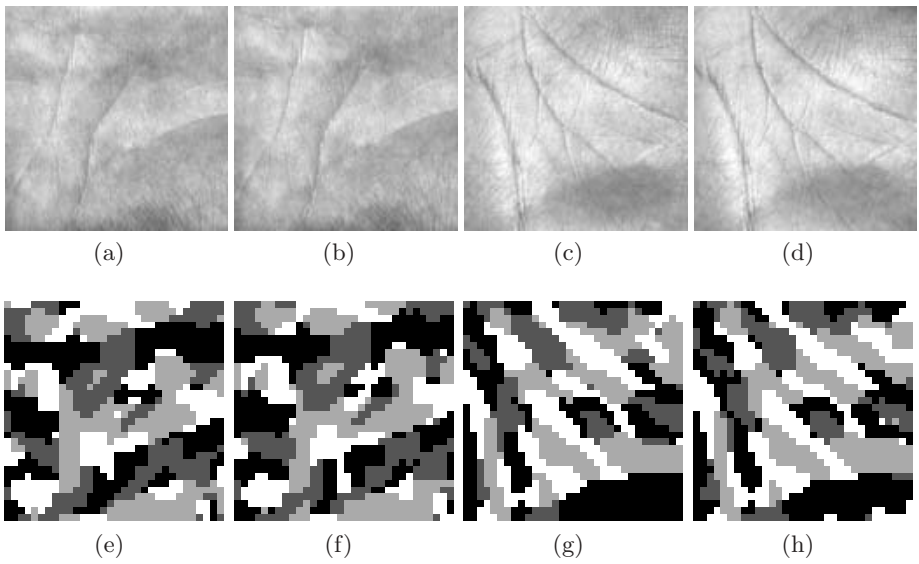


Fig. 3. Some examples of POCs. (a) and (b) are two palmprint samples from a palm; (c) and (d) are two palmprint samples from another palm; (e)-(h) are the POCs of (a)-(d), respectively

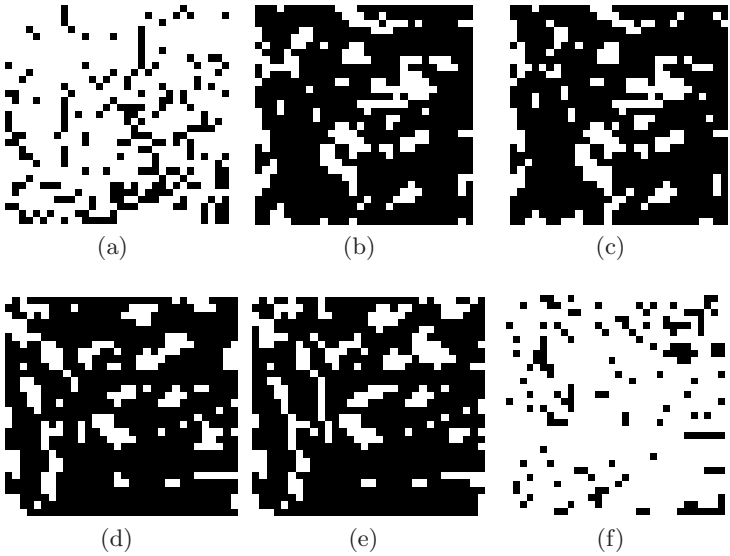


Fig. 4. Matching Results of Figure 3. (a)-(f) are the matching results of Figure 3(e) and 3(f), Figure 3(e) and 3(g), Figure 3(e) and 3(h), Figure 3(f) and 3(g), Figure 3(f) and 3(h), Figure 3(g) and 3(h), respectively

match is 1. Figure 4 is the matching results of Figure 3. In this figure, the white points of the images represent that the orientation of the corresponding places in C_1 and C_2 are same. Their matching scores are listed in Table 1. This figure and table show that the matching scores of the POCs from the same palms are much larger than that of the ones from different palms.

Table 1. Matching Scores of the POCs in Figure 2

No. of POCs	Figure 3(e)	Figure 3(f)	Figure 3(g)	Figure 3(h)
Figure 3(e)	1	0.80	0.21	0.21
Figure 3(f)	-	1	0.21	0.23
Figure 3(g)	-	-	1	0.86
Figure 3(h)	-	-	-	1

4 Experimental Results

We employed the PolyU Palmprint Database [12] to test our approach. This database contains 600 grayscale images captured from 100 different palms by a CCD-based device. Six samples from each of these palms were collected in two sessions, where three samples were captured in the first session and the other three in the second session. The average interval between the first and the second collection was two months. Some typical samples in this database are shown in Figure 5, in which the last two samples were captured from the same palm at

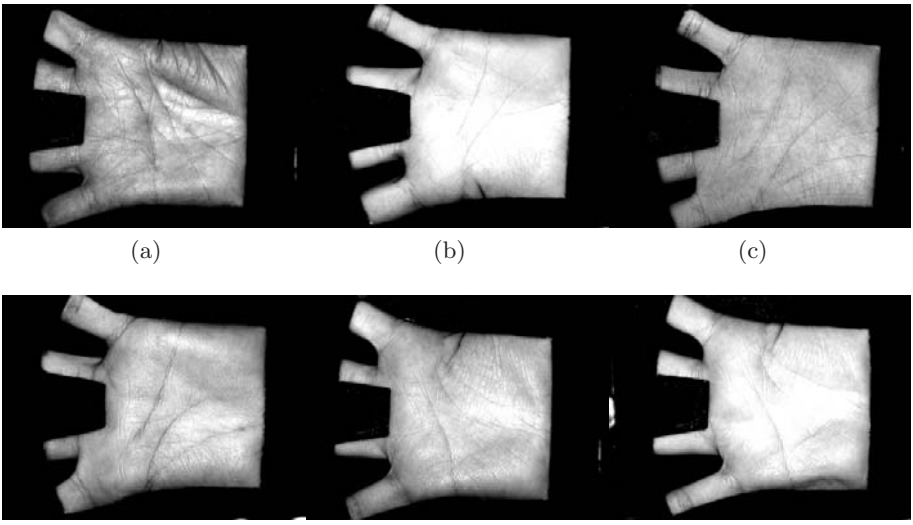


Fig. 5. Some typical samples in the Polyu Palmprint Database

different sessions. According to this figure, the lighting condition in different sessions is very different.

In order to investigate the performance of the proposed approach, each sample in the database is matched against the other samples. The matching between palmprints which were captured from the same palm is defined as a genuine matching. Otherwise, the matching is defined as an impostor matching. A total of $179,700$ $(600 \times 599/2)$ matchings have been performed, in which 1500 matchings are genuine matchings. Figure 6 shows the genuine and impostor matching scores distribution. There are two distinct peaks in the distributions of the matching scores. One peak (located around 0.7) corresponds to genuine matching scores while the other peak (located around 0.3) corresponds to impostor matching scores. The Receiver Operating Characteristic (ROC) curve, which plots the pairs (FAR, FRR) with different thresholds, is shown in Figure 7. For comparisons, the FusionCode method [11], which is an improvement of the PalmCode algorithm [10], is also implemented on this database. In the FusionCode method, each sample is also matched with the others. The ROC curve of the FusionCode method is also plotted in Figure 7 and the corresponding equal error rates (EERs) are listed in Table 2. According to the figure, the whole curve of the POC approach is below that of the FusionCode method, which means that

Table 2. EERs of Different Palmprint Recognition Methods

Method	POC	FusionCode
EER (%)	0.73	0.77

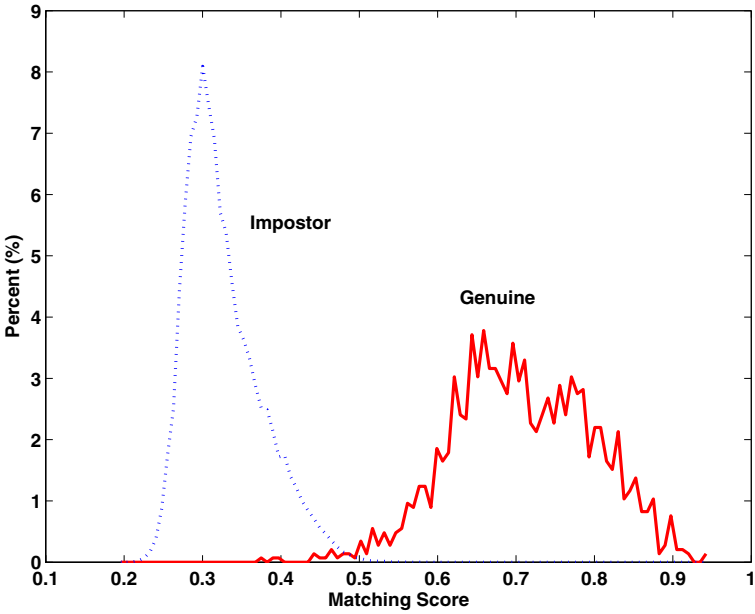


Fig. 6. The Distributions of Genuine and Impostor Matching Scores

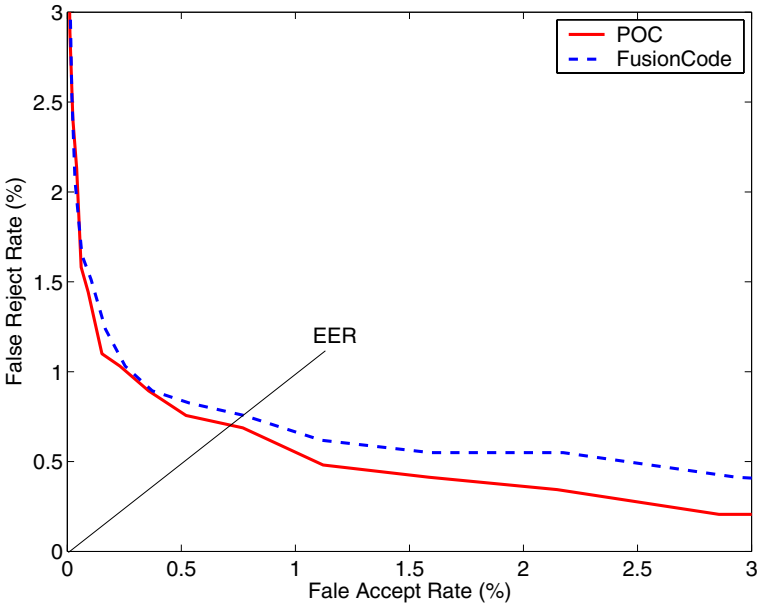


Fig. 7. The ROC Curve of the Proposed Approach

the performance of the proposed approach is better than that of the FusionCode method. In the future, we will investigate the fusion of the POC and the FusionCode, which is expected to further improve the performance of both the POC approach and the FusionCode method.

5 Conclusion and Future Work

A novel approach to palmprint authentication is presented in this paper. The palmprint orientation code (POC) is extracted using four directional templates. The similarity of the POCs is defined using their Hamming distance. The experimental results clearly shows the effectiveness of this approach.

In future, we will test the POC approach on a large database and investigate the fusion of the POC and FusionCode.

Acknowledgements

This work is supported by National Natural Science Foundation of China (60441005).

References

1. Zhang, D.: Automated Biometrics–Technologies and Systems. Kluwer Academic Publishers (2000)
2. Jain, A., Bolle, R., Pankanti, S.: Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers (1999)
3. Jain, A., Hong, L., Bolle, R.: On-line fingerprint verification. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 302–313
4. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.: Fvc2000: Fingerprint verification competition. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 402–412
5. Wildes, R.: Iris recognition: an emerging biometric technology. Proceedings of the IEEE **85** (1997) 1348–1363
6. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transaction on Circuit and System for Video Technology **14** (2004) 4–20
7. Han, C., Chen, H., Lin, C., Fan, K.: Personal authentication using palm-print features. Pattern Recognition **36** (2003) 371–381
8. Kumar, A., Wong, D., Shen, H., Jain, A.: Personal verification using palmprint and hand geometry biometric. Lecture Notes in Computer Science **2688** (2003) 668–678
9. Wu, X., Wang, K., Zhang, D.: Fisherpalm based palmprint recognition. Pattern Recognition Letters **24** (2003) 2829–2838
10. Zhang, D., Kong, W., You, J., Wong, M.: Online palmprint identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003) 1041–1050
11. Kong, A.W., Zhang, D.: Feature-level fusion for effective palmprint authentication. Internatrional Conference on Biometric Authentication, LNCS **3072** (2004) 761–767
12. Zhang, D.: Polyu palmprint palmprint database. http://www.comp.polyu.edu.hk/biometrics/

A Novel Palm-Line Detector

Li Liu and David Zhang

Biometric Research Centre, Department of Computing,
The Hong Kong Polytechnic University, Kowloon, Hong Kong
{csliliu,csdzhang}@comp.polyu.edu.hk

Abstract. Palmprint is a new emerging biometric feature for personal recognition. Palm lines, which consist of principal lines and wrinkles, are stable and essential traits for palmprint-based individual identification and can be extracted in low-resolution images. Therefore, it is the natural and reliable way to extract palm lines for personal authentication. However, the research on palm-line detection has done little. Due to special properties of palmprint, in addition to the structure feature, width of the palm-line, which generally reflects strength information, is important to identify palms especially when various palmprints have similar structures. In this paper, a novel palm-line detector is proposed to simultaneously extract structure and strength features of palm lines by minimizing a local image area which is of similar brightness to each individual pixel. A stable and sensible similarity function for brightness comparison is used to obtain the initial line responses. In order to give smooth and isotropic responses, a Gaussian weighting mask is defined as the local image area. Further, the relation between the size of Gaussian weighting mask and the width of detected palm lines is analyzed. The presented method has been tested on the PolyU Palmprint Database. Experimental results illustrate the effectiveness of this palm-line detector.

1 Introduction

The need for automatic human authentication is ever increasing in today's information and networked society. Biometrics has become an important and powerful means [1-3]. Palmprint, as a relatively new and developing biometric trait, has several advantages compared with others: low cost capture device, low-resolution imaging, non-fake, stable line feature and easy self positioning, etc. Palmprint authentication is drawing more and more researchers' attention [4-10].

Palm-line based palmprint identification schemes, such as interesting points, line segments and line features, have been presented. In the offline palmprint verification, Zhang *et al.* [4] approximated line features by using several straight-line segments. Duta *et al.* [5] extracted a set of feature points along palm lines and the associated line orientation to represent line features. You *et al.* [6] detected interesting points by applying the Plessey operator, a corner detector, to achieve higher performance than edge points by eliminating the redundancy. Han et al. [7] extracted the line-like features of a palmprint by using a morphological edge detector. In the online palmprint recognition, You et al. [8] further extracted the "interest" lines based on the interesting points by applying a fuzzy rule. Zhang et al. [9] obtained a set of statistical features of palm lines to characterize a palmprint by employing the overcomplete wavelet expansion. However, none of these offline and online palmprint recognition methods directly extracted palm lines which are fundamental and most important to characterize a palm.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 563–571, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Observing palmprint, we can find some principal lines, wrinkles and ridges on a palm. Usually, there are three principal lines in a palm which are most notable and vary little over time. Wrinkles are generally much thinner than principal lines and much more irregular. Ridges exist all over the palm just like the ridges do in a fingerprint and cannot be observed in low-resolution images. Palm lines, *which refer to principal lines and wrinkles*, are stable and reliable for individual identification and can be obtained from a low-resolution palmprint image. Fig. 1 shows an example of palmprint images captured by a CCD camera based on a special device for online palmprint acquisition [10]. From Fig.1 we can see that palm lines have their own properties. Firstly, palm lines are negative lines on which pixels' intensities are lower than neighbours'. Secondly, palm lines, especially principal lines, consist of many short line segments and curves which form many corners, junctions, branches, rings and chains. Finally, principal lines and some thick wrinkles are very strong and look *wider* than other light palm lines. That is, palm lines have different width which generally reflects strength information. The strength feature of palm lines is very important to describe a palmprint clearly especially when different palmprints have similar line structures. Hence, not only structure features but also strength characters of palm lines are necessary and important for palmprint recognition. However, the existing line or edge detection methods only focus on extracting the structure feature which is not enough for palmprint identification. Therefore, it is necessary to design a line detector to simultaneously extract structure and strength features of palm lines.

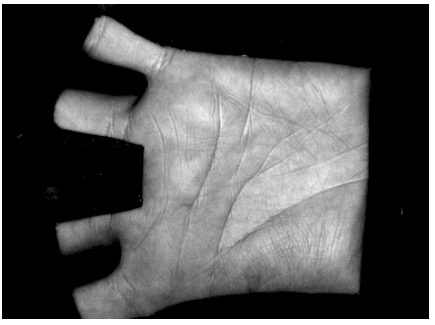


Fig. 1. Example of palmprint images with principal lines and wrinkles

In this paper, a novel palm-line detector is proposed to simultaneously extract palm-line structure and strength features by minimizing a local image area which is of similar brightness to each individual pixel. It is implemented by using circular masks to give isotropic responses. Digital approximations to circles are realized with Gaussian weighting. A stable and sensible similarity function for brightness comparison is used to obtain the local image area within circular masks. The line response is determined by the inverted area. An analysis of the relation between the size of the Gaussian weighting mask and the width of detected lines is also given.

The paper is organized as follow. Section 2 describes the main idea of the palm-line detector. In Section 3, the palm-line detector is defined and the parameter selection is discussed. Finally, we summarize the experimental results and draw conclusions in Section 4 and 5, respectively.

2 The Principle for the Palm-Line Detector

The principle of the proposed palm-line detector is illustrated in Fig. 2 showing a dark line on a white background. A circular mask, having a center pixel called "nucleus", is shown at six image positions. The detector firstly examines the intensity of the nucleus of the mask and counts pixels that have similar brightness to the nucleus into a 'univalue segment assimilating nucleus' (USAN) [11]. The USAN is at a maximum when the nucleus lies in a flat region of the image, decreases to about half of this maximum very near a straight edge, and decreases even further when inside a corner. That is, the smaller the USAN area is, the larger edge or corner response is given, as mentioned in [11]. Thereby, in order to detect a line, the USAN area of any pixel on the line should be less than those of background pixels. In other words, when the maximum of USAN areas of line pixels is less than those of background pixels, the line will be detected completely.

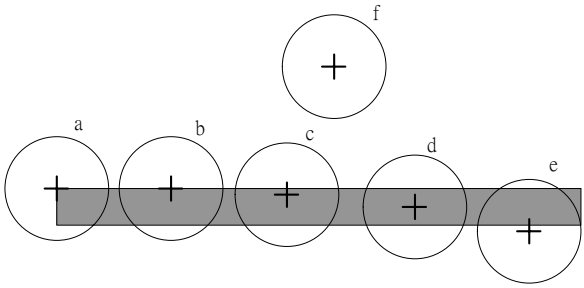


Fig. 2. Five circular masks at different positions of a line with width of 3

In Fig. 3, a test image with a 3 pixels wide line has been processed to give USAN area as output. All of USAN areas can be divided into two groups: one is obtained from pixels on the line and the other is from background. From the three dimensional plot, we can see that the USAN of the line reaches a local maximum when the line passes through the center of the circular mask, just like the mask *d* shown in Fig. 2, while the USAN of background is at a local minimum when the nucleus of the circular mask is very near the edge of the line. Obviously, in order to completely extract the line, the local maximum of USAN areas given by pixels on the line should be less than the local minimum obtained from background. This can be realized by using a circular mask with a proper radius.

According to the above analysis and observation of examples, the line detector principle can be formulated: lines are enhanced by giving inverted USAN area and can be completely extracted by using a mask with a proper size.

3 The Palm-Line Detection Approach

3.1 Palm-Line Detector

Denote the palmprint image as *I*. The palm-line detector is defined as follows:

$$L(x_0, y_0) = \begin{cases} g - s(x_0, y_0) & \text{if } s(x_0, y_0) < g \\ 0 & \text{otherwise} \end{cases}, \quad g = s_{\max} / 2 \tag{1}$$

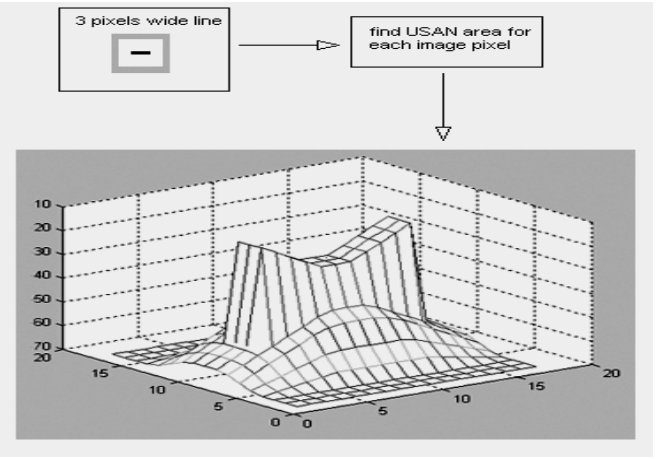


Fig. 3. A three dimensional plot of USAN area given a test image with a 3 pixels wide line

$$s(x_0, y_0) = \sum_{\substack{y_0 - r \le y \le y_0 + r \\ x_0 - r \le x \le x_0 + r}} w(x, y, x_0, y_0) \tag{2}$$

$$w(x, y, x_0, y_0) = \begin{cases} G(x, y, x_0, y_0, r) \bullet e^{-\left(\frac{I(x,y) - I(x_0, y_0)}{t}\right)^6} & \text{if } I(x, y) \ge I(x_0, y_0) \\ G(x, y, x_0, y_0, r) & \text{otherwise} \end{cases} \tag{3}$$

$$G(x, y, x_0, y_0, r) = e^{-\frac{(x - x_0)^2 + (y - y_0)^2}{r^2}} \tag{4}$$

where (x_0, y_0) is the coordinate of the nucleus, (x, y) is the coordinate of any other pixel within the mask, $s(x_0, y_0)$ is the USAN area and $s_{\max}$ is the maximum value which s can take, t is the brightness difference threshold, r is the radius of the Gaussian weighting mask, $w(x, y, x_0, y_0)$ is the output of brightness comparison weighted by a Gaussian function $G(x, y, x_0, y_0, r)$, and $L(x_0, y_0)$ is the line response.

This is a formulation of the principle for the palm-line detector. According to Eq. (3), if surrounding pixels' brightness is less than the nucleus, i.e., the nucleus is brighter than its neighbors, these pixels will be directly counted into USAN to decrease the line response of the nucleus, which can eliminate the affection of local exposure in palmprint images. In Eq. (3), a stable and sensible function $e^{-\left(\frac{I(x,y) - I(x_0, y_0)}{t}\right)^6}$ is used to count USAN which allows a pixel's brightness not to have too large an effect on w even if it is near the threshold t. The use of the sixth power has been proved to be the theoretical optimum in [11]. In order to give smooth and isotropic responses, a 2-D Gaussian mask is used as the weight function of brightness comparison. Fig. 4 shows the segmented palmprint images and the corresponding palm-line detection results.

Fig. 5 shows palm-line detection results of a segmented palmprint image by using Gaussian weighting masks with different radius. It can be seen that with the increase of the mask radius, more and more palm-line strength features (or width information)

can be extracted. In other words, the higher degree of completeness of detected palm lines can be achieved as the radius of Gaussian weighting mask increases. When using a Gaussian mask with a small size such as 3-pixel radius, only structure features of palm lines can be detected as shown in Fig. 5 (b). As the radius of Gaussian mask increases, the detected palm lines, especially thick lines in original palmprint images, become thicker and close to the real width of palm lines. Therefore, palm-line structure and strength features can be extracted simultaneously by using the proposed palm-line detector and, further, the detected strength features are determined by the size of Gaussian weighting mask.

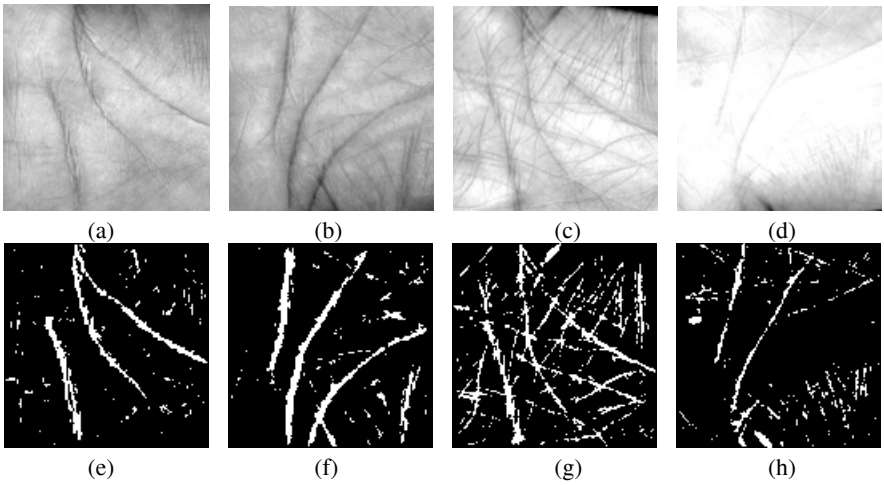


Fig. 4. The segmented palmprint images (a, b, c, d) and the corresponding palm-line detection results (e, f, g, h) by using the proposed approach

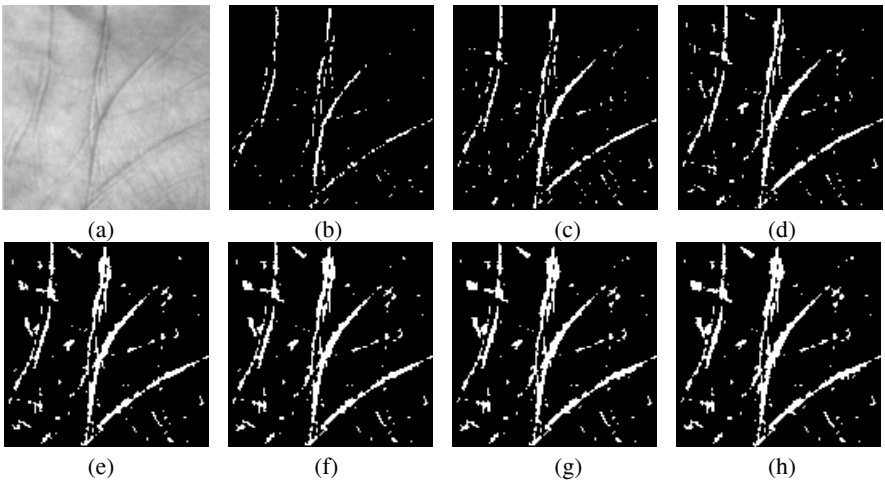


Fig. 5. Palm-line extraction by using the proposed negative palm-line detector; (a) segmented palmprint image, palm-line extracted images by using Gaussian weighting mask with radius of 3 in (b), 5 in (c), 7 in (d), 9 in (e), 10 in (f), 11 in (g), 12 in (h)

3.2 Parameter Selection

The Gaussian weighting mask determines the strength features, i.e., the width of palm lines, detected by using the proposed approach. Therefore, the radius of the Gaussian weighting mask is an important parameter to this line detection approach. In this section, a brief analysis of the parameter selection is given to show the relationship between the width of detected palm lines and the size of Gaussian weighting mask.

As mentioned in section 2, the USAN area on a palm line reaches a local maximum when the line passes through the center of the circular mask. Assume that a circle C with a radius r has a Gaussian density $e^{-\frac{x^2+y^2}{r^2}}$ and a line of width $2\times w$ traverses the center of the circle as shown in Fig. 6. Let L denote the part of the line within the circle. According to the definition of the palm-line detector, if a line of width $2\times w$ is fully detected by using a Gaussian mask with radius r, it requires:

$$\iint_L e^{-\frac{x^2+y^2}{r^2}} dxdy < \frac{1}{2}\iint_C e^{-\frac{x^2+y^2}{r^2}} dxdy , \tag{5}$$

which is equivalent to,

$$\int_0^w \int_0^{\sqrt{r^2-x^2}} e^{-\frac{x^2+y^2}{r^2}} dxdy < \frac{1}{2}\int_0^r \int_0^{\sqrt{r^2-x^2}} e^{-\frac{x^2+y^2}{r^2}} dxdy . \tag{6}$$

The right hand side can be simplified as

$$\begin{aligned}\int_0^r \int_0^{\sqrt{r^2-x^2}} e^{-\frac{x^2+y^2}{r^2}} dxdy &= \int_0^{\frac{\pi}{2}} d\theta \int_0^r e^{-\frac{\rho^2}{r^2}} \rho d\rho \\ &= \frac{\pi}{4} r^2 (1-e^{-1}) .\end{aligned} \tag{7}$$

Therefore,

$$\int_0^w \int_0^{\sqrt{r^2-x^2}} e^{-\frac{x^2+y^2}{r^2}} dxdy < \frac{\pi}{8} r^2 (1-e^{-1}) . \tag{8}$$

The relationship between the width of detected line and the radius of Gaussian mask can be determined based on Eq. (8). Given the mask with radius r, the critical value of detected line of width $2\times w$ is obtained when the left and right arguments of Eq. (8) are equal. As the analytic form of the left function is not available, we only give the approximate critical value of detected line of width, as shown in Table 1.

Table 1. The relation between the radius of Gaussian mask and the critical width of detected line

Radius of Gaussian mask	3	4	5	6	7	8	9	10	11	12	13
Approximate critical width of line	2.1	2.8	3.5	4.2	4.9	5.6	6.3	6.9	7.7	8.4	9.1
Digital approximation to width of line	2	2	3	4	4	5	6	6	7	8	9

Therefore, given the mask with radius r, a palm line can be definitely detected if the width is not larger than the corresponding critical value of detected line of width.

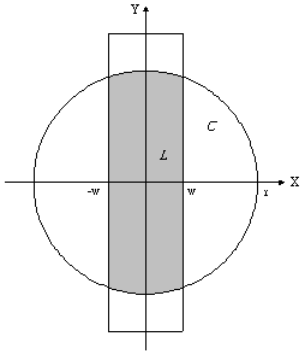


Fig. 6. A line of width $2\times w$ passing through the centre of a circle with radius r

4 Experimental Results

The proposed palm-line detector has been tested on the public palmprint database [12] built by the Biometric Research Center at the Hong Kong Polytechnic University. This database contains 600 palmprint images from 100 different palms. Six samples from each of these palms were collected in two sessions, where 3 samples were captured in the first session and the other 3 in the second session. The average interval between the first and the second collection was two months. The size of these images is 384×284. In pre-processing, key point detection approach [10] is used to align different palmprint images by segmenting a central part (128×128 pixels) which is used for palm-line detection. In the stage of palm-line detection, Eq. (3) is implemented as a look up table for speed. The threshold t, which determines the minimum contrast of edges which will be picked up, set to 10 by experience. First, we introduce the similarity measurement for palm-line matching, followed by palm-line verification.

4.1 Palm-Line Matching

Let P and Q be two palm-line images which are logical matrices. In order to provide translation invariance matching, the matching score between P and Q is defined as below:

$$M_{\max} = \max_{|s|\le sr,|t|\le sr} \frac{2}{M_P + M_Q} \times \sum_{i=\max(1,1-s)}^{\min(N,N-s)} \sum_{j=\max(1,1-t)}^{\min(N,N-t)} P(i,j) \cap Q(i+s, j+t) \tag{9}$$

Where $\cap$ is the logical "AND" operation; $\boldsymbol{M_P}$ and $\boldsymbol{M_Q}$ are the number of pixels on detected palm lines in P and Q, respectively; $\boldsymbol{N\times N}$ is the size of palm-line image; sr, the search radius set to 6, controls the range of translation in the matching process. Obviously $M_{\max}$ is between 0 and 1. For perfect matching, the matching score is 1.

4.2 Palm-Line Verification

Each palmprint in the public database was matched with the other palmprints in the same database. Before matching, a post processing has been done to discard weak

line responses and to smooth detected palm-line images by using a Gaussian filter with standard deviation σ.

A matching was labeled correct if the matched palmprint was among the five other palmprints of the same individual, and incorrect otherwise. A total of 179,700 (600×599) matchings have been performed on the public database including 1,500 genuine's matching attempts and 178,200 imposter's matching attempts. Table 2 gives the equal error rates (EER) corresponding Gaussian weighting masks with different radii (r) and Gaussian smooth filters with different standard deviations (σ). Fig. 7 shows the relationship between the EER and the radius of a Gaussian weighting mask. It can be seen that the EER reaches at the minimum 1.00% by using a weighting mask with radius 12 and a Gaussian smooth filter with standard deviation 1.0. The corresponding probability distributions for genuine and imposter and the Receiver Operating Characteristic (ROC) curve are shown in Fig. 8 (a) and (b), respectively. Considering Table 1, we can conclude that based on the proposed negative palm-line detector, a Gaussian weighting mask with radius 12 is proper for palmprint authentication.

Table 2. Equal error rates (%) based on Gaussian weighting masks with different radius (r) and Gaussian smooth filters with different standard deviation (σ)

σ \ r	6	7	8	9	10	11	12	13
0	1.67	1.68	1.60	1.60	1.54	1.34	1.26	1.25
0.5	1.74	1.72	1.55	1.29	1.21	1.30	1.26	1.18
0.75	1.47	1.44	1.26	1.31	1.36	1.25	1.27	1.29
1.0	**1.30**	1.38	**1.22**	1.24	1.23	**1.10**	**1.00**	**1.11**
1.25	1.32	**1.25**	1.22	**1.22**	**1.19**	1.11	1.10	1.11

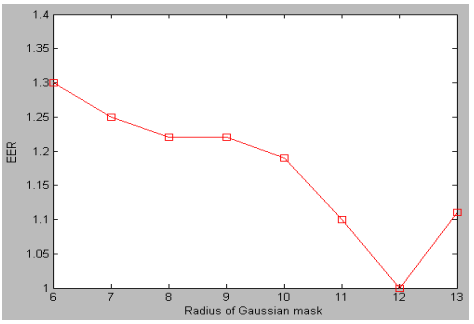
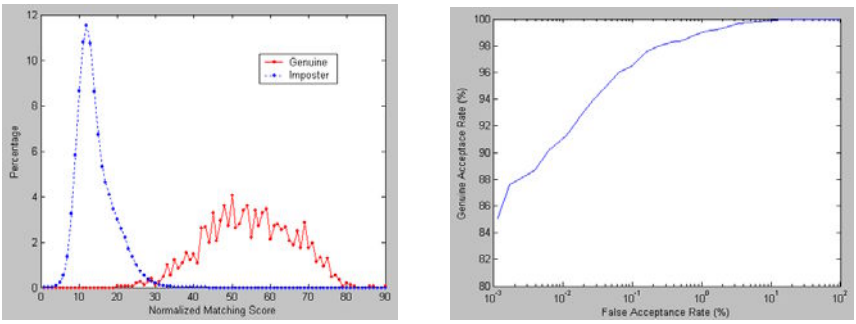


Fig. 7. The relationship between the EER and the radius of Gaussian weighting mask

5 Conclusions

We have presented a palm-line detector to simultaneously extract structure and strength features of palm lines. This line detector is based on the brightness comparison in a local area. A stable and sensible function is introduced to obtain line response. In order to give a smooth and isotropic response, a Gaussian weighting mask



(a) (b)

Fig. 8. Verification test results. (a) Genuine and imposter distributions and (b) the receiver operator characteristic curve

is used as the local area for brightness comparison. An analysis of the algorithm shows the size of the Gaussian weighting mask determines the degree of completeness of extracted palm lines. Experimental results show that a Gaussian weighting mask with radius 12 is proper for palm-line based palmprint recognition. The low EER (1%) represents the effectiveness of the proposed palm-line detectior.

References

1. Jain A., Bolle R., Pankanti S .(ed.): Biometrics: Personal Identification in networked Society. Kluwer Academic, Dordrecht, 1999
2. Zhang D.: Automated Biometrics-Technologies and Systems. Kluwer Academic, Dordrecht, 2000
3. Zhang D. (ed.): Biometric Solutions for Authentication in an e-World. Kluwer Academic, Dordrecht, 2002
4. Zhang D., Shu W.: Two novel characteristics in palmprint verification: datum point invariance and line feature matching. Pattern Recognition 32 (1999) 691-702
5. Duta N., Jain A.K., Mardia K.V.: Matching of palmprints., Pattern Recognition Letters 23 (2002) 477-485
6. You J., Li W., Zhang D.: Hierarchical palmprint identification via multiple feature extraction. Pattern Recognition 35 (2002) 847-859
7. Han C.C., Cheng H.L., Fan K.C., Lin C.L.: Personal authentication using palmprint features. Pattern Recognition 36 (2003) 371-381
8. You J., Kong W.K., Zhang D., Cheung K.H.: On hierarchical palmprint coding with multiple features for personal identification in large databases. IEEE Trans. Circuits and Systems for Video Technology 14 (2004) 234-243
9. Zhang L., Zhang D.: Characterization of palmprints by wavelet signatures via directional context modeling. IEEE Trans. SMC-B 34 (2004) 1335-1347
10. Zhang D., Kong W.K., You J., Wong M.: Online Palmprint Identification. IEEE Trans. PAMI 25 (2003) 1041-1050
11. Smith S.M., Brady J.M.: SUSAN – A new approach to low level image processing. Int. J. of Computer Vision 23 (1997) 45-78
12. PolyU Palmprint Palmprint Database, http://www.comp.polyu.edu.hk/~biometrics/

A New On-Line Model Quality Evaluation Method for Speaker Verification

Javier R. Saeta[1] and Javier Hernando[2]

[1] Biometric Technologies, S.L., 08007 Barcelona, Spain
j.rodriguez@biometco.com
[2] TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain
javier@talp.upc.es

Abstract. The accurate selection of the utterances is very important to obtain right estimated speaker models in speaker verification. In this sense, it is important to determine the quality of the utterances and to establish a mechanism to automatically discard or accept them. In real-time speaker verification applications, it is decisive to obtain on-line measures to ask the speaker for more data if necessary. In this paper, we introduce a new on-line quality method based on a male and a female Universal Background Model (UBM). These two models act as a reference for new incoming utterances in order to decide if they can be used to estimate the speaker model or not. Text-dependent experiments have been carried out by using a telephonic multi-session database in Spanish. The database has been recorded by the authors and has 184 speakers.

1 Introduction

In a Speaker Verification (SV) system, the user enrolls the system by pronouncing some utterances in order to estimate a speaker model. The enrollment procedure is one of the most critical stages of a SV process. At the same time, it becomes essential to carry out a successful training process to obtain a good performance. The importance and sensitiveness of the process force us to pay special attention on it. Consequently, it is necessary to protect the enrollment procedure by giving the user some security mechanisms, like extra passwords or by providing a limited physical access. There are a lot of cases where the training process is vulnerable. One of the most common ones is when the enrollment is done by phone.

In such cases, occasional impostors could seriously damage the speaker model especially if we are training the model in several sessions. This could also be applied to models which are adapted from new utterances coming from the speaker.

But sometimes background noises, distortions or heavy colds can produce similar effects to the ones with real impostors. For this reason, it is convenient to control the quality of the speaker utterances to detect low quality ones and prevent from errors in the estimation of the speaker model.

The quality of a model mainly depends on the reliability and variability of the utterances and on the training and test conditions. It is crucial that the speaker model includes the most discriminative speaker characteristics. In real applications, one can normally afford one or two enrollment sessions only. In this context, it is important to control the content and quality of the recorded voice samples, when the enrollment process is 'open', i.e., when the speaker is talking and the utterances are being recorded.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 572–579, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Quality model measures evaluate how discriminative a model is by comparing client and/or impostor utterances against the model. Some approaches to the problem of model quality evaluation have traditionally dealt with outliers, i.e., those client scores which are distant with respect to the mean in terms of Log-Likelihood Ratio (LLR). They use the distance between the training model and the utterances used to estimate the model. The 'leave-one-out' method [1] has the problem of an excessive computational cost. The Z method [2] uses impostor data. The method introduced by the authors in [3] overcomes these two problems but, as it happens with the first two methods, it needs the speaker model to evaluate quality.

In this paper, we introduce a new on-line quality method to detect non-profitable or non-representative utterances coming from an impostor or from the own speaker. When an undesired utterance is located, the system asks the user for a new one. The method compares an utterance against a male and a female UBM, previously estimated from a collected corpus. Two scores are obtained. These scores are used to locate the utterance with respect to the UBMs. In principle, utterances from the same speaker are similar enough between them so when a new utterance is compared against the UBMs, the score should be similar to the ones obtained before for the rest of the speaker utterances. This is the basis of the on-line quality model method.

A theoretical view of the state-of-the-art is reported on the next section. New proposers are developed in section 3. The experimental setup and the evaluation with empirical results are described in section 4, followed by conclusions in section 5.

2 Theoretical Approach

Some approaches have been previously shown in literature concerning the evaluation of quality models. In [1], a model quality checking method called 'leave-one-out' is introduced. It uses N-1 utterances from a total of N utterances to train the model. N scores are obtained by testing every utterance against the model. The model that yields the highest score on the test utterance is the most representative model. The lowest scores belong to utterances which can be considered as outliers. The main problem of the 'leave-one-out' is that it estimates the model N times to detect the best representative one. It implies a huge computational cost.

Another different approach [2] to check model quality introduces the distance Z between LLR scores from clients and from impostors for a given model:

$$Z = \frac{\max\left\{0, \mu_C - \mu_I\right\}}{\sigma_I} \tag{1}$$

where μ_C is the mean LLR score on client utterances of the given model and μ_I and σ_I are, respectively, the mean and standard deviation of LLR scores on a set of impostor utterances. Z shows how discriminative a model is. If Z is close to zero, a low discrimination is expected. Z method has the problem of using data from impostors and it is common to deal only with client data in some applications.

Another measure which has been introduced by the authors in [3] uses an algorithm to determine the quality level of a speaker model. This algorithm decides if an utterance is a good representation of the model according to an iterative process. We define s_n as a LLR score obtained by testing an utterance against its own model. We

assume that an utterance has an acceptable degree of quality when it surpasses the following interval:

$$s_n \geq \mu_C - \alpha\sigma_C \tag{2}$$

where μ_C and σ_C are the mean and standard deviation of LLR scores on the utterances used to train the model. The coefficient α is empirically determined.

The method is applied to the enrollment data in combination with an algorithm to find the less representative utterances for every speaker. Once these outliers are located, they can be suppressed or replaced by new ones coming from the same speaker. It classifies the speaker models according to their quality. The classification will detect reduced quality models. Models will be placed into different groups depending on the degree of similarity of their utterances with their respective models.

The use of this method with client data is especially useful when it is difficult to obtain data from impostors, for instance in phrase-prompted cases. When using words or phrases as passwords – except in connected digits-, this method will be generally more suitable than the one explained before which employed Z to determine the model discrimination, because that method used data from impostors.

On the other hand, in comparison with the 'leave-one-out' method, the last method is more effective in terms of computational cost. If N is the number of client model utterances, the 'leave-one-out' method trains N models per client to evaluate quality while the method showed in (2) trains a maximum of the whole part of N/5 models.

But the problem of the last method – and also of the first two ones – is that it is not possible to ask the user for new data until the model is estimated. And this inconvenience is especially critical when we use only one session for training or when we are in the second session of a two-session enrolment process. If there are some low quality utterances, we lost the opportunity of obtaining more voice samples from the speaker when (s)he is just recording them. It could lead to wrong estimated or undertrained models.

3 New On-Line Method

Until now, the evaluation of the quality of the speaker model took place once the model was estimated but the use of quality measures has several disadvantages then. The main problem is that, in real-applications, we do not have the option of asking for more utterances to the speaker. This is basically because modern systems use to train the speaker in one – maximum two – session(s). Furthermore, the number of utterances tends to be small.

In this case, even when we detect that an utterance has a bad quality or comes from an intentional impostor, it is not possible to ask the speaker for a new one.

With on-line model quality measures we solve this problem because we decide if an utterance has a sufficient degree of quality before estimating the model and, what is more important, before adding this utterance to the speaker model.

The algorithm works as follows:

1. Obtain LLR scores $\{s_{1m}, s_{2m}, s_{3m}...\}$ and $\{s_{1f}, s_{2f}, s_{3f}...\}$ from incoming utterances $\{U_1, U_2, U_3...\}$ against $\{UBM_m, UBM_f\}$
2. Estimate $\{\mu_m, \mu_f\}$ from the previous scores

3. Ask for a new utterance U_n and obtain $\{s_{nm}, s_{nf}\}$ against $\{UBM_m, UBM_f\}$
4. Calculate a distance $d_{mf} = |\mu_m - s_{nm}| + |\mu_f - s_{nf}|$
5. If $d_{mf} > \Theta$, quality is considered as sufficient. If $d_{mf} \leq \Theta$, then go to 3

First of all, we obtain a pair of scores for every utterance $\{U_1, U_2, U_3...\}$, one against a male UBM_m and another one against a female UBM_f. From the moment we obtain some new utterances, we estimate the mean $\{\mu_m , \mu_f\}$ for every pair of scores. Thus, a comparison takes place when new incoming utterances (U_n) are obtained for the speaker. They should not be far – in terms of LLR – from that estimated mean if they really belong to the speaker. The process is shown in the following scheme:

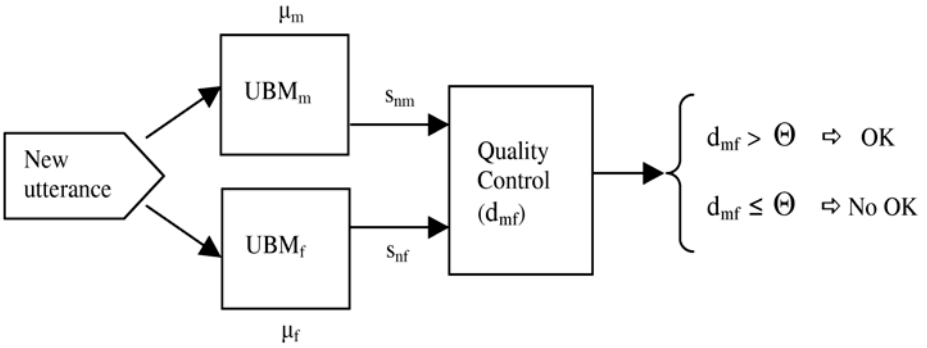


Fig. 1. Block diagram for the on-line quality algorithm

Finally, we set a maximum distance d_{mf} and reject utterances that surpass that distance because they have not reached the minimum degree of quality required, fixed by a threshold Θ. D_{mf} is a conventional distance which has been shown as suitable in our experiments. Of course, more work could be done to find a more optimized one.

The threshold Θ is empirically determined. It is obvious that the quality estimation becomes more robust if using as more utterances as possible to established the maximum distance allowed to considerate an acceptable degree of quality.

The on-line quality method has similarities to the Tnorm [4] normalization technique because the score is obtained on-line by comparing – in the Tnorm case – the test utterance to the client model and to some impostor models.

4 Experiments

4.1 Database

The database used in this work has been recorded by the authors and has been especially designed for speaker recognition. It includes land-line and mobile telephone sessions. 184 speakers were recorded by phone, 106 male and 78 female. It is a multi-session database in Spanish, with 520 calls from the Public Switched Telephone Network (PSTN) and 328 from mobile telephones. One hundred speakers have at least 5 or more sessions. The average number of sessions per speaker is 4.55. The average time between sessions per speaker is 11.48 days.

Each session includes:

a) 4 different sequences of 8-digit numbers, repeated twice.
b) 2 different sequences of 4-digit numbers, repeated twice.
c) 6 different isolated words.
d) 5 different sentences.
e) 1 minute long read paragraph.
f) 1 minute of spontaneous speech.

4.2 Experimental Setup

In our experiments, utterances are processed in 25 ms frames, Hamming windowed and pre-emphasized. The feature set is formed by 12^{th} order Mel-Frequency Cepstral Coefficients (MFCC) and the normalized log energy. Delta and delta-delta parameters are computed to form a 39-dimensional vector for each frame. Cepstral Mean Subtraction (CMS) is also applied.

Left-to-right HMM models with 2 states per phoneme and 1 mixture component per state are obtained for each digit. Client and world models have the same topology. The UBM for each digit is estimated with data from a balanced subset of speakers of the database, concretly those speakers which have recorded one to four enrollment sessions only (over 25 speakers).

The speaker verification is performed in combination with a speech recognizer for connected digits recognition. During enrollment, those utterances catalogued as "no voice" are discarded. This ensures a minimum quality for the threshold setting.

Clients have a minimum of 5 sessions. It yields 100 clients. We use 4 sessions for enrollment and the rest of sessions to perform client tests or for adding more data to the speaker in quality model experiments. Speakers non-included in digit UBMs with more than one session and less than 5 sessions are used as impostors. 4 – and 8-digit utterances are employed for enrollment and 8-digit for testing. We apply verbal information verification [5] as a filter to remove low quality utterances. The total number of training utterances per speaker goes from 8 to 48. The exact number depends on the number of utterances discarded by the speech recognizer. During test, the speech recognizer discards those digits with a low probability and selects utterances which have exactly 8 digits.

It is important to note that fixed-line and mobile telephone sessions are used indistinctly to train or test. This factor increases the error rate.

4.3 Verification Results

Our verification experiments with connected digits show the Equal Error Rate (EER) for the baseline, the 'leave-one-out' method [1], the method introduced in [3] and the on-line quality method introduced in this paper.

The 'leave-one-out' method has been used here without predefined thresholds. Like the other experimental methods presented here, it uses the Speaker Dependent Threshold (SDT) method of the following equation [6, 7]:

$$\Theta_x = \alpha\, \mu_I + (1-\alpha)\, \mu_C \tag{3}$$

where μ_C is the client scores mean, μ_I is the impostor scores mean and α is a constant which has to be optimized from a pool of speakers.

Table 1. Error rates for a set of speakers in connected digit verification experiments with SDT defined in (3)

Quality methods	EER (%)
Baseline	2.23
Leave-one-out	2.02
Without outliers	5.86
On-line method	2.00

As we can see in Table 1, the baseline experiments give an EER over 2%. The 'leave-one-out' method slightly improves the baseline experiments, but its enormous computational cost makes it unaffordable.

In the third method, an average of 2.3 utterances per speaker were removed for the 44 speakers with low quality. The error rates dramatically increased by removing only a few utterances considered as outliers. That reflects the importance of data when estimating a model. In our case, we have found that it is better to keep data even when we have realized that they are not the best representation of the speaker. This is especially important when we do not use too much data to estimate the speaker model or when the handsets for training and testing are different because it can cause errors in the selection of outliers.

On the other hand, in case we replace outliers by new and more representative data from the speaker, we reduce error rates by around 40% and the system performs better than the baseline (EER = 1.39%).

The on-line quality measure consists of a simulation for an enrollment procedure with 4 training sessions per speaker. The algorithm tests the quality of the utterances by means of the on-line quality method and decides if there are non-representative utterances. If the measure reveals bad quality utterances, they are replaced by new ones from the fifth session of the speaker. If the number of non-representative samples exceeds the number of valid utterances of the fifth session, bad quality utterances are removed anyway. In this case, some models are trained with a smaller number of utterances than initially – a reduction of 8% of the data with respect to the baseline. This increases the error rates.

The whole process can be done in real-time because the model is not estimated until the minimum number of utterances is reached. The use of on-line quality measure reduces the error although not very significantly because the threshold is estimated using impostor data. In this case, the influence of non-representative utterances can be better minimized than in cases when only material from clients is available. Furthermore, not every utterance discarded by the on-line method was replaced by a new one from the fifth session. Some of them could not be replaced because of the bad quality of the utterances of the fifth session for some speakers. Anyway, the on-line quality method has the advantage of determining the quality before the creation of the model.

The following table shows a comparison of the EER (%) for threshold estimation methods with client data only, without impostors:

The baseline SDT method for Table 2 is defined as follows [8]:

$$\Theta_x = \mu_C - \alpha\, \sigma_C \tag{4}$$

where μ_C is the client scores mean, σ_C is the standard deviation from clients and α is a constant empirically determined.

Table 2. Comparison of threshold estimation methods in terms of EER (%) with data from clients only

Quality methods	Baseline	On-line method
Baseline	5.89	4.50
Baseline + 2 impostor utterances	6.19	4.72

Two intentional impostor utterances per speaker are added here to the baseline during training to taint the enrollment process. We add two utterances from a male voice for men and two female utterances for women.

The on-line quality method discards the 94% of these utterances. At the same time and despite the presence of intentional impostors and the elimination of some training data, the on-line method reduces the error rate with respect to the baseline.

As we can see from table 2, the on-line measures, with and without 2 impostors, perform better than their respective baselines.

5 Conclusions

A new on-line model quality evaluation algorithm has been introduced here. It outperforms the 'leave-one-out' method in terms of computational cost and it has the advantage of using only data from clients, which is strongly recommended when dealing with words or phrases as passwords and it is difficult to obtain data from impostors.

The new algorithm has the advantage of estimating quality without needing the speaker model. This implies that the quality can be measured on-line. In our experiments, the method was capable of rejecting 94% of intentional impostor utterances while preserving client utterances. The best on-line quality performance was achieved with a threshold that used impostor data.

Although the improvement is not very sensitive when adding two impostor utterances, further work will show that if the number of intentional impostor utterances is increased, the use of the on-line quality evaluation method will result in a substantial improvement.

References

1. Gu, Y., Jongebloed, H., Iskra, D., Os, E., Boves, L.: Speaker Verification in Operational Environments-Monitoring for Improved Service Operation, ICSLP'00, Vol. II, 450-453, Beijing (2000).
2. Koolwaaij, J., Boves, L., Os, E. den, Jongebloed, H.: On Model Quality and evaluation in Speaker Verification, ICASSP'00, 3759-3762, Istanbul (2000).
3. Saeta, J.R., Hernando, J.: Model Quality Evaluation during Enrollment for Speaker Verification, 8th International Conference on Spoken Language Processing (ICSLP), 352-355, Jeju (South Korea), October (2004).
4. Auckentaler, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Verification Systems, Digital Signal Processing, Vol. 10, 42-54, 2000.
5. Li, Q., Juang, B.H., Zhou, Q., Lee, C.H.: Verbal Information Verification, Proc. Eurospeech'97, 839-842.

6. Pierrot, J.B., Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Blomberg, M., Bimbot, F.: A Comparison of A Priori Threshold Setting Procedures for Speaker Verification in the CAVE Project, Proc. ICASSP'98, 125-128.
7. Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Pierrot, J.B., Blomberg, M., Bimbot, F., Techniques for A Priori Decision Threshold Estimation in Speaker Verification, Proc. RLA2C, Avignon (1998) 89-92.
8. Saeta, J.R., Hernando, J.: Automatic Estimation of A Priori Speaker Dependent Thresholds in Speaker Verification, Proc. 4th International Conference in Audio- and Video-based Biometric Person Authentication (AVBPA), ed. Springer-Verlag, 70-77 (2003).

Improving Speaker Verification Using ALISP-Based Specific GMMs

Asmaa El Hannani[1,⋆] and Dijana Petrovska-Delacrétaz[1,2]

[1] DIVA Group, Informatics Dept., University of Fribourg, Switzerland
asmaa.elhannani@unifr.ch
[2] Institut National des Télécommunications, 91011 Evry, France
dijana.petrovska@int-evry.fr

Abstract. In recent years, research in speaker verification has expended from using only the acoustic content of speech to trying to utilise high level features of information, such as linguistic content, pronunciation and idiolectal word usage. Phone based models have been shown to be promising for speaker verification, but they require transcribed speech data in the training phase. The present paper describes a segmental Gaussian Mixture Models (GMM) for text-independent speaker verification system based on data-driven Automatic Language Independent Speech Processing (ALISP). This system uses GMMs on a segmental level in order to exploit the different amount of discrimination provided by the ALISP classes. We compared the segmental ALISP-based GMM method with a baseline global GMM system. Results obtained for the NIST 2004 Speaker Recognition Evaluation data showed that the segmental approach outperforms the baseline system. It showed also that not all of the ALISP units are contributing to the discrimination between speakers.

1 Introduction

Traditional speaker verification systems are limited to the use of frame-based spectral features that are basically modeled globally via Gaussian Mixture Models (GMM). In such systems the linguistic structure of the speech signal is not taken into account and all sounds are represented using a unique model. Hence the phoneme-specific information is ignored. Various studies [1–4] have shown that voiced phones and fricatives are the most effective broad speech classes for speaker discrimination.

Among the previous work, [5] used phoneme-specific Hidden Markov Models (HMMs) for modeling the target speakers. [6] and [7] used a speaker verification system based on broad phonetic categories and achieved an improvement over the baseline system. [8] compared GMM and HMM system across different phonemes. In [8], GMM and HMM were compared, and unlike in the above cited works, phonetic information was used only during the scoring phase. [9] introduced a phonetic class-based GMM system based on a tree-like structure, which

⋆ Supported by the Swiss National Fund for Scientific Research, No. 2100-067043.01/1

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 580–587, 2005.
© Springer-Verlag Berlin Heidelberg 2005

outperformed a single GMM modeling. Closer to what is presented in [9], are the works done by [10] and [11], where phoneme-adapted GMMs were built for each speaker. [10] and [11] concluded that the phoneme-adapted GMM system outperformed the phoneme independent GMM system. In order to apply such techniques, a phone recognizer is needed. Building the phone recognizer, requires transcribed databases. Transcribing databases is an error-prone and expensive task. To avoid this problem, we propose to use Automatic Language Independent Speech Processing (ALISP) tools [12]. The segmentation can be obtained automatically on speech data without any transcriptions.

In [13] and [14] we have already used the ALISP data-driven speech segmentation method for speaker verification, clustering the speech data in 8 classes. Classifying speech in only 8 speech classes, did not lead to a good coherence of the speech classes. In [15], we have used a finer segmentation of the speech data into 64 speech classes, and a Dynamic Time Warping (DTW) distortion measure for the distance between two speech patterns. In the present work we present an ALISP-based GMM system in which 64 models are built for each speaker. This paper focuses on building and testing ALISP-specific GMMs and secondly how the independent ALISP units scores can be combined to achieve a better performance compared to modeling all speech classes in a single model.

Even though it is possible to use structural Gaussian Mixture Models [16, 17] to perform better and efficient (fast) speaker verification systems, the idea of using data-driven speech segmentation is to be able to further exploit high level informations [18, 19] with no need of transcribed databases.

The outline of this paper is the following: In Section 2 more details about the proposed method are given. Section 3 describes the database used and the experimental protocol. The evaluation results are reported in Section4. The conclusions and perspectives are given in Section 5.

2 Systems Description

2.1 Baseline GMM System

The baseline system is based on Gaussian Mixture Models [20] in which the multivariate distribution distribution of the feature vectors is modeled with a weighted distributions. Two gender-dependent background models are created and each speaker model is obtained by adaptation of the matching gender background model.

For each frame y_t in the test segment a score is calculated using the log-likelihood ratio of the speaker likelihood to the background likelihood

$$s_{y_t} = log\, p(y_t|X) - log\, p(y_t|\overline{X}) \tag{1}$$

where X and $\overline{X}$ denote the client and the world models respectively. The final score Λ^j is obtained by summing the frames' scores and normalizing by T; the total number of frames in the test utterance j:

$$\Lambda^j = \frac{1}{T}\sum_{t=1}^{T} s_{y_t} \tag{2}$$

2.2 ALISP-Based GMM System

This approach aims to model the different speech sounds separately by using a data-driven speech segmentation. The goal is to further enhance the performance of the system by exploiting the speaker discriminating properties of individual speech classes. This system is composed of the following four stages.

First the speech data is segmented using a data-driven segmentation Automatic Language Independent Speech Processing (ALISP) tools [12]. This technique is based on units acquired during a data-driven segmentation, where no phonetic transcription of the corpus is needed. In this work we use 64 classes. The modeling of the set of data-driven speech units, denoted as ALISP units, is achieved through the following stages. After the pre-processing step for the speech data, first Temporal Decomposition is used, followed by Vector Quantization providing a symbolic transcription of the data in an unsupervised manner. Hidden Markov Modeling is further applied for a better coherence of the initial ALISP units.

Secondly and after the segmentation of the speaker and non-speaker speech data, the ALISP-specific background models and speaker models are built using the feature vectors for the given ALISP class. In this segmental approach we represent each speaker by 64 GMMs each of them models an ALISP class. The speaker specific 64 models were adapted from the 64 gender and ALISP class dependent background models.

During the test phase, each test speech data is first segmented with the 64 ALISP HMM models. Then, each ALISP segment found in the test utterance is compared to the hypothesized speaker model and to the background model of the corresponding ALISP class. The segmental scores are calculated using the log-likelihood ratio of the speaker likelihood to the background likelihood.

Finally, and after the computation of a score for each ALISP segment, the segmental scores are combined together to form a single recognition score for the test utterance. In this work linear summation and Multi-Layer Perceptrons (MLP) [21] are used to combine the individual scores for the ALISP segments.

3 Experimental Setup

All experiments are done on the NIST'2004 data which is split into two different subsets: the *Development-set* and the *Evaluation-set*, used to test the performance of the proposed system. We use the "8sides-1side" NIST'2004 protocol in which we dispose of 40 minutes of data to build the speaker model and 5 minutes for the test data (including silences).

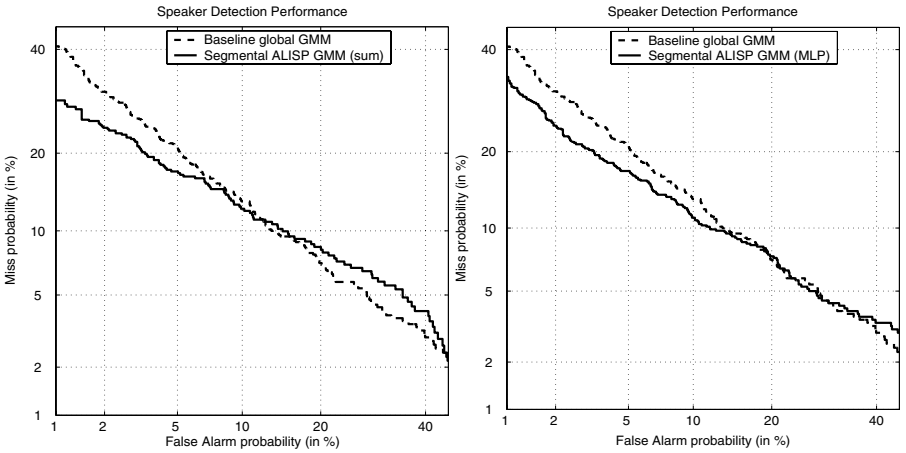
The speech parameterization is done with Mel Frequency Cepstral Coefficients (MFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame a 15-element cepstral vector is computed and appended with first order deltas.

Cepstral mean substraction is applied to the 15 static coefficients and only bands in the 300-3400 Hz frequency range are used. The energy and delta-energy are used in addition during the ALISP units recognition.

During the preprocessing step, after the speech parametrization, we separated the speech from the non-speech data. The speech activity detector is based on a bi-Gaussian modeling of the energy of the speech data [22]. Only frames higher than a certain threshold are chosen for further processing. Using this method, 56% of the original NIST 2004 data are removed.

In the baseline GMM[1] system two gender-dependent background models are built and for each target speaker, a specific GMM with diagonal covariance matrices is trained via maximum a posteriori (MAP) adaptation of the Gaussian means of the matching gender background model. The two gender-dependent background models (with 512 Gaussians) are trained using 5 iterations of the Expectation Maximization (EM) algorithm.

In the ALISP-based GMM system, the ALISP-specific models are trained using the same configuration as the global GMM system except the number of mixtures. In the segmental GMM system, each ALISP unit model had only 32 mixtures, because there was not enough data to keep the mixture count at 512. If an ALISP class does not occur in the training data for a target, the background model of this class becomes that target's model.



(a) (b)

Fig. 1. Speaker verification results for the global GMM system and the segmental system using (a) a linear summation (b) an MLP for the segmental scores fusion on the evaluation data set (subset of NIST'04)

The gender dependent background models for the GMMs and the gender dependent ALISP recognizers, are trained on a total of about 6 hours of data from (1999 and 2001) NIST data sets.

[1] Based on the BECARS package [23]

The MLP is trained on the development set. Since the ALISP units do not always occur in the test data, not all of the ALISP units scores were available for each speaker. For training the MLP, the missing scores were replaced by zero.

4 Experimental Results

We present in this section results for "8sides-1side" NIST 2004 task on the evaluation data set. For this task we dispose of 40 minutes to build the speaker model and 5 minutes for the test data (including silences). Performance is reported in term of the Detection Error Tradeoff (DET) curve [24]. Results are compared via Equal Error Rates (EER): the error at the threshold which gives equal miss and false alarm probabilities.

Figure 1 (a) shows the speaker verification results for the global and the segmental ALISP-based GMM systems. The best performing system in the figure 1 (a) is a linear summation of the segmental scores from the segmental system. The EER was reduced from 12.4% to 11.7% and improvement in the region favoring false alarms is also visible.

The individual performances of the ALISP classes are given in Figure 2. This shows that certain ALISP classes perform better than others. Hence, we can say that a linear summation of the segmental scores does not lead to an optimal solution. Therefore an MLP was applied to the merging of the ALISP segmental scores in order to improve the performance of the segmental system. Figure 1 (b) shows that using an MLP instead of the simple summation brings 17% of improvement in performance, in term of Equal Error Rates, over the baseline system.

As interesting question is, whether all ALISP segments are useful for speaker modeling or whether it is better to ignore some of them when doing speaker

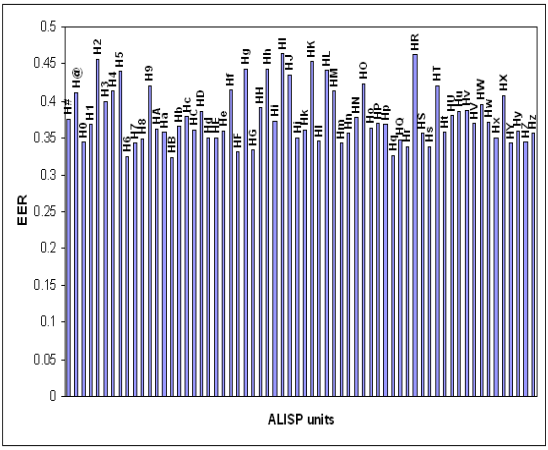


Fig. 2. Individual performances for each ALISP class for male in the development data set (subset of NIST'04)

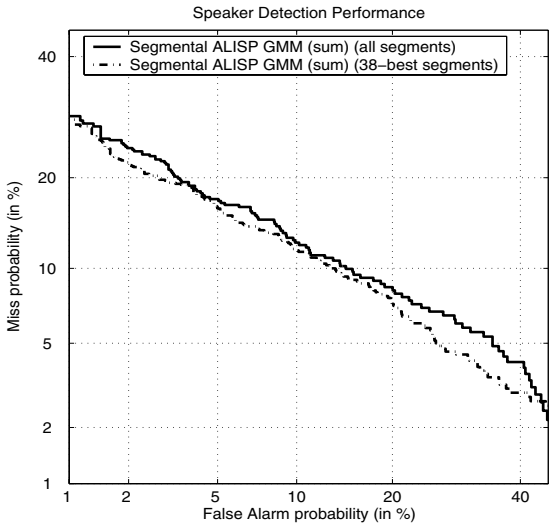


Fig. 3. Speaker verification results for the segmental ALISP-based GMM system using the sum rule for the fusion of all ALISP classes and of only the 38-best classes on the evaluation data set (subset of NIST'04)

verification. Figure 3 shows that using a subset of the most discriminative ALISP units (the top 38) gives better performance than using all of the ALISP classes. Here the ALISP units are sorted by their individual performances and the score of the 38 best ALISP units are combined using a linear summation to produce the final score. These 38 ALISP classes account for about 56% of the total segments in the corpus. Since correct transcriptions of the evaluation data are not available, we cannot compare the correspondence of ALISP units and the usual phonetic units. Figure 4 summarizes the results of the global and the segmental ALISP-based systems.

5 Conclusions and Perspectives

In this paper we have presented a speaker verification system based on data-driven (ALISP) segmentation, instead of the phonetic segmentation. We demonstrated that the ALISP segments could capture speaker information. Thirty eight ALISP units were in fact contributing most to the discrimination between speakers. The segmental ALISP GMM system provided better performance compared to the global GMM system. We have shown that applying both linear weighting to the ALISP units and non linear weighting by using an MLP gave an improvement in performance over the baseline system which models all speech sound with a single model.

This system could be improved by varying the mixture sizes across the ALISP classes and by normalizing the segmental scores using the Tnorm and Znorm techniques. HMMs could also be used instead of the GMMs in order to capture

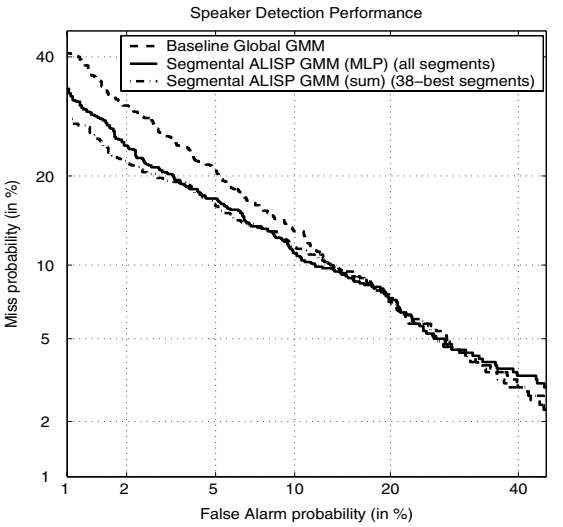


Fig. 4. Speaker verification results for the global GMM system, the segmental system using an MLP for combining scores of all ALISP classes and the segmental system using a linear summation of the 38 best ALISP units on the evaluation data set (subset of NIST'04)

and take advantage of sequential information. We are also investigating the fusion of this system with others systems treating high level information provided by the ALISP sequence and which is related to the speaker style. The great advantage of the proposed method is that it is not grounded on the usage of transcribed speech data.

References

1. Parris, E.S., Carey, M.J.: Discriminative phonemes for speaker identification. In ICLSP (1994) 1843–1846
2. Eatock, J., Mason, J.: A quantitative assessment of the relative speaker discriminant properties of phonemes. Proc. ICASSP **1** (1994) 133–136
3. Olsen, J.: A two-stage procedure for phone based speaker verification. In G. Borgefors J. Bigün, G. Chollet, editor, First International Conference on Audio and Video Based Biometric Person Authentication (1997) 199–226
4. Petrovska-Delacretaz, D., Hennebert, J.: Text-prompted speaker verification experiments with phoneme specific MLP's. In Proc. ICASSP (1998) 777–780
5. Mastui, T., Furui, S.: Concatenated phoneme models for text-variable speaker recognition. Proc. ICASSP (1994) 133–136
6. Koolwaaij, J., de Veth, J.: The use of broad phonetic class models in speaker recognition. Proc. ICSLP (1998)
7. Kajarekar, S.S., Hermanskey, H.: Speaker verification based on broad phonetic categories. 2001: A Speaker Odyssey - The Speaker Recognition Workshop (2001)
8. Auckenthaler, R., Parris, E.S., Carey, M.J.: Improving a GMM speaker verification system by phonetic weighting. Proc. ICASSP (1999)

9. Hébert, M., Heck, L.P.: Phonetic class-based speaker verification. Proc. Eurospeech (2003)
10. Hansen, E.G., Slyh, R.E., Anderson, T.R.: Speaker recognition using phoneme-specific GMMs. Proc. Odyssey (2004)
11. Gutman, D., Bistritz, Y.: Speaker verification using phoneme-adapted gaussian mixture models. Proc. EUSIPCO (2002)
12. Chollet, G., Černocký, J., Constantinescu, A., Deligne, S., Bimbot, F.: Towards ALISP: a proposal for Automatic Language Independent Speech Processing. In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag (1999)
13. Petrovska-Delacrétaz, D., Černocký, J., Hennebert, J., Chollet, G.: Text-independent speaker verification using automatically labeled acoustic segments. In ICLSP (1998)
14. Petrovska-Delacrétaz, D., Černocký, J., Chollet, G.: Segmental approaches for automatic speaker verification. DSP, Special Issue on the NIST'99 evaluations **vol. 10(1-3)** (2000) 198–212
15. Petrovska-Delacretaz, D., El-Hannani, A., Chollet, G.: Searching through a speech memory for text-independent speaker verification. In proc. of Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA) (2003)
16. Xiang, B., Berger, T.: Efficient text-independent speaker verification with structural gaussian mixture models and neural network. IEEE Transactions on Speech and Audio Processing. **vol. 11, no. 5** (2003)
17. Chaudhari, U.V., Navrátil, J., Maes, S.H.: Multi-grained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition. IEEE Transactions on Speech and Audio Processing. **vol. 11, no. 1** (2003)
18. Reynolds, D., Andrews, W., Campbell, J., Navratil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, J., Xiang, B.: The supersid project: Exploiting high-level information for high-accuracy speaker recognition. In Proc. ICASSP (2003)
19. Andrews, W., Kohler, M., Campbell, J., Godfrey, J.: Phonetic, idiolectal, and acoustic speaker recognition. Speaker Odyssey Workshop (2001)
20. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. DSP, Special Issue on the NIST'99 evaluations **vol. 10(1-3)** (2000) 19–41
21. Haykin, S.: Neural Networks: A Comprehensive Foundation. IEEE Computer society Press (1994)
22. Magrin-Chagnolleau, I., Gravier, G., Blouet, R.: Overview of the 2000-2001 elisa consortium research activities. Speaker Odyssey Workshop (2001)
23. Blouet, R., Mokbel, C., Mokbel, H., Sanchez, E., Chollet, G., Greige, H.: Becars: A free software for speaker verification. Proc. Odyssey (2004)
24. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The det curve in assessment of detection task performance. Proc. Eurospeech'97 **vol. 4** (1997) 1895–1898

Multimodal Speaker Verification Using Ear Image Features Extracted by PCA and ICA

Koji Iwano, Taro Miyazaki, and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{iwano,taro,furui}@furui.cs.titech.ac.jp
http://www.furui.cs.titech.ac.jp

Abstract. This paper first compares performances of two authentication methods using ear images, in which feature vectors are extracted by either principal component analysis (PCA) or independent component analysis (ICA). Next, the effectiveness of combining PCA- and ICA-based ear authentication methods is investigated. In our previous work, we proposed an audio-visual person authentication using speech and ear images with the aim of increasing noise robustness in mobile environments. In this paper, we apply the best ear authentication method to our audio-visual authentication method and examine its robustness. Experiments were conducted using an audio-visual database collected from 36 male speakers in five sessions over a half year. Speech data were contaminated with white noise at various SNR conditions. Experimental results show that: (1) PCA outperforms ICA in the ear authentication framework using GMMs; (2) the fusion of PCA- and ICA-based ear authentication is effective; and (3) by combining the fusion method for ear images with the speech-based method, person authentication performance can be improved. The audio-visual person authentication method achieves better performance than ear-based as well as speech-based methods in an SNR range between 15 and 30dB.

1 Introduction

In the IT/network society of today, accurate and convenient person authentication has become increasingly necessary. In order to achieve a high performance, various multimodal biometric authentication methods have been proposed[1–7].

Speech is one of the most useful and effective biometrics for authentication in mobile/ubiquitous environments. However, since its performance deteriorates due to additive noise and session-to-session variability of voice quality, combination with other biometric features is needed for improving the performance.

Along this line, various audio-visual biometric authentication methods have been proposed[1–4]. Although most of them use "face" information in combination with speech, the face features also change due to make-up, mustache, beard, hair styles and so on, and derives degradation of the performance. Therefore, it is worth investigating other biometric features with high permanence.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 588–596, 2005.
© Springer-Verlag Berlin Heidelberg 2005

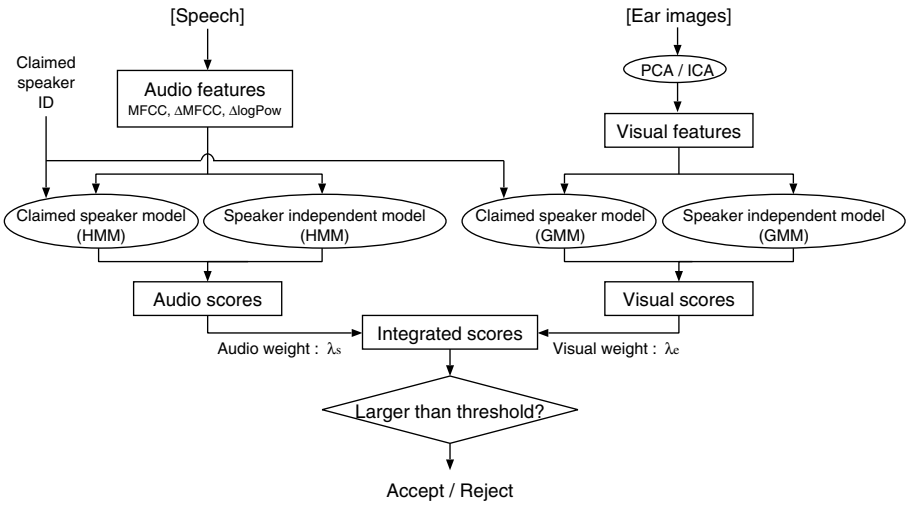


Fig. 1. Multimodal person authentication system using speech and ear images

Since ear shapes are unique to each individual and hardly change over time[8, 9], utilizing ear information is expected to achieve reliable and robust person authentication. [6] reported that combination of ear and face information is effective in biometric authentication. Since ear images can be captured using a small camera installed in a mobile phone, ear information can be more easily used in mobile environments than other biometrics, such as fingerprints, irises, or retinas, which require special equipment. From this point of view, we previously proposed an audio-visual speaker verification method using ear shape information in combination with speech signal[7].

In our previous method[7], ear features were built by an "eigen-ear" approach based on principal component analysis (PCA) in the same way as eigen-face approach is used in face recognition[10], and improved the performance of authentication obtained using only speech. On the other hand, it has been reported that independent component analysis (ICA)[11] can outperform PCA in face recognition[12] and palmprint recognition[13]. This paper applies the ICA to extraction of ear features in our multimodal authentication method, and compares the performances of ICA- and PCA-based authentication methods. Additionally, the two ear authentication methods are combined, and the effectiveness is investigated.

Our authentication method and audio-visual database are described in Section 2. Section 3 reports experimental results and Section 4 concludes this paper.

2 System Structure and Database

Figure 1 shows our multimodal person authentication system using speech and ear images. Audio and visual data are respectively converted into feature vectors. Each set of features is compared to both a claimed person model and a speaker

independent (SI) model to calculate a posterior probability. The audio and visual scores are integrated after appropriate weighting, and a decision is made whether he/she should be accepted as a true speaker or rejected as an impostor, by comparing the integrated score with a preset threshold value.

2.1 Integrated Score

The posterior probability of being a claimed speaker S^c after observing a biometric feature set x, is denoted by $p(S^c|x)$. Since x is composed of speech (audio) features x_s and ear (visual) features x_e, $p(S^c|x)$ can be transformed as follows:

$$p(S^c|x) = p(S^c_s|x_s) \cdot p(S^c_e|x_e) \tag{1}$$

where S^c_s and S^c_e represent the claimed speaker's speech and ear models, respectively. The Bayes' Rule derives the following equation:

$$p(S^c|x) = \frac{p(x_s|S^c_s)p(S^c_s)}{p(x_s)} \cdot \frac{p(x_e|S^c_e)p(S^c_e)}{p(x_e)} \tag{2}$$

where $p(x_s|S^c_s)$ and $p(x_e|S^c_e)$ are likelihood values with claimed speaker's speech and ear models, respectively. The probabilities in the denominator are approximated by using likelihood values for the general speaker's speech model $p(x_s|S^g_s)$ and ear model $p(x_e|S^g_e)$:

$$p(S^c|x) \approx \frac{p(x_s|S^c_s)p(S^c_s)}{p(x_s|S^g_s)p(S^g_s)} \cdot \frac{p(x_e|S^c_e)p(S^c_e)}{p(x_e|S^g_e)p(S^g_e)} \tag{3}$$

$$\propto \frac{p(x_s|S^c_s)}{p(x_s|S^g_s)} \cdot \frac{p(x_e|S^c_e)}{p(x_e|S^g_e)} \tag{4}$$

Equation (4) is derived based on the assumption that observations of all the claimed speakers are equally probable. Since an SI model made by many speakers can be used as the general model, a posterior probability for each claimed speaker's model is calculated by the product of likelihood values normalized by the SI models. By defining authentication scores for speech (p_s) and ear (p_e) as

$$p_m = \log p(x_m|S^c_m) - \log p(x_m|S^g_m) \quad (m = s, e) \tag{5}$$

an integrated score p_{se} which balances the effectiveness of speech and ear features can be modeled by the following equation:

$$p_{se} = \lambda_s p_s + \lambda_e p_e \quad (\lambda_s + \lambda_e = 1) \tag{6}$$

where λ_s and λ_e are audio and visual weights, respectively.

2.2 Audio-Visual Database

Recording Conditions. Audio-visual data[7] were recorded in five sessions at intervals of approximately one month. In this paper, data from 36 male speakers

were used, in which each speaker uttered 50 strings of four connected digits in Japanese at each session. Speech data were sampled at 16kHz with 16bit resolution. One right ear image for each speaker, with no obscuring hair, was taken by a digital camera with 720×540 pixel resolution at each session. The camera was positioned perpendicular to the ear with the distance of approximately 20cm to take the whole ear image.

Training and Testing Data. The database was divided into three groups of 12 speakers. The three groups were assigned to training, testing, and development sets. The development set was used for adjusting parameters such as weights and thresholds. By rotating the roles of the three groups, six experiments were conducted. The set of data recorded at sessions 1∼3 was used for training; and that recorded at sessions 4 and 5 was used for testing and development. The SI model was trained using the utterances by 12 speakers in the training set. All the speakers in the testing data other than the claimed speaker himself were used as impostors.

White noise was added to the audio data for training at a 30dB SNR level to increase the robustness against noisy speech, and testing data were contaminated with white noise at 5, 10, 15, 20, and 30dB SNR conditions.

As image data, we first extracted gray-scaled ear images with 80×80 pixel resolution. The ear location and rotation in the image were manually adjusted. In order to increase the robustness of visual models, the following variations were given to training data:

1. Shifting the ear location in vertical and horizontal directions within ±6 pixels at a 2 pixel interval. Consequently, 49 variations were made for each ear image.
2. Rotating the ear images within ±30 degrees at one degree intervals. Accordingly, 61 variations were made for each ear image.

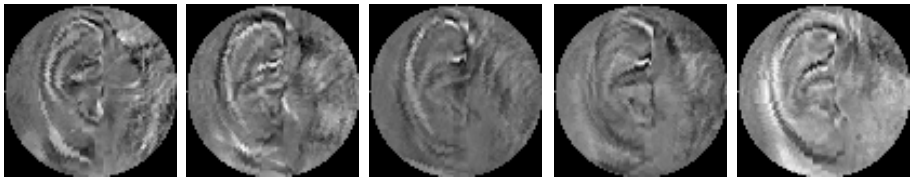
Both operations together made approximately 9,000 (= 3 sessions × 49 × 61) ear images for training each speaker's model. For testing and development data, we applied only the rotating operation.

All ear images were filtered by Laplacian-Gaussian filtering, and then circularly sampled and digitized for reducing hair effects as well as avoiding window shape effects caused by rotation of the images.

2.3 Audio and Visual Features

Audio features were 25-dimensional vectors consisting of 12 MFCCs, 12 ΔMFCCs, and Δ log energy. The frame shift and the analysis window length were 10ms and 25ms, respectively.

For ear images, subspaces were built by using principal component analysis (PCA) and independent component analysis (ICA). Although there are a number of algorithms for performing ICA, the FastICA algorithm[14] was chosen, since it provides the best results for face authentication[12]. Before applying the ICA



(a) PCA



(b) ICA

Fig. 2. Examples of the five basis vectors obtained by PCA and ICA

algorithm to the data, centering and whitening processes were conducted for reducing the computational complexity of the ICA. PCA and ICA were applied to the 12 ear images in the training set recorded at the first session. The original ear images with no shifting or rotation were used for the analysis. Figure 2 shows examples of the five basis vectors obtained by PCA and ICA.

Effective basis vectors were selected from 12 vectors by the greedy algorithm to minimize the equal error rate (EER) for the development set. Then, ear images were converted into ear features by projecting the images on a subspace built by the selected basis vectors. Consequentially, the average number of dimensions for ear features extracted by PCA and ICA were 8.5 and 9.0, respectively.

2.4 Speech and Ear Models

Audio features were modeled by digit-unit HMMs. Each digit HMM has standard left-to-right topology with $n \times 3$ states, where n is the number of phonemes in the digit. The authentication score for the speech features represented in Eq. (4) is calculated as follows:

$$\begin{aligned} \frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} &= \frac{\Sigma_w p(x_s|S_s^c, w)p(w)}{\Sigma_w p(x_s|S_s^g, w)p(w)} \\ &\approx \frac{\max_w p(x_s|S_s^c, w)}{\max_w p(x_s|S_s^g, w)} \end{aligned} \tag{7}$$

where w is a string of four connected digits.

Visual features were modeled using GMMs. In each testing experiment, 61-feature vectors converted from the rotated images were input to the GMMs. Log likelihood values calculated for the claimed speaker and the SI models were used to obtain the authentication score for each ear image according to the Eq. (5).

Table 1. Error rates (%) in person authentication using ear images extracted by PCA and ICA for development and test sets. "PCA+ICA" indicates error rates of a fusion method integrating visual scores obtained from the PCA- and ICA-based methods

	dev. set	test set
PCA	6.5	11.3
ICA	12.4	14.2
PCA+ICA	6.3	9.8

3 Experimental Results

3.1 Authentication Using Ear Images

First, an experiment using only ear images was conducted. Table 1 shows error rates by PCA- and ICA-based authentication for the development and test sets. The number of mixtures for speaker GMMs and SI GMM, the number of dimension of feature vectors, and the threshold for authentication were optimized to minimize the EER for the development set. "PCA+ICA" indicates error rates by the fusion method; that is, PCA- and ICA-based visual scores were integrated with appropriate weights, λ_{PCA} and λ_{ICA} ($\lambda_{PCA} + \lambda_{ICA} = 1$), and used for the authentication. λ_{PCA} and λ_{ICA} were also optimized for the development set.

The table shows that PCA outperforms ICA in person authentication using ear images. The best performance is obtained by using the fusion technique; 6.3% error rate for the closed condition and 9.8% error rate for the open condition were obtained.

3.2 Multimodal Authentication Using Speech and Ears

Audio-visual authentication results for the test set in various SNR conditions are shown in Fig. 3. Since the fusion method "PCA+ICA" showed the best performance for authentication using ear images, it was used in this experiment. The audio and visual weights (λ_s and λ_e) and the threshold were optimized to minimize the EER for the development set at each condition. The number of mixtures in the audio HMMs was determined based on experimental results using only speech at 30dB SNR condition. Results using only speech ($\lambda_s = 1.0$) and only ear ($\lambda_s = 0.0$) are also shown for the purpose of comparison.

It is clearly shown that multimodal authentication is more robust than speech-only methods where the performance significantly degrades with noise. The multimodal method outperforms both speech- and ear-only methods when the SNR is 30, 20, and 15dB. The combination of speech and ear images is most effective when the SNR is 30dB; an error rate of 0.26% is obtained which is 67.9% lower than the speech-only method and 97.3% lower than the visual-only method.

Figure 4 shows examples of DET curves with various authentication methods at 15dB SNR. In this experiment, all the control parameters except thresholds were determined to minimize the EER for the development set. Effectiveness of combining speech and ear information is also shown in this figure.

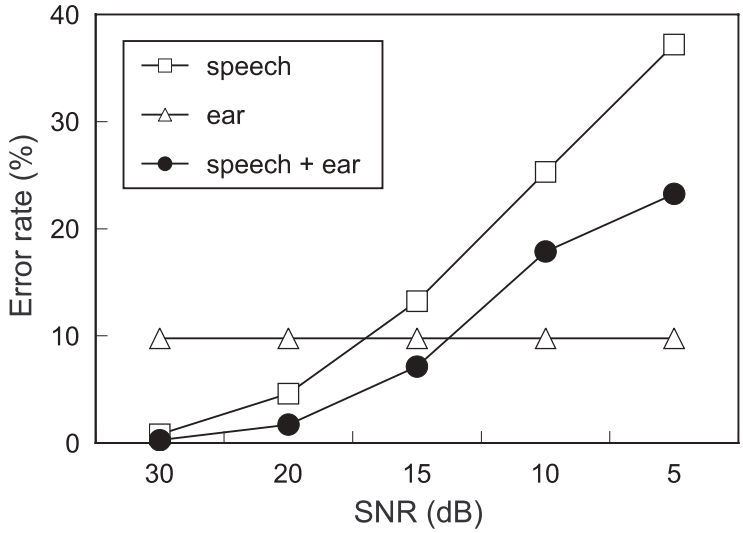


Fig. 3. Results of multimodal person authentication for test set in various SNR conditions

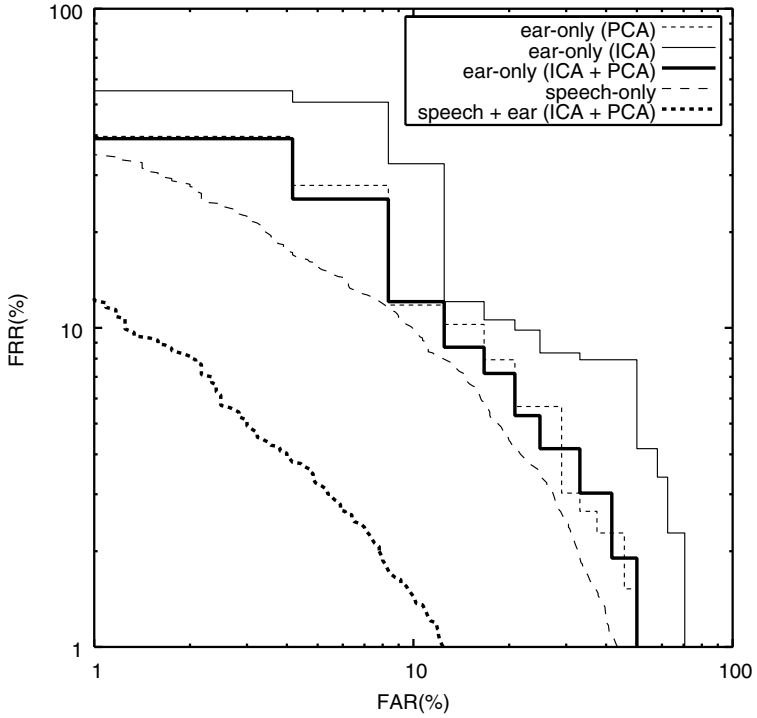


Fig. 4. DET curves of various person authentication methods at 15dB SNR condition

4 Conclusions

This paper has applied an ICA-based feature extraction technique to our multimodal authentication method using speech and ear images.

First, an ICA-based ear authentication method was compared with a PCA-based method. Experimental results show that PCA outperforms ICA in the framework using GMMs. Since the performance of the ICA method varies largely according to task, data size, and details of the algorithm, we need to conduct further experiments before generalizing the conclusion. Our future research includes applying an alternative ICA algorithm rather than the FastICA and increasing the ear image data for evaluation.

It was confirmed that the fusion of PCA- and ICA-based ear authentication methods is effective, which indicates that the two authentication methods are fairly independent.

It was also confirmed that, by using the fusion method for ear authentication and combining it with speech information, our multimodal method becomes more robust than the speech-only method in various SNR conditions.

Our future work also includes improving authentication performance using ear information by increasing the robustness against ear image variation caused by tilt of camera and developing an automatic ear-area detection method.

References

1. Brunelli, R., Falavigna, D.: Personal identification using multiple cues, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.17, no.10 (1995) 955–966.
2. Duc, B., Bigun, E.S., Bigun, J., Maitre, G., and Fischer, S.: Fusion of audio and video information for multi modal person authentication, Pattern Recognition Letters, vol.18, no.9 (1997) 835–843.
3. Jourlin, P., Luettin, J., Genoud, D., and Wassner, H.: Acoustic-labial speaker verification, Pattern Recognition Letters, vol.18, no.9 (1997) 853–858.
4. Poh, N., Korczak, J.: Hybrid biometric person authentication using face and voice features, In: Audio- and Video-Based Biometric Person Authentication, Third International Conference, AVBPA 2001, Bigun, J. and Smeraldi, F., Eds., Springer (2001) 348–353.
5. Ross, A., Jain, A., and Qian, J.-Z.: Information fusion in biometrics, In: Audio- and Video-Based Biometric Person Authentication, Third International Conference, AVBPA 2001, Bigun, J. and Smeraldi, F., Eds., Springer (2001) 354–359,
6. Chang, K., Bowyer, K.W., Sarkar, S., and Victor, B.: Comparison and combination of ear and face images in appearance-based biometrics, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.25, no.9 (2003) 1160–1165.
7. Iwano, K., Hirose, T., Kamibayashi, E., and Furui, S.: Audio-visual person authentication using speech and ear images, Proc. Workshop on Multimodal User Authentication (2003) 85–90.
8. Iannarelli, A.: Ear Identification, Forensic Identification series. Paramont Publishing Company, Fremont, California (1989).
9. Burge, M., Burger, W.: Ear biometrics, In: Biometrics: Personal Identification in Networked Society, Jain, A., Bolle, R., and Pankanti, S., Eds., Kluwer Academic, Boston, MA (1999) 273–285.

10. Turk, M., Pentland, A.P.: Face recognition using eigenface, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1991) 586–591.
11. Common, P.: Independent component analysis – A new concept?, Signal Processing, vol.36, no.3 (1994) 287–314.
12. Draper, B.A., Baek, K., Bartlett, M.S., and Beveridge, J.R.: Recognizing faces with PCA and ICA, Computer Vision and Image Understanding, vol.91, iss.1–2 (2003) 115–137.
13. Connie, T., Teoh, A., Goh, M., and Ngo, D.: Palmprint recognition with PCA and ICA, Proc. Image and Vision Computing New Zealand, Palmerston North, New Zealand (2003) 227–232.
14. http://www.cis.hut.fi/projects/ica/fastica/

Modelling the Time-Variant Covariates for Gait Recognition

Galina V. Veres, Mark S. Nixon, and John N. Carter

Department of Electronics and Computer Science
University of Southampton, Southampton, SO17 1BJ

Abstract. This paper deals with a problem of recognition by gait when time-dependent covariates are added, i.e. when 6 months have passed between recording of the gallery and the probe sets. We show how recognition rates fall significantly when data is captured between lengthy time intevals, for static and dynamic gait features. Under the assumption that it is possible to have some subjects from the probe for training and that similar subjects have similar changes in gait over time, a predictive model of changes in gait is suggested in this paper, which can improve the recognition capability. A small number of subjects were used for training and a much large number for classification and the probe contains the covariate data for a smaller number of subjects. Our new predictive model derives high recognition rates for different features which is a considerable improvement on recognition capability without this new approach.

1 Introduction

Recently much attention has been devoted to use of human gait patterns as biometric. Gait recognition aims to discriminate individuals by the way they walk and has the advantage of being non-invasive, hard to conceal, being readily captured without a walker's attention and is less likely to be obscured than other biometric features. Approaches to gait recognition can be broadly classified as being model-based and model-free. Model-based methods [3, 5, 10] model the human body structure and extract image features to map them into structural components of models or to derive motion trajectories of body parts. Model-free methods [1, 2, 4, 23] generally characterise the whole motion pattern of the human body by a compact representation regardless of the underlying structure. In this paper we employ the model-based (dynamic) method of Wagg and Nixon [22] and the model-free (static) method of Veres et al [20].

However, in these works only databases recorded over a short time interval were evaluated. Some studies over a more lengthy time interval were reported for face recognition. In [19] images of 240 distinct subjects were acquired under controlled conditions, over a period of ten weeks. They showed that there was not a clearly decreasing performance trend over a period of ten weeks and concluded that reduction in degradation is small enough as to be nearly flat over this time period. Other studies have shown that over a period of years, face recognition performance degrades linearly with time [16]. Some studies were done to

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 597–606, 2005.
© Springer-Verlag Berlin Heidelberg 2005

show effects of ageing on face recognition [6–8]. In [6] a systematic method for modelling appearance variation due to ageing is presented. It was shown that ageing variation is specific to a given individual, it occurs slowly and it is affected significantly by other factors, such as the health, gender and the lifestyle of the individual. Taking this into consideration, reasonably accurate estimates of age can be made for unseen images. In [7, 8] face identification experiments are presented, where the age of individuals in the gallery is significantly different than the age of individuals in the probe. It was demonstrated that automatic age simulation techniques can be used for designing face recognition systems, robust to ageing variation. In this context, the perceived age of subjects in the gallery and probe is modified before the training and classification procedures, so that ageing variation is eliminated. O'Toole et al. [14, 15] use three-dimensional facial information for building a parametric $3D$ face model. They use a caricature algorithm in order to exaggerate or deemphasize distinctive $3D$ facial features; in the resulting caricatures, the perceived age is increased or decreased according to the exaggeration level, suggesting that $3D$ distinctive facial features are emphasized in older faces. Some recent efforts [9] were made to improve age estimation by devoting part of the classification procedure to choosing the most appropriate classifier for the subject/age range in question, so that more accurate age estimates can be obtained.

In this paper we consider a gait recognition problem when two databases (the gallery and probe) were recorded with a time interval of 6 months between the finish of recording the first database (gallery) and the start of recording the second database (probe), i.e. time-dependent covariates are added. Moreover, some extra covariates were added in the second database such as different shoes, clothes, carrying different bags. In real life the need to analyse such databases arises in security of access to a company or an embassy for example. It is possible to record people walking normally as a gallery, but later it will be necessary to recognize these people in different clothes, shoes, possibly carrying luggage and when time passes. It is shown that in this case correct classification rates fall significantly and recognition becomes unreliable. Similar results were obtained for the HumanID Gait Challenge Problem [17], where recognition fell from 82% to 6% after 6 months. Some other recent works reported a significant fall in recognition capability over a lengthy time interval [11–13]. Under the assumptions that we can have records of people walking normally from the probe and similar people have similar changes in gait, the predictive model of gait changes is suggested in this paper as way to increase CCRs when analysis is needed over time. The predictive model is based on available records both from the gallery and the probe and a prediction matrix is constructed for these subjects. Then prediction matrix is generalised for all subjects in the gallery and predicted gallery is obtained. The probe is analysed via the predicted gallery and CCRs are calculated. To show robustness of the suggested approach the predictive model was appllied both on static and dynamic feature sets. We show that CCRs can be increased by several times when using the new predictive model.

Section 2 describes the suggested prediction model for changes in gait over lengthy time interval. The methodology of constructing feature sets is presented in Section 3. Experimental results are presented and described in Section 4. Section 5 concludes this paper.

2 Prediction of Time-Variant Covariates

The idea of our approach was inspired by work of Lanitis et al [8] where it was shown that reasonably accurate estimates of age can be made for unseen images. In case of recognition by gait over a lengthy time interval we assume that it is possible to predict the gallery over the given time interval and achieve good recognition results by analysing the probe via the predicted gallery. In this case the training set consists of a set of subjects from the gallery and the same set of subjects from the probe. The probe was recorded some time later after finishing recording the gallery. In the general case the predicted gallery can be defined as

$$\hat{\mathbf{G}} = \mathbf{G} + \mathbf{Q}, \tag{1}$$

where $\hat{\mathbf{G}}$ is the predicted gallery, $\mathbf{G}$ is the gallery and $\mathbf{Q}$ is a prediction matrix.

Let the gallery and the probe be divided into groups, where a number of groups corresponds to a number of subjects and each group represents feature vectors for a given subject. Let us consider at first the case when the number of groups (subjects) in the gallery equals the number of groups (subjects) in the probe and the groups (subjects) are the same. At first we sort the records of the subjects according to their groups and note the number of records per subject. Then the prediction matrix $\mathbf{Q}$ is constructed as follows. At first for each group in the probe and gallery the mean of the group is calculated

$$\bar{\mathbf{x}}_p^j = \frac{1}{n_p^j} \sum_{i=1}^{n_p^j} \mathbf{x}_{pi}^j \quad \text{and} \quad \bar{\mathbf{x}}_g^j = \frac{1}{n_g^j} \sum_{i=1}^{n_g^j} \mathbf{x}_{gi}^j, \tag{2}$$

where $\bar{\mathbf{x}}_p^j$ is the mean of group j in the probe, $\bar{\mathbf{x}}_g^j$ is the mean of group j in the gallery, $j = 1, \ldots, n_g$, where n_g is a number of groups, n_p^j and n_g^j is a number of records in the jth group of the probe and of the gallery, respectively, $\mathbf{x}_{pi}^j$ and $\mathbf{x}_{gi}^j$ are records for ith subject in jth group in the probe and gallery respectively.

Then the prediction matrix for each group is calculated as

$$\mathbf{Q}^j = \mathbf{e}(\bar{\mathbf{x}}_p^j - \bar{\mathbf{x}}_g^j), \quad j = 1, \ldots, n_g, \tag{3}$$

where $\mathbf{e}$ is a positive unit vector of $(n_g^j \times 1)$, and the final prediction matrix is

$$\mathbf{Q} = [\mathbf{Q}^1; \mathbf{Q}^2; \ldots; \mathbf{Q}^{n_g}]. \tag{4}$$

For the more general case when a number of groups in the gallery is not the same as a number of groups in the probe and/or subjects are not the same both in the gallery and in the probe, the prediction matrix is constructed as follows. Here we present a case when a number of groups in the gallery is more than a number of groups in the probe. Two assumptions are made in this case

1. Every subject in the probe exists in the gallery.
2. The gait of the similar subjects will change in a similar manner with time.

We are looking forward to gathering more data to provide a theoretical analysis or statistical observation to support the second assumption.

The gallery and probe are divided into groups. The gallery is rearranged in such a way that the first n_g will be the groups which are in the probe, and the last n_{dg} groups are not in the probe. The number of groups in the gallery is $n_{gg} = n_g + n_{dg}$. Then the final prediction matrix is calculated as

$$\mathbf{Q} = [\mathbf{Q}^1; \mathbf{Q}^2; \ldots; \mathbf{Q}^{n_g}, \mathbf{Q}^{n_g+1}, \ldots, \mathbf{Q}^{n_{gg}}]. \tag{5}$$

The means of all groups in the probe and the gallery are calculated and prediction matrices for coincidental groups are taken as (3) in this case. The differences will be for groups in the gallery which do not exist in the probe. Further the calculation of the prediction matrices for such groups is presented. To be able to distinguish means of groups belonging only to the gallery from the means of the groups existing in the probe we will denote them as $\bar{\mathbf{x}}_g^{d_j}$, where $d_j = 1, \ldots, n_{dg}$. Then taking into consideration assumption 2, we can first compare the means of groups existing both in the probe and in the gallery with the means of groups existing only in the gallery. For each $\bar{\mathbf{x}}_g^{d_j}$ find the first nearest neighbour from $\bar{\mathbf{x}}_g^j$ by using formula

$$\text{find } k \text{ such as } \; k = j : \{\min_j |\bar{\mathbf{x}}_g^j - \bar{\mathbf{x}}_g^{d_j}|, \;\; j = 1, \ldots, n_g\}. \tag{6}$$

Then the predicted matrix for a given group is

$$\mathbf{Q}^{n_g+d_j} = \mathbf{Q}^k + \mathbf{e}|\bar{\mathbf{x}}_g^k - \bar{\mathbf{x}}_g^{d_j}|, \tag{7}$$

where $\mathbf{e}$ is of $(n_g^k \times 1)$.

After the prediction matrix is obtained, the predicted gallery is calculated as (1) and the probe is classified via the predicted gallery. In some cases it is possible that the number of records per subject is much higher than the number of records used for calculation of prediction matrix. In this case the probe is divided into parts: one used for training and classification and second used only for classification. The suggested approach to predict changes of gait over lengthy time interval was applied to recognition by gait later in the paper.

3 Methodology

Two databases were analysed in the paper, both comprising indoor (studio) data since the purpose of this paper is to investigate a gait recognition problem. The first database, called the large database (LDB), consists of 115 subjects walking normally. The database arrangements are described elsewhere [18]. The LDB can be used to determine which image information remains unchanged for a subject in normal conditions and which changes significantly from subject to subject, i.e.

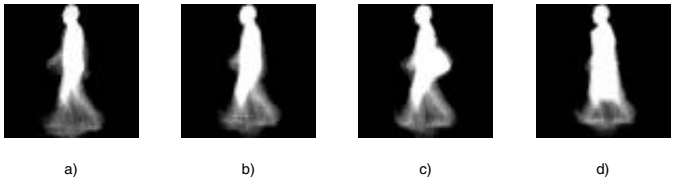


Fig. 1. Average silhouette representing subject 46: a) LDB, b) SDB, walking normally, c) SDB, carrying a bag, d) SDB, wearing trench coat

it represents subject-dependent covariates. The small database (SDB) consists of a sample of 10 subjects from the LDB. Each subject was filmed wearing a variety of footwear, clothes and carrying various bags. They were also filmed walking at different speeds. In this case the covariate factor is more complicated than decribed in Section 2, since it does not only depend on the subjects, but on the other factors mentioned above. Each subject's data was captured during one continuous filming session. Examples of features from the LDB and the SDB is presented in Fig. 1. The figure shows subject 46's data recorded in the LDB and in the SDB walking normally, carrying a bag and wearing a trench coat.

The SDB is intended to investigate the robustness of biometric techniques to imagery of the same subject in various common conditions (carrying items, wearing different clothing or footwear). It worth noticing that sequence of filming data was LDB first and then SDB with approximately 6 months difference between recording LDB and SDB, i.e. time-dependent covariates are also added.

For brevity we shall not fully describe the extraction methods, complete explanations of the techniques used can be found in [20, 22]. These techniques yield two feature vectors for each database: the dynamic vector consists of 73 parameters describing joint rotation models of the limbs together with normalized information about the subject's speed and gait frequency; the static method consists of a 4096 dimensional vector derived from the subject's silhouette accumulated over a sequence of images.

4 Experimental Results

In this section we present the results of an experimental assessment of the performance of the suggested predictive model in the tasks of recognizing people over lengthy time interval by gait, i.e when SDB was analysied via LDB S/L, and of predicting the future changes in gait. The system was tested on five different conditions. For the first we performed leave-one-out experiments using all our training silhouettes. The training set consists of 10 subjects belonging both to LDB and SDB, i.e. S_{10}/L_{10}. This experiment tested the accuracy of our approach in prediction of gait changes over time in subjects who had already provided training silhouettes. In the second experiment we are testing the predictive model for a case when the testing set was 10 subjects from the SDB walking normally but recorded 1 hour later in comparision to the training set, i.e. $S_{10}(1)/L_{10}$. This experiments shows how robust the predictive model is

to extra small time intervals, which were not taking into consideration during training. For the third experiment we added 105 new subjects to the gallery and tested how our approach will work when many more subjects are available for recognition in gallery than were used in training, i.e. S_{10}/L_{all}. In fourth experiment the training set was kept the same, but at the classification stage non time-dependent covariates are added to the probe, i.e. S_{all}/L_{10}. This test showed how adding extra covariates which were not used in training will affect the performances of the predictive model. The last experiment investigates the performance of the predictive model when only 10 subjects from the gallery and the probe are available for training, but 105 subjects are added to the gallery and 12 different experiments affecting a gait of a subject are added to the probe at classification stage, i.e. S_{all}/L_{all}.

To show the robustness of the suggested approach the prediction was applied to both the static *stat* and dynamic *dyn* features of gait. The reduced datasets were used for representing both static and dynamic part of gait, since it was shown [20] that recognition rates are little changed if a subset of the features is used. Different dimensionality reduction techniques were chosen for dynamic and static parts of gait due to their different performance on different datasets. The reduction techniques were applied only to training sets (not on whole databases) and choice of reduction technique was made accordingly, i.e. the best from the point of view of the training set. Static feature set dimensionality reduction was achieved by one-way ANOVA (analysis of variance), and the backwards feature selection algorithm [21] was used for the dynamic feature set. In case of dimensionality reduction by ANOVA the features are selected which satisfy the condition $i : F_i > k_1(\max(F) - min(F))$, where F is the F-statistic, and k_1 is a coefficient. In our case $k_1 = 0.2$ was chosen for the best classification results. And by correct classification rate, CCR is understood a correct classification rate obtained by by comparison SDB (to mean probe) via LDB (gallery) if not mentioned otherwise. The CCR was calculated using Euclidean distance as it is the most popular distance metric in gait recognition literature and the 1-nearest neighbour rule. To show the influence of time and not a degradation of data quality on recognition, the CCR is calculated for each database separately, i.e. the LDB is analysed via LDB (L_{all}/L_{all}) and SDB via SDB (S_{all}/S_{all}). The results are presented in Table 1 and show acceptable CCRs for both dynamic and static feature sets. The CCR's are consistent with the size of the database and we are able to recognise people, that means that both sets of data are good.

Table 1. Analysis of databases without time-dependent covariates

	L_{all}/L_{all}	S_{all}/S_{all}
stat	98.47%	99.90%
dyn	72.32%	90.24%

CCRs for original and reduced datasets when SDB via LDB analysed are presented in Table 2, i.e. when LDB is considered as the gallery and SDB is

Table 2. CCRs for static and dynamic features before training

Dataset	number of features	S_{10}/L_{10}	$S_{10}(1)/L_{10}$	S_{10}/L_{all}	S_{all}/L_{10}	S_{all}/L_{all}
stat	4096	62.10%	64.93%	43.84%	47.61%	22.54%
stat	174	70.32%	68.24%	42.47%	51.73%	20.74%
dyn	73	19.18%	22.27%	8.68%	16.41%	5.30%
dyn	34	28.77%	33.65%	20.09%	28.81%	13.50%

considered as the probe. A subject in the probe can wear normal shoes, clothes and walk normally or can walk slower/faster than normal, wear different shoes, raincoat or even carry a bag/rucksack. We try to match this subject in the probe to a subject in the gallery who walks normally, wears normal shoes and normal clothes (no raincoat) and does not carry any bags. Training is done only for 10 subjects walking normally in both databases. Analysis is done when all features are taken into consideration in the probe and the gallery and when a reduced set of features is considered. In Tables 2 it can be seen that as soon as time-dependent covariates are added to analysis the fall in CCR is very noticeable. In some case the recognition rate approaches chance. The significant reduction in feature space is achieved without significant loss in CCR, and in some cases the visible increase in CCR can be seen. At the same time the CCRs are very low even in the best case especially when all subjects and all experiments are considered and something should be done to improve CCR. One of the ways is to use the suggested predictive model. It was noticing that if for static features the small time interval in experiments S_{10}/L_{10} and $S_{10}(1)/L_{10}$ does not cause noticable change in the recognition cability, in case of dynamic features the difference in CRR is almost 17% whcih can cause the noticable reduction in CCR on the testing set in comparison with the training set. However CCRs are affected not only by time-dependent covariates but adding extra subjects in the gallery and by adding non time-dependent covariates to the probe. In this paper we try to remove time-dependent covariates, but the influence of non time-dependent covariates will remain since it was assumed that only some records/subjects are available for training.

Results of applying the predictive model to the training set and later to the whole databases are presented in Table 3. When only the training set is considered (S_{10}/L_{10}), 99.54% CCR is achieved for static features and 90.41% CCR for dynamic features which is expected as the predictive model was built on these sets and these numbers verify the linear model suggested for prediction in the paper. However, there is a drop in CCRs when experiment $S_{10}(1)/L_{10}$ is considered, i.e. when the testing and the training set recorded with time interval 1 hour. In the case of static features the drop can be considered insignificant and we can say that this experiment is verified the predictive model for static features. In the case of dynamic features it is almost a 10% drop, which can be explained by the lower CCR due to differences between the training set and testing set, see Table 2. However, for both static and dynamic features, CCRs are much higher after training than they were before training for this experiment. Moreover, when more subjects are added to the gallery, the predictive model

can cope with this situation quite well. Only insignificant reduction in CCRs are presented for experiment S_{10}/L_{all} for both the dynamic and static feature sets. Also CCRs are decreasing with adding extra covariates to the training sets (more subjects, different experiments or both), the final results are much better than before training. When all subjects and experiments are analysed the CCR of 65.44% is achieved for static features which is practically three times more than was before training. In case of dynamic features the increase in CCR is almost four times which is very significant.

Table 3. CCR for different datasets after applying predictive model

Dataset	S_{10}/L_{10}	$S_{10}(1)/L_{10}$	S_{10}/L_{all}	S_{all}/L_{10}	S_{all}/L_{all}
stat	99.54%	94.79%	96.35%	76.80%	65.44%
dyn	90.41%	82.46%	88.21%	62.66%	50.24%

Table 3 shows that the suggested prediction approach is able to correct time-dependent variances and even produces good results when the gallery is much bigger than training set. However, it is not able to take into account the non time-dependent covariates in the probe. Therefore some extra efforts are needed to improve CCRs in such cases, which can be different predictive models, different classifiers, or fusion algorithms. We did not use sophisticated classifiers in this paper so we can verify the suggested linear model for the given problem.

5 Conclusions

This paper deals with a problem of increasing correct classification rate when time-dependent covariates (6 months passed between the finish of recording the gallery and the start of recording the probe) together with some other covariates such as variety of footwear, clothes and carrying different bags are added to an analysed database for gait recognition. We have shown that CCRs are very low in this case. In this paper we suggest to use the prediction of gait over the given time interval. One assumption made for the predictive model is that similar subjects will have similar changes in gait over a lengthy interval. The predictive model is based on estimation of differences between the means of subjects in the gallery and the probe and incorporation of these differences in prediction of the feature vectors in the gallery over the given time interval. Then the predicted gallery is compared with the probe and cross-validation is done. The experimental results showed that good results can be achieved both on the training set and the testing set and when extra subjects are added to the gallery and/or extra covariates are added to the probe. The predictive model allows good estimation of gait even when extra subjects are added to the gallery. However, the results are less impressive (though very good in comparison with the results achieved without prediction) when extra covariates are added to the probe such as wearing different shoes/clothes, wearing bags and walking faster/slower. To increase the CCR still further in such cases different approaches can be used: better predictive model, different classifiers or fusion.

Acknowledgments

We gratefully acknowledge partial support by the Defence Technology Centre 8 – 10 supported by General Dynamics.

References

1. A.F. Bobick, and A. Johnson Gait extraction and description by evidence-gathering *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 423-430, 2001
2. R. Collins, R. Gross, and J. Shi Silhouette-based human identification from body shape and gait *Proceedings of the International conference on Automatic Face and Gesture Recognition*, Washington,DC, 2002
3. D. Cunado, M.S. Nixon, and J. N. Carter Automatic extraction and description of human gait models for recognition purposes *Computer Vision and Image Understanding*, **90**(1): 1-41, 2003
4. P.S. Huang, C.J. Harris, and M. S. Nixon Recognizing humans by gait via parametric canonical space *Artificial Intiligence in Engineering*, **13**(4): 359-366, 1999
5. A. Kale, N. Cuntoor, and R. Chellapa A framework for activity-specific human recognition *Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing*, Orlando, Fl,, 2002
6. A. Lanitis, C.J. Taylor and T.F. Cootes Modeling the process of ageing in face images, *Proceedings of the Seventh IEEE International Conference on Computer Vision*, **1**: 131-136 , 1999.
7. A. Lanitis and C.J. Taylor Towards automatic face identification robust to ageing variation, *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 391-396 , 2000.
8. A. Lanitis, C.J. Taylor, and T. F. Cootes Toward Automatic Simulation of Aging Effects on Face Images *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(4): 442-455, 2002
9. A. Lanitis, C. Draganova and C. Christodoulou Comparing different classifiers for automatic age estimation, *IEEE Transactions on Systems, Man, and Cybernetics-PartB: Cybernetics* , **34**(1): 621-628, 2004.
10. L. Lee, and W.E.L. Grimson Gait analysis for recognition and classification *Proceedings of the IEEE International Conference on Face and Gesture*, 155-161, 2002
11. Z. Liu, and S. Sarkar Simplest Representation Yet for Gait Recognition: Averaged Silhouette, *Proceedings of the 17th International Conference on Pattern Recognition* , **2**: 704-711, 2004.
12. Z. Liu, L. Malave and S. Sarkar Studies on Silhouette Quality and Gait Recognition, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* , **4**: 211-214, 2004.
13. Z. Liu, and S. Sarkar Effect of Silhouette Quality on Hard Problems in Gait Recognition, Accepted for future publication *IEEE Transactions on Systems, Man, and Cybernetics- PartB: Cybernetics* , 2005.
14. A.J. O'Toole, T. Vetter, H. Volz and E. Salter Three-dimensional caricatures of human heads: distinctiveness and the perception of age, *Perception*, **26**: 719-732, 1997.
15. A.J. O'Toole, T. Price, T. Vetter, J.C. Bartlett and V. Blanz 3D shape and 2D surface textures of human faces: The 'Role'of 'averages in atractiveness and age' , *Image and Vision Computing*, **18**(1): 9-20, 1999.

16. P. Phillips, D. Blackburn, M. Bone, P. Grother, R. Micheals and E. Tabassi Face recognition vendor test 2002 (FRVT 2002), *Available at http://www.frvt.org/FRVT2002/default.htm*, 2002.
17. S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.V. Bowyer The HumanID Gait Challenge Problem: Data sets, Performances, and Analysis *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**(2): 162-177, 2005.
18. J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter On a large sequence-based human gait database *Proc. 4th International Conf. on Recent Advances in Soft Computing*, Nottingham, UK,66-72, 2002
19. D.A. Socolinsky and A. Selinger. Thermal face recognition over time , *Proceedings of Proceedings of the 17th International Conference on Pattern Recognition*, **4**: 187-190, 2004.
20. G.V. Veres, L. Gordon, J.N. Carter and M.S. Nixon What image information is important in silhouette-based gait recognition? *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA, **II**: 776-782 2004.
21. G.V. Veres, L. Gordon, J.N. Carter, M.S. Nixon. Feature extraction or feature selection in silhouette-based gait recognition? Submitted to *Pattern Recognition*.
22. D.K. Wagg and M.S. Nixon Automated markless extraction of walking people using deformable contour models *Computer Animation amd Virtual Worlds*, **15**(3): 399-406, 2004.
23. L. Wang , W.M. Hu and T.N. Tan A new attempt to gait-based human identification *IEEE Transactions on Circuits and Systems for Video Technology*, **14**(2): 149-158, 2002.

Robust Face Recognition Using Advanced Correlation Filters with Bijective-Mapping Preprocessing

Marios Savvides[1], Chunyan Xie[1], Nancy Chu[1], B.V.K. Vijaya Kumar[1], Christine Podilchuk[2], Ankur Patel[2], Ashwath Harthattu[2], and Richard Mammone[2]

[1] Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh PA 15213
msavvid@ri.cmu.edu, {chunyanx,nlnchu,kumar}@ece.cmu.edu
[2] Electrical and Computer Engineering, Rutgers University, Piscataway NJ 08854
{chrisp1,ankurrp,aithal,mammone}@caip.rutgers.edu

Abstract. In this paper we explore performing robust face verification using Advanced Correlation Filters on Bijective-Mapping preprocessed face images. We show that using the proposed Bijective-Mapping preprocessing method we can increase verification performance (at 0.1% FAR) significantly in our experiments using the Face Recognition Grand Challenge (FRGC) database collected by the University of Notre Dame consisting of 152 subjects. This recognition experiment is challenging as the results reported on these experiments utilize only a single gallery image from each subject during the training phase and the probe images are captured in different time-lapsed sessions which vary mostly in pose and facial expression.

1 Introduction

Face recognition[1] is one of the least intrusive biometrics requiring the least amount of user-cooperation, more importantly the camera surveillance infrastructure is already in place. There are many face recognition applications, one of the more popular applications include the criminal/terrorist watch-list face recognition systems; this is particularly challenging as typically in such situations we may only have a single face image for training. Thus this paper focuses on performing robust face recognition using single gallery images for training. We report recognition performance on the Face Recognition Grand Challenge (FRGC)[2] dataset collected by the University of Notre Dame. We report results on Experiment 1 that is challenging since there are 152 subjects with only a single gallery image person to train from. The probe set consists of image captured from each person up to 8 different time-lapsed sessions. Other variations in probe set include pose and facial expression. We show that we can get improved face verification performance at 0.1% False Acceptance Rate (FAR) using Advanced Correlation Filters(ACF) approach with Bijective-Mapping preprocessed face images. We show that a parts-based representation works best compared to whole-face representation. We also show that using Advanced Correlation Filters clearly outperforms the Matched Filter. In all experimental configurations that use the novel preprocessing method based on Bijective-Mapping[3] and applying Advanced Correlation Filter methods provide the highest recognition accuracy.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 607–616, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Preprocessing

2.1 Introduction

A new preprocessing technique is proposed that is robust to variations in pose, illumination and expression[3]. The framework is similar to the ubiquitous block matching algorithm used for motion estimation in video compression but has been modified to compensate for illumination differences. A mapping between the probe (image to be preprocessed) and the gallery images is obtained using a variation of the block matching algorithm that compensates for illumination differences. Once the mappings are found, the degree of bijectivity that each mapping produces is used to determine which mapping is to be used to preprocess the probe image. The image is then pose and illumination compensated using the appropriate mapping.

2.2 Proposed Algorithm

The block matching algorithm used for motion estimation in current video coding standards such as MPEG[4, 5] is the basic framework for the registration algorithm in order to compensate for variations in pose, expression and illumination between the captured probe images and the stored gallery images in a face recognition system. The mapping between probe and gallery images is found which converts the probe image into the gallery image. The block matching algorithm was first introduced to perform motion estimation and compensation for video compression in order to take advantage of temporal correlations between video frames by estimating the current frame from the previous frame[4, 5]. An estimate of the current block can be obtained by searching similar blocks in the previous encoded (or original image) frame in a predetermined search area. The block matching algorithm is used for motion estimation between two video frames for compression and in our case, the block matching algorithm is used for disparity estimation between the probe image and each gallery image. Strictly speaking, there is no concept of time between the gallery and probe image as in the video compression problem nor is there a concept of disparity between two camera views as found in the stereo correspondence problem. However, the basic idea is the same in that we are trying to get a disparity map or correspondence between the probe image and test image. The block matching algorithm assumes simple translational motion or disparity within the image plane which is constant over a small block size. A straightforward variation of the BMA is the full search algorithm (FS) or exhaustive search algorithm that finds the best match by exhaustively searching every pixel location within a predetermined search range. The image to be represented is partitioned into distinct blocks, and a match is found for each block within a specified search area in the search image.

In order to make the BMA robust to changes in illumination and pose we modify the MAD cost function to include a multiplicative term *a(i,j)* for illumination variations, so that the modified block matching criterion becomes:

$$E(d_i,d_j)=\sum_{\substack{search\\region}}\left|Y(i,j)-a(i+d_i,j+d_j)X(i+d_i,j+d_j)\right| \tag{1}$$

where Y is the image to be represented, X is the search image to be mapped into the image Y, and a is assumed to be a constant over a small region (MxN), i.e.

$$Y(i,j) = a(i+d_i, j+d_j)X(i+d_i, j+d_j)$$
$$0 \le i \le M-1 \qquad (2)$$
$$0 \le j \le N-1$$

Note that the block size for illumination and disparity do not have to be identical. For the images that we have processed with a resolution of 256x256, we have found that a illumination block size of 8x8 and a disparity block size of 16x16 yields good results. Similarly for the images with a resolution of 864x864, we have found that a illumination block size of 27x27 and a disparity block size of 54x54 yields good results.

Let $Y'(i,j)$ represent a vector of length $L(M \times N)$ of the concatenated block $Y(i,j)$ and $X'(i+d_i,j+d_j)$ represent a vector of length L of the concatenated block $X(i+d_i,j+d_j)$. The least squares solution for a can be expressed as

$$a(i+d_i, j+d_j) = \frac{\left\langle Y'(i,j), X'(i+d_i, j+d_j) \right\rangle}{\left\langle X'(i+d_i, j+d_j), X'(i+d_i+j+d_j) \right\rangle} \qquad (3)$$

In order to compensate for illumination variations between the two images, we modify the traditional similarity metric to incorporate a multiplicative term that is constant over a small area, solved using least squares and represents illumination differences between the images.

2.3 Preprocessing Principle

A bijective function (one-to-one correspondence or bijection) is a function that is both injective ("one-to-one") and surjective ("onto"). More formally, a function $f: X \rightarrow Y$ is bijective if for every y in the codomain Y there is exactly one x in the domain X with $f(x) = y$.

This concept of a bijective function can be extended to the mapping (vector field) that is obtained using the modified block matching algorithm to find correspondence between the probe and gallery images.

Let us consider the case in which the probe image is the searched image and the gallery image is the image to be represented. A function 'f 'is found which maps blocks in the probe image to blocks in the gallery image,

$$Y = f(X) \qquad (4)$$

where X is the probe image, Y is the gallery image and f represents the mapping between probe and gallery image.

For each distinct block Y_i $(1 \le i \le p)$, we generate a vector representing the location of the best match found in X in a predetermined search region resulting in a disparity field of vectors for every block location. There are three scenarios for the vector field representations. A block in X uniquely represents only one block in Y, which we refer to as one-to-one correspondence. A block in X matches multiple blocks in Y, which we refer to as one-to-many correspondence and a block in X does not represent any block in Y that we refer to as one-to-none correspondence. The main idea behind our approach is that we expect the mapping between two images containing a matched face to have a higher percentage of one-to-one correspondence than the mapping between mismatched faces. Because the block matching algorithm is based on non-overlapping blocks only in the image to be represented (Y), we measure one-

to-one correspondence on a pixel level or the percentage of the entire image *X* (probe image) that has been matched to *Y* (gallery image) as a one-to-one correspondence. This is illustrated in figure 1 where the shaded areas correspond to the bijective mappings.

The measure of bijectivity is given by

$$M_f = \frac{\eta[\{f(X_1) \cup \ldots \cup f(X_p)\} - \{f(X_1) \cap \ldots \cap f(X_p)\}]}{\eta[X]} \tag{5}$$

where *p* is the total number of distinct blocks in *Y*, the numerator in equation (5) is the total number of pixels having one-one correspondence in the search image and the denominator in equation (5) is the total number of pixels in the search image. The mapping that produces the maximum one to one correspondence (bijectivity) is chosen to obtain the preprocessed probe image. The preprocessing step consists of using the mapping obtained with the block matching algorithm to correct for pose/expression variations as well as using the least squares solution for the illumination pattern given by the a's to correct for illumination differences.

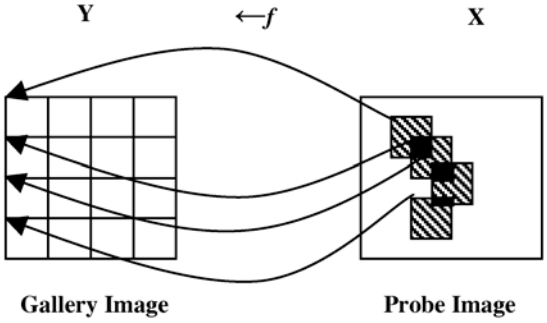


Fig. 1. Mapping between the probe image and gallery image

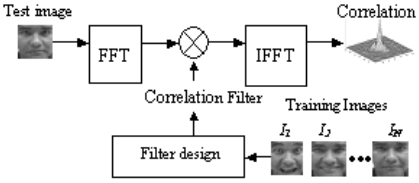


Fig. 2. Correlation filter block diagram shows a filter designed on *N* images for class I. When a test image from class I is input to the system, then the correlation output yields a sharp peak

3 Advanced Correlation Filters (ACF)

3.1 Minimum Average Correlation Energy (MACE) Filters

The MACE filter[6] is designed to minimize the average correlation plane energy resulting from the training images, while constraining the value at the origin to certain pre-specified values. This control of correlation plane energy enables us to control the correlation output shape, not just the peak value. Correlation outputs from well-designed MACE filters typically exhibit sharp correlation peaks making the

peak detection and location relatively easy and robust. Let $c(x,y)$ represent the correlation output resulting from the correlation of the input image $f(x,y)$ using the frequency domain filter $H(u,v)$. Here u and v denote the spatial frequencies associated with x and y.

$$c(x,y) = \text{2D-InvFFT}\{[\text{2D-FFT}\{f(x,y)\}] \,.\, H^*(u,v)\} \tag{6}$$

Here superscript asterisk denotes complex conjugation. Minimizing average correlation plane energy is done efficiently in the frequency domain. Parseval's Theorem allows us to write the energy E_i of the ith spatial correlation plane $c_i(x,y)$ in the frequency domain representation $C_i(u,v)$ as follows:

$$\begin{aligned} E_i &= \sum_{x=0}^{d-1}\sum_{y=0}^{d-1}\left|c_i(x,y)\right|^2 = \frac{1}{d^2}\sum_{u=0}^{d-1}\sum_{v=0}^{d-1}\left|C_i(u,v)\right|^2 \\ &= \frac{1}{d^2}\sum_{u=0}^{d-1}\sum_{v=0}^{d-1}\left|H(u,v)\right|^2\left|X_i(u,v)\right|^2 = \mathbf{h}^+\mathbf{D}_i\mathbf{h} \end{aligned} \tag{7}$$

where input images, frequency domain arrays and correlation outputs are all assumed to be of size dxd and i=1,2,…,N with N denoting the number of training images. Here $\mathbf{D}_i$ is a d^2xd^2 diagonal matrix containing the power spectrum of training image i along its diagonal and $\mathbf{h}$ is a d^2x1 column vector containing the 2-D correlation filter $H(u,v)$ lexicographically reordered to 1-D. MACE filter seeks to minimize the average correlation plane energy defined as follows:

$$E_{average} = \frac{1}{N}\sum_{i=1}^{N}E_i = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}^+\mathbf{D}_i\mathbf{h} = \mathbf{h}^+\mathbf{D}\mathbf{h} \tag{8}$$

where $\mathbf{D} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{D}_i$.

This minimization of $E_{average}$ is done while satisfying the linear constraints that the correlation values at the origin due to training images take on pre-specified values (stored in row vector $\mathbf{u}$), i.e.,

$$\mathbf{X}^+\mathbf{h}=\mathbf{u} \tag{9}$$

Where $\mathbf{X}$ is a d^2xN complex matrix, whose ith column contains the 2-D Fourier transform of the ith training image lexicographically re-ordered into a column vector**.** Minimizing Eq. (8) while satisfying Eq. (9) leads to the following closed form solution for the MACE filter **h.**

$$\mathbf{h} = \mathbf{D}^{-1}\mathbf{X}(\mathbf{X}^+\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{u} \tag{10}$$

The MACE filter in Eq. (10) yields sharp correlation peaks in response to training images and near-training images from the desired class and small output values in response to images from all the other classes. Since $\mathbf{D}$ is a diagonal matrix, we can see from Eq. (10) that the most computationally demanding aspect of determining $\mathbf{h}$ is the inversion of the NxN matrix $(\mathbf{X}^+\mathbf{D}^{-1}\mathbf{X})$. MACE-type filters typically use the peak sharpness measured by the peak-to-sidelobe ratio [7, 8] as a fitness metric rather than the correlation peak height. This advanced correlation filter design has the advantage of the option to include impostor class images in the training set for improved discrimination by specifying a very low peak response in Eq.(9) for the impostor training images.

3.2 Optimal Tradeoff Correlation Filters (OTCF)

Most face images have more energy in low spatial frequencies than in high frequencies and as a result, the MACE filter in Eq. (10) emphasizes high spatial frequencies whereas the maximally noise-tolerant filter[9] typically emphasizes low spatial frequencies. Optimally trading off between noise tolerance and peak sharpness results in the following optimal trade-off filter[10]:

$$\mathbf{h} = \mathbf{T}^{-1}\mathbf{X}(\mathbf{X}^{+}\mathbf{T}^{-1}\mathbf{X})^{-1}\mathbf{u} \tag{11}$$

where $\mathbf{T} = \left(\alpha\mathbf{D} + \sqrt{1-\alpha^2}\,\mathbf{C}\right)$, and $1 \geq \alpha \geq 0$. It is important to note that when α=1, the optimal tradeoff filter reduces to the MACE filter in Eq. (6) and when α=0, it simplifies to the maximally noise-tolerant filter[9].

4 Results

In this section we present our results which are partitioned in to two sub-sections. We first show the results of pre-processing stage then show the results of applying advanced correlation filters for recognition on these pre-processed images. We also compare the verification performance without any preprocessing.

4.1 Preprocessing Results

Figures 3 illustrate two preprocessing examples. The images used are from the UND Biometrics Database[2].

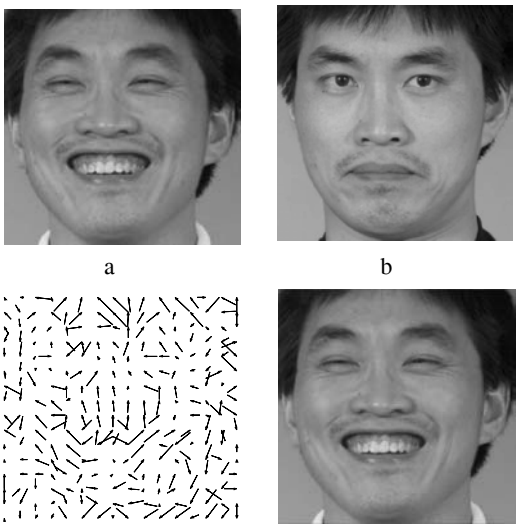


Fig. 3. Preprocessing Example: a) Gallery Image b) Probe Image c) Maximally Bijective Vector Mapping d) Preprocessed Image

Since we require a probe and gallery image in order to preprocess the probe image, we need to obtain a mapping between a probe image and every gallery image in the database and determine the "maximally bijective mapping". In doing this we can make errors in our preprocessing step, i.e. a wrong mapping may turn out to be maximally bijective. In terms of a RoC curve, for optimum performance of our preprocessor, we would like to minimize the number of false accepts (FA) (preprocessing the image with the incorrect mapping). By setting an appropriate threshold, we can minimize the FA rate while maximizing the number of true accepts. In the preprocessing scenario, false rejects (FR) correspond to probe images that are not preprocessed (pass through unchanged) because the bijectivity scores are too low. For each probe image, we determine the ratio of the second best bijectivity score to the top bijectivity score. A threshold based on this ratio is used to determine which images get preprocessed and which images are passed through unchanged. We have found experimentally that a threshold of approximately 0.98 yields good results in terms of the FA rate and FR rates. Figure 4 illustrates that for the 608 probe images in the FRGC -UND biometrics database, a threshold of 0.98 results in 44 probe images which are not preprocessed (FR) and 5 probe images that are preprocessed incorrectly (FA).

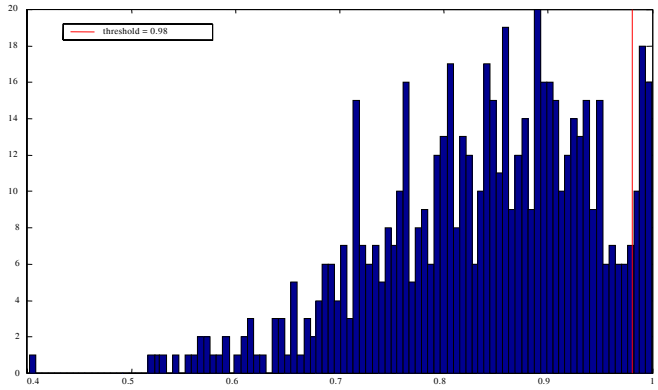


Fig. 4. Histogram of the ratios



Fig. 5. Example how a face image (shown on the left) is partitioned into 9 overlapping image segments (shown on the right). Each part segment is used to synthesize an OTC filter

4.2 Verification Results Using Advanced Correlation Filters

To show the verification improvement using the Bijective-Mapping preprocessing we applied Advanced Correlation Filter approach to the original data and the Bijective-Mapping preprocessed data. We also compared our results to that obtained by the Matched Filter. When synthesizing the correlation filters we followed two approaches, a parts-based and a whole-face approach. In the parts-based approach we segmented the faces into 9 overlapping regions as shown in Fig 5. In the verification phase, the resulting peak-to-sidelobe ratios are averaged across all regions to give a single PSR score (in the case of Matched Filter we used the correlation peak). Figures 6 and 7 show the RoC curves achieved by using different correlation filter types on un-processed and the Bijective-Mapping preprocessed face images correspondingly. Table 1 summarizes the verification rate obtained at 0.1% False Acceptance Rate, showing that parts-based Advanced Correlation Filter configuration performs the best and obtains a performance gain of 9.4% using the Bijective-Mapping preprocessing.

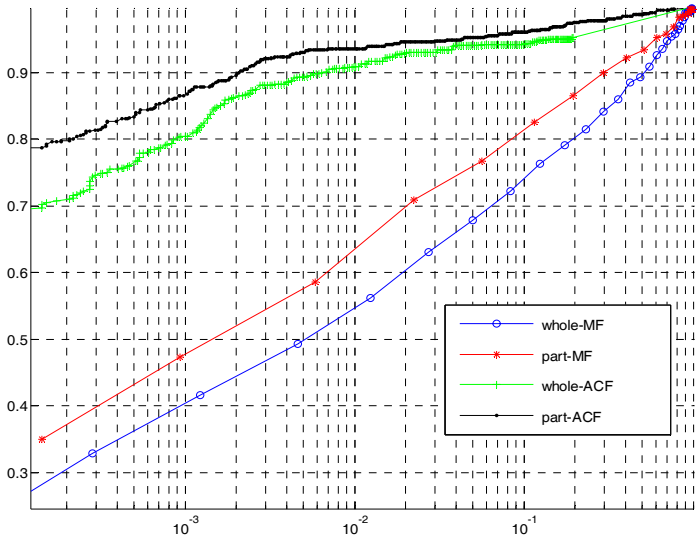


Fig. 6. ROC plots showing the performance of Matched Filter, Part-based Matched Filter, Advanced Correlation Filter (ACF) and Parts-based Advanced Correlation Filter on original unprocessed images

Table 1. Verification accuracy (un-normalized) of the different correlation filter approaches at 0.1% FAR

Correlation Filter Type	Recognition Rate at 0.1% False Acceptance Rate(FAR)	
	Original Images	**Bijective-Mapping Preprocessed Images**
Matched Filter (whole-face)	40.4 %	90.1 %
Matched Filter (parts-based)	47.8 %	93.2 %
Optimal Trade-off Filter (whole-face ACF)	80.4 %	92.7 %
Optimal Trade-off Filter (parts-based ACF)	86.6 %	96 %

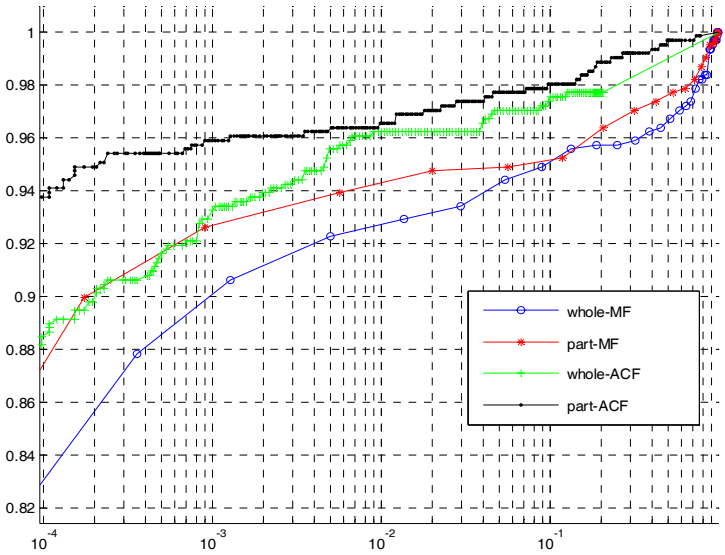


Fig. 7. (un-normalized) ROC plots showing the performance of Matched Filter, parts-based Matched Filter, Advanced Correlation Filter (ACF) and parts-based Advanced Correlation Filter using the Bijective-Mapping preprocessed images

5 Conclusions

The Face Recognition Grand Challenge dataset is challenging as the experiment we are working on only allows for a single gallery image to train on for each person. In such a scenario, we show that a parts-based Advanced Correlation Filter approach can produce significantly higher verification accuracy compared to the parts-based and whole-face Matched Filter. We also show that verification performance can be further improved using the novel Bijective-Mapping preprocessing approach by bringing the best overall verification rate to 96% with a 9.4% increase in accuracy. In the future we plan to explore the verification performance using Advanced Correlation Filters and Bijective-Mapping on harsh overhead illumination experiments.

Acknowledgements

This research is sponsored by the Technical Support Working Group (TSWG).

References

1. R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, pp. 705-741, 1995.
2. P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, "Overview of the Face Recognition Grand Challenge," NIST Interagency Report 2004.
3. C. Podilchuk, A.Patel, A. Harthattu and R. Mammone, "Face Recognition using Bijective Mappings," submitted to Computer Vision and Pattern Recognition(CVPR), 2005.

4. S.C. Han and. C.I. Podilchuk, "Video Compression using Dense Motion Fields," *IEEE Transactions on Image Processing*, vol. 10, pp. 1605-1612, 2001.
5. T. Sikora, "MPEG Digital Video-Coding Standards," *IEEE Signal Processing Magazine*, vol. 14, pp. 82-100, 1997.
6. A. Mahalanobis, B.V.K. Vijaya Kumar, D. Casasent, "Minimum average correlation energy filters," *Applied Optics*, vol. 26, pp. 3633-3640, 1987.
7. M. Savvides, B.V.K. Vijaya Kumar, P.K. Khosla, "Face verification using correlation filters," Proc. of the Third IEEE Automatic Identification Advanced Technologies (Auto-ID), Tarrytown, New York, 2002.
8. M. Savvides and B. V. K. Vijaya Kumar, "Efficient design of advanced correlation filters for robust distortion-tolerant face recognition," Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance(AVSS), 2003.
9. B. V. K. Vijaya Kumar, "Minimum variance synthetic discriminant functions," *J.Opt.Soc. Am.*, vol. A 3, pp. 1579-1584, 1986.
10. P. Refregier, "Optimal trade-off filters for noise robustness, sharpness of the correlation peak, and Horner efficiency," *Optics Letters*, vol. 16, pp. 829-831, 1991.

Photometric Normalisation for Face Verification

James Short, Josef Kittler, and Kieron Messer

Centre for Vision, Speech and Signal Processing
University of Surrey, UK
{j.short,j.kittler,k.messer}@surrey.ac.uk

Abstract. Further to previous work showing the superiority of the preprocessing algorithm developed by Gross and Brajovic, we propose improvements that remove the need for parameter selection. In extensive experimentation on the XM2VTS database, the Yale B database and the BANCA database, we show that our method of automatic parameter selection can produce better results than setting the parameter to a single value for the whole database.

1 Introduction

The benefits of using a face as a biometric is in the ease with which a face image can be captured. That is, from a distance with a non-intrusive camera. Unfortunately, the ease of capturing an image of a persons face is twinned with associated problems due to the nature of the image capture. The image of the face varies with the angle of the camera and with the illumination of the face. In fact, it is true that in general the variation between images of different faces is smaller than that of the same face taken in a variety of environments [2]. As a result, the popular appearance based methods suffer. Although they perform well on images of faces captured in environments similar to that of the training set, they lack the ability to extrapolate to novel conditions.

There are two approaches to this problem. Firstly, we can attempt to model the variation caused by changes in illumination, so as to generate a template that encompasses all possible environmental changes. Secondly, we can try and remove the variation and normalise the input images to some state where comparisons are more reliable. Belhumeur and Kriegman used the former approach [5]. The illumination cone method attempts to model the set of images of an object that can be generated by all possible light source combinations. They showed that this set of images forms a low dimensional convex cone, and demonstrated that the cone could be derived from nine images [14]. Unfortunately, this method requires training images to be illuminated by point light sources in a particular configuration. Basri and Jacobs and also Ramamoorthi have shown in recent work, that spherical harmonics can be used to represent more complicated lighting configurations [4, 17]. Using a second order model it is possible to represent over 98% of the possible variation in the reflectance function in a nine dimensional linear subspace. Three dimensional information about the face is required for a subject template using this method and as such a large amount of training data is

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 617–626, 2005.
© Springer-Verlag Berlin Heidelberg 2005

needed. Vasilescu and Terzopolous have extended the principal component analysis method, eigenfaces, as developed by Turk and Pentland [19], into a multimodel representation [20]. The eigenfaces method cannot differentiate between variance caused by inter-subject variation from variance caused by intra-subject variation. As a solution, the tensorfaces representation uses N-mode singular value decomposition where image changing factors such as subject, illumination, pose and expression are represented independantly. They have shown that this approach can represent a face in fewer components than the eigenfaces method. This method also requires a large amount of training data.

The second approach to the illumination problem was to use pre-processing to remove the effects of illumination before verification. This removes the need for large amounts of training data. Land [13] showed that an image could be thought of as the product of two functions, reflectance and luminance. The reflectance function is dependant only upon the albedo of the face and the luminance function is determined by the surface normals of the face and by the position of the light source. The luminance function was then estimated as a low pass version of the original image, thus finding the reflectance function by dividing the image by the luminance function. Rahman [16] improved upon Lands work by estimating the luminance function as a weighted combination of images generated by convolving the original image with Gaussians of varying widths. Two well known methods for photometric normalisation are homomorphic filtering and histogram equalisation. Homomorphic filtering takes the log of the image to seperate the reflectance and luminance functions before low pass filtering is carried out and the exponential of the result is taken. The output image has normalised brightness and amplified contrast [8]. Histogram equalisation improves the contrast in an image. The histogram of pixel intensities in a poorly illuminated image is generally skewed towards the lower values. As a result, the majority of pixel intensities occur over a small range with little contrast. Using histogram equalisation, the pixel intensities are mapped to an even distribution, thus improving the contrast. Gross and Brajovic estimated the luminance function as an anisotropically smoothed version of the original image [9]. The smoothing was modulated by edge features in the original image, in such a way as to preserve the structure of important features. Unfortunately, as we shall show, face verification rates are very sensitive to the amount of smoothing and as yet there is no method of calculating the amount of smoothing needed.

1.1 Previous Work

Previous work by the authors [18] compared five photometric normalisation algorithms. A method based on principal component analysis, multiscale retinex [16], homomorphic filtering [8], a method using isotropic smoothing to estimate the luminance function and Gross and Brajovic's method using anisotropic smoothing [9]. Three contrasting databases were used in the experiment. The Yale B database has only ten subjects but contains a large range of illumination conditions [7]. The images in the XM2VTS database were all captured under a controlled environment in which illumination variation is minimised [15]. The

BANCA database contains much more realistic illumination conditions [12]. The methods were tested extensively on the three databases using numerous protocols. The results showed that the anisotropic method yields the best results across all three databases.

1.2 Overview of This Paper

In this paper, we present and evaluate an automatic method for selecting the smoothing parameter for the pre-processing algorithm of Gross and Brajovic. The next section details the photometric normalisation algorithm. Section 3 describes the face databases how they were used for testing. Section 4 presents the results of the various face verification tests and we conclude with section 5.

2 The Photometric Normalisation Algorithm

2.1 The Existing Method

The existing normalisation method as presented by Gross and Brajovic [9] is based on the assumption that an image, $I(x, y)$, of a scene is the product of two functions, namely reflectance $R(x, y)$ and luminance $L(x, y)$ [10].

$$I(x, y) = L(x, y).R(x, y) \tag{1}$$

Firstly, the luminance function is dependant only on the geometric properties of the scene (i.e. the surface normals and the position of the light source or sources). Secondly, the reflectance function is dependant on the reflectivity or albedo of the surface. Clearly of these two functions, the reflectance function is invariant to illumination and therefore a desriable function to use as a representation.

The method proposed, first estimates the luminance function, and uses that to modify the image to estimate the reflectance function. The luminance function was estimated as an anisotropically smoothed version of the original image. This smoothing is modulated by a measure of local contrast, such as Webers local contrast, in such a way as to not smooth along prominent features. This is a more sophisticated method than Land's gaussian smoothing [13] and produces a more realistic luminance function. Whereas gaussian smoothing removes all high frequency components, this varying smoothing function retains the structure of the edges of important features, such as the outline of the nose and eyes.

The smoothing of the original image is carried out by minimizing the cost function in equation 2. This is carried out using multigrid methods [1, 6].

$$J(L) = \int_y \int_x \rho(x, y)(L - I)^2 dxdy + \lambda \int_y \int_x (L_x^2 + L_y^2) dxdy \tag{2}$$

where ρ is Webers local contrast between a pixel a and its neighbour b in either the x or y directions. It is the contrast between two neighbouring pixels, weighted by the local illumination value.

$$\rho_{\frac{a+b}{2}} = \frac{|I_a - I_b|}{\min(I_a, I_b)} \tag{3}$$

The cost function has two conflicting terms. The first is the difference term, which increases cost as the luminance function differs from the original image. The second cost is a gradient term, which penalizes large gradients in the luminance function. Two variables control the balance between these two costs. The local contrast variable ρ is determined by the original image. Changing the value of λ therefore allows us to control the amount of smoothing carried out.

An illustration of the process is shown in figure 1.



Fig. 1. Example of the anisotropic smoothing method. Left: Original image. Middle: Estimate of luminance function. Right: Estimate of reflectance function. Image taken from the YaleB database

We found that including histogram equalisation as a final step in the image processing significantly improved verification results.

2.2 Drawbacks of the Method

There is a single parameter, λ, that controls the performance of the normalisation. λ determines the amount of smoothing used in estimating the luminance function. In first presenting the method, Gross and Brajovic tuned the value of the smoothing parameter by hand for each individual image. For the purposes of comparing this normalisation with other methods, we found a single optimal value of the parameter for each database tested. Clearly, neither of these approaches are sufficient for a fully automated face verification system. In this paper, we show that the error rates of verification are sensitive to correct parameter selection and present a method of finding the optimal value of λ for each probe image presented to the system.

2.3 The Face Verification Software

For the purposes of verification we generated scores using a system developed by Kittler et al [12]. When a probe image claims an identity, the data is first projected into a PCA subspace and then into a client specific LDA subspace.

In this subspace, the probe image is then compared with both the mean of the data representing the client whose identity is being claimed and the mean of the data representing an imposter class. These two comparisons are then fused to generate a score. The score is then compared with a threshold (see section 3) and the claim either accepted or rejected.

2.4 The Modified Method

We propose that the score function described in section 2.3 can be used to evaluate the standard of processing as a function of λ. Figure 2 shows an example of how the score changes as a function of λ.

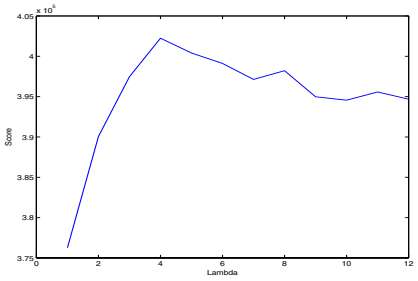


Fig. 2. Claim score as a function of λ

We can see that the score quickly rises to a maximum value and then gradually declines. This maximum value corresponds to the value of λ where the probe image is most similar to the gallery image.

When a claim is made, the probe image would be pre-processed a number of times, with differing values of λ. The score would be calculated for each processed image and the maximum used for comparison with the threshold. For the purpose of this paper, a fixed set of λ values were used, but it would be possible for a search routine to find the maximum score without processing the probe image for every value of λ in the set. However, in addition to finding the maximum score, we also evaluated a similar method that found the mean of all of the scores generated. In this way, the score is in effect the integral of the graph, shown in figure 2, over the range of the set of λ values.

We now assess the effect on face verification rates of using the two methods.

3 Testing

3.1 The Databases

The normalisation was tested on three contrasting databases, The Yale B database (using frontal poses only), the XM2VTS database and the BANCA database.

The Yale B database contains 64 different illumination conditions for 10 subjects. The illumination conditions are a single light source, the position of which

varies horizontally (from -130° to 130°) and vertically (from –40° to 65°). The database was split in to two groups of five subjects. Because of the limited size of the database, the verification software was trained (PCA and Client Specific LDA subspaces were generated) on the BANCA database. The frontally illuminated images were used as the gallery images and the remaining images were used as probes. Each probe image made a claim to each of the five gallery images in its group. A score was generated describing the level of belief in each claim (see section 2.3). Using the resulting scores for each group, the threshold corresponding to the equal error rate, where false acceptance rate equals false rejection rate, was found. This threshold was applied to the alternate subset to find the error rate of verification for that group. The error rates for the two groups were averaged to find a final total error rate.

The XM2VTS database contains images of 295 subjects, captured over 4 sessions in a controlled environment. The database uses a standard protocol. The Lausanne protocol splits the database randomly into training, evaluation and test groups [15]. The training group contains 200 subjects as clients, the evaluation group contains an extra 25 subjects as impostors and the testing group another 70 subjects as impostors. There are two configurations of the XM2VTS database, C1 and C2, differing in the way in which images were selected for the training, evaluation and test sets.

The BANCA database [3] was captured over twelve sessions in three different scenarios and has a population of 52 subjects (26 male and 26 female). Sessions 1–4 were captured in a controlled scenario, sessions 5–8 were captured in a degraded scenario which was captured using a simple web cam and session 9–12 were captured in an adverse scenario. The controlled scenario was a well lit (frontally) environment and the subjects maintain a constant pose. Images were captured with a high quality camera. The degraded scenario was similar to the controlled scenario, except the images were captured with an inexpensive web camera. The adverse scenario again used the high quality camera, but the lighting and the pose of the subject were not constrained. The BANCA database has seven configurations of training and testing data incorporating different permutations of data from the twelve sessions. The seven configurations are Matched Controlled (MC), Matched Degraded (MD), Matched Adverse (MA), Unmatched Degraded (UD), Unmatched Adverse (UA), Pooled test (P), and Grand test (G). The content of each configuration is described by table 1. T represents clients for training, I impostors for testing and C represents clients for testing.

As with the Yale B test, the XM2VTS and BANCA databases are split into two groups. The subspaces for each group were generated using images from the other and the thresholds were found in the same manner as with the YaleB database.

4 Results

In this section we present results showing the accuracy of face verification. We show how this accuracy varies with the value of the parameter λ and compare

Table 1. How different sessions are used for the protocols of the BANCA database

Session	MC	MD	MA	UD	UA	P	G
1	TI			T	T	TI	TI
2	CI					CI	CI
3	CI					CI	CI
4	CI					CI	CI
5		TI		I		I	TI
6		CI		CI		CI	CI
7		CI		CI		CI	CI
8		CI		CI		CI	CI
9			TI		I	I	TI
10			CI		CI	CI	CI
11			CI		CI	CI	CI
12			CI		CI	CI	CI

these results with those generated by taking the maximum score as a function of λ (denoted Max) and with those generated by averaging over the range of values of λ (denoted Mean).

Table 2 shows the error rates for the Yale B database. The lowest error rate for a fixed value of λ occurs at $\lambda = 2$, giving an error rate of 14.28%. Taking the λ corresponding to the maximum score for each image, gives an error rate of 14.31%. Although this value is slightly higher, it does not require any additional training. Taking the mean score across the images yields the best result of 14.16% error rate.

Table 2. Error rates (%) for fixed values of lambda and for the two combination methods of the Yale B database

Lambda	0.5	1	1.5	2	2.5	3	Max	Mean
T	18.95	16.33	14.32	14.28	14.43	15.70	14.31	14.16

Table 3 shows the error rates for the two configurations of the XM2VTS database. The best accuracy for configuration one, attained through using a single value of λ occurs at $\lambda = 7$ giving an error rate of 3.73%. For configuration two, the lowest error rate is 3.44%, but this occurs at $\lambda = 8$. Clearly, it is not possible to achieve both these results by using a single value of λ. Using the value of λ corresponding to the maximum score, gives error rates of 3.15% and 2.83%, an improvement over using the optimum fixed λ values for either configuration one or two. Using the mean score across the different values of lambda also yields improved results, but not better than the using the maximum score.

Table 4 shows the error rates for the seven different protocols of the BANCA database. Using a single value of λ for the whole database can yield some very good results for certain protocols. However, that value of λ does not necessarily generate the best result on another protocol. For example, taking λ equal to

Table 3. Error rates (%) for fixed values of lambda and for the two combination methods for both configurations of the XM2VTS database

Lambda	3	4	5	6	7	8	9	10	Max	Mean
C1	9.22	7.41	5.94	5.41	3.73	3.97	6.80	4.50	3.15	3.56
C2	7.39	7.25	5.41	4.40	3.70	3.44	3.48	3.65	2.83	3.08

11 gives the best score for the Matched Controlled protocol (MC), but using this value of λ on the Pooled protocol (P) gives error of 13.18% in comparison with 11.79%, the best error rate for that protocol. The strategy of taking the maximum value yields very good results for each protocol. Although it does not always show higher accuracy than using an individual value of λ, it does remove the need for finding λ and also removes the problem of the variation in results across different protocols.

Table 4. Error rates (%) for fixed values of lambda and for the two combination methods for each of the protocols of the BANCA database

λ	4	5	6	7	8	9	10	11	12	13	14	Max	Mean
MC	5.82	5.22	5.22	4.79	4.78	4.92	4.73	4.65	5.42	5.69	5.53	4.84	4.44
MD	5.10	4.70	4.41	4.28	4.15	3.91	3.81	3.49	4.46	4.34	3.78	4.15	3.51
MA	8.13	7.68	7.60	7.58	8.25	7.44	6.78	7.26	6.97	6.97	6.60	6.27	6.20
UD	8.51	8.61	7.63	8.54	8.30	7.85	8.93	9.05	9.47	9.34	9.01	7.98	7.58
UA	19.10	19.79	18.97	18.89	19.04	20.34	20.59	20.77	20.51	20.69	20.03	17.95	18.14
P	12.07	12.52	11.79	11.85	11.94	12.15	12.87	13.18	13.09	13.25	13.02	11.64	11.35
G	3.71	3.55	3.46	3.26	3.22	3.16	3.01	3.00	3.17	3.41	3.33	2.77	2.56

The results demonstrate the large variation of the optimal value of λ. The optimal value of λ is shown to be 2 for the Yale B database, the two configurations of the XM2VTS database give an optimal value of 7 and 8. For the BANCA database the optimal value of λ varies between 6 and 14 depending on the chosen protocol. Also, the results show that a change in λ can give rise to a large change in verification error rate. For example, the optimal value of λ for the unmatched degraded protocol of the BANCA database occurs at $\lambda = 6$. For a change in λ of $\pm$ 1, the error rate increases by more than 10%.

5 Conclusions

We have presented a method of selecting the parameter value for the photometric normalisation algorithm on an image by image basis. The parameter is selected by evaluating the quality of a number of pre-processed images of varying λ, and the image with the maximum score is used for verification. This evaluation is carried out using a score function based on client specific linear discriminant analysis. In addition we have evaluated a method that uses a number of processed images and finds the mean of the resulting scores.

The methods have been extensively tested on three contrasting databases. Results from all three databases illustrate the sensitivity to the selection of the value of λ. It has also been shown that for using a fixed value of λ, optimal parameter selection differs greatly as a results of different configurations of the training, evaluation and testing data. As such, although the results for fixed λ show the error rates given by the optimally tuned pre-processing algorithm, it is not necessarily the case that the correct value of λ will be found accurately.

Results from the Yale B database have shown that using the mean method yields better results than using a fixed value of λ. Using the max method yields results that are nearly as good as using the fixed optimal value of λ, however does not suffer from the previously mentioned problem of finding this value. The XM2VTS and BANCA databases show significant improvements using both the max method and the mean method. In the case of the XM2VTS database, the max method performs better and in the case of the BANCA database, the mean method performs better.

The max method has additional advantages over the mean method. For the purposes of the evaluation, a number of pre-defined values of λ were used, from which the max value of the score was found and the mean value of the score calculated. For the mean value to be a valid comparison it must always be calculated over the same range of values of λ. In contrast, the max value can be found using a search for the peak in the score versus λ function. Once the peak has been found, it is no longer necessary to process more images. As such it may require fewer evaluations.

References

1. S. Acton, "Multigrid Anisotropic Diffusion" IEEE Trans. Image Processing **vol. 7** (1998)
2. Y. Adini, Y. Moses, and S. Ullman, "Face recognition: the problem of compensating for illumination changes" IEEE Trans. Pattern Anal. Mach. Intelligence **vol. 19(7)** (1997) 721–732
3. E. Bailly-Bailliere, et al., "The BANCA Database and Evaluation Protocol" AVBPA (2003)
4. R. Basri, D. Jacobs, "Lambertian Reflectance and Linear Subspaces" IEEE Trans. Pattern Anal. Mach. Intelligence **vol. 25(2)** (2003) 218–233
5. P. Belhumeur, D. Kriegman, "What is the Set of Images of an Object Under All Possible Lighting Conditions?" IEEE Proc. Conf. Computer Vision and Pattern Recognition (1996)
6. W. Briggs, V. Henson, S. McCormick, "A Multigrid Tutorial" Siam, Second ed.
7. A. Georghiades, P. Belhumeur, D. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose" IEEE Trans. Pattern Anal. Mach. Intelligence **vol. 23** (2001) 643–660
8. R. Gonzalez, R. Woods, "Digital Image Processing" Prentice Hall, Second ed.
9. R. Gross, V. Brajovic, "An Image Preprocessing Algorithm" AVBPA (2003) 10–18
10. B.Horn, "Robot Vision" MIT Press (1998)
11. S. Kee, K. Lee and S. Lee, "Illumination Invariant Face Recognition Using Photometric Stereo" IEICE Trans. Inf & Syst **Vol.E83-D No.7** (2000)

12. J. Kittler and Y. P. Li and J. Matas, "Face verification using client specific Fisher faces" The Statistics of Directions, Shapes and Images (2000) 63–66
13. E. Land, J. McCann, "Lightness and Retinex Theory" Journal of the Optical Society of America **vol. 61** (1971) 1–11
14. K. Lee, J. Ho, D. Kriegman, "9 Points of Light: Aquiring Subspaces for Face Recognition Under Variable Lighting" IEEE Proc. Conf. Computer Vision and Pattern Recognition (2001)
15. K. Messer, J. Matas, J. Kittler, "XM2VTSDB: The extended M2VTS Database" AVBPA (1999)
16. Z. Rahman, G. Woodell, D. Jobson, "A Comparison of the Multiscale Retinex with other Image Enhancement Techniques" Proceedings of the IS&T 50th Anniversary Conference (1997)
17. R. Ramamoorthi, "Analytic PCA Construction for Theoretical Analysis of Lighting Variability, including Attached Shadows, in a Single Image of a Lambertian Object" IEEE Trans. Pattern Anal. Mach. Intelligence **vol. 24(10)** (2002) 1322–1333
18. J. Short, J. Kittler, K. Messer, "A Comparison of Photometric Normalisation Algorithms for Face Verification," Proc. Automatic Face and Gesture Rocgnition (2004) 254–259
19. M. Turk, A. Pentland, "Eigenfaces for Recognition" J. Congitive Neuroscience **vol. 3** (1991) 71–86
20. M. Vasilescu, D. Terzopoulos, "Multilinear Anaysis of Image Ensembles: TensorFaces" Proc. European Conf. Computer Vision (2002) 447–460

Experimental Evaluation of Eye Location Accuracies and Time-Lapse Effects on Face Recognition Systems

Haoshu Wang and Patrick J. Flynn

Computer Science and Engineering Department
University of Notre Dame
Notre Dame, IN., USA

Abstract. It is claimed that eye location accuracy is very important to face recognition system performance. In most systems, the eye locations are the most significant facial landmark for the preprocessing step. Eye location estimates can be assessed in absolute terms (e.g., proximity to known eye location) and also in application-specific terms (e.g., performance of a system that employs the location). This paper assesses an automatic commercial eye-finding system in absolute and application-specific terms, using four different face recognition systems and a database of thousands of images. A pilot study on the time-lapse effect suggests that with the time-lapse increasing, the face recognition performance will degrade. Our experiments examine this effect by using a large image dataset, which has a time-lapse up to two years, with 250 subjects and 64300 probes. Experiment results show eye location accuracy is significant to face recognition system performance. Different systems can have different level of sensitivity, and the system using local feature analysis is less sensitive to eye location accuracy. Also all the algorithms tested in this study show that time-dependency exists in face recognition system.

1 Introduction

Appearance-based and geometry-based techniques are the major approaches employed in face recognition. Regardless of the approaches, accurate registration is a crucial issue. In most cases, eyes can be the most reliable features for the image registration, in that eye positions are not easily affected by other face changes; the interocular distance can be used to normalize the face image and the orientation of the interocular line can be used to correct the head pose. Also, the eye is often viewed as the most important feature of the face. Hjelmas *et al.* [1] proposed a face recognition system, which employed eyes as the only facial feature to recognize the face, and it obtained a surprisingly high correct classification rate of 85%.

As the results of the importance of accurate eye locations, there have been many novel automatic eye extraction or eye location detection algorithms developed during the last twenty years, such as eigenfeature-based methods [2], de-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 627–636, 2005.
© Springer-Verlag Berlin Heidelberg 2005

formable template-based methods [4] [5] [6] [7], Gabor wavelet filter methods [8], variance projection function algorithms [9], and the neural network method [2].

In recent years, there are some researchers starting to explore how the eye location accuracy affects the face recognition system performance. Marques *et al.* [14] conducted a study on eigenface-based face recognition. Their experimental results suggest that the eigenface algorithm is more sensitive to eye positions that deviate above or below the enrolled reference than those that deviate left or right from the enrolled reference. Riopka and Boult [15] conducted an eye perturbation sensitivity analysis. The experimental results strengthened the conclusion that the eye localization is important to the accuracy of face recognition systems and also suggested that the correct measurement of eye separation is more important than the correct eye location. As the result of that, they predicted that for better recognition performance, improving the eye localization accuracy might be more effective than improving the face recognition engine itself. In this paper, we examine the effect of eye location accuracy on face recognition systems' performance and sensitivities of different face recognition systems to the eye location accuracy.

Besides eye location accuracy, the time-laspe effect is another important factor, which is crucial to the face recognition system to be used in the really world. Previous pilot study, based on a relatively small and short time-lapse database, suggests that with time-lapse increasing, the face recognition performance will degrade[3] [12]. Our experiment examined this effect by using a large image dataset, which has a time-lapse up to two years for 250 subjects and 64300 probes.

To examine both the eye location accuracy and time-lapse effects, four different face recognition systems are chosen for this study, Principal Components Analysis (PCA), Elastic Bunch Graph Matching (EBGM), Principle Component Analysis followed by Linear Discriminant Analysis (PCA+LDA), and the FaceIt system (version G5). The first three systems are implemented in software distribution at the Colorado State University [10], while the fourth one is a commercial face recognition system from Identix [11]. In this study, all the facial images are drawn from the database developed at the University of Notre Dame [12].

The remainder of this paper is organized as follows. Section 2 summarizes the four different face recognition systems used in this study. Section 3 describes the data collection, eye locating techniques employed and the evaluation of the automatic eye location system's performance. Section 4 shows the extensive experimental results and discussion. The conclusion and future work are summarized in Section 5.

2 Face Recognition Algorithms

Four different systems, PCA, PCA+LDA, EBGM and FaceIt, are chosen for this study. They are described briefly below.

2.1 PCA

PCA method implements nearest-neighbor search in a projection space to perform recognition. By using the training images, PCA projects images into a subspace, where the first dimension of this subspace captures the greatest amount of variance among the images and the last dimension of this subspace captures the least amount of variance among the images. After the subspace is created, both gallery images and probe images are projected into the eigenspace. Finally, probe images are identified by comparing them with the projected gallery images using nearest-neighbor search and the one with the highest similarity is the identified subject.

2.2 PCA+LDA

The PCA approach optimizes variance among the images, while Fisher's linear discriminant optimizes discrimination characteristics. The LDA method groups images of the same class and separates images of different classes. Images are projected from N-dimensional space (where N is the number of pixels in the image) to C-1 dimensional space (where C is the number of classes of images). The LDA method finds a linear transformation that maximizes the between-class variance and minimizes the within-class variance, to distinguish different subjects. By using the PCA+LDA algorithm [13], a PCA subspace will be built by a set of training images, and then all the training images are projected to the PCA subspace and are grouped according to subject identity. Each subject is treated as a distinct class and the LDA basis vectors are computed. The test images are projected into this subspace and identified using a nearest-neighbor search.

2.3 Elastic Bunch Graph Matching (EBGM)

This method recognizes faces by comparing the facial features and computing the similarity of two images [16]. First, the algorithm locates landmarks on the image, such as the eyes, nose and mouth, which are referred as facial features. The Gabor jet, which is the Gabor wavelet convolution at these points, is computed are used to represent a landmark. Then a face graph is created for each image. The nodes of the face graph are placed at the landmark locations, and each node contains a Gabor jet extracted from that location. Finally, the similarity of two images is computed, which is a function of the corresponding face graph, and the classification is based on this similarity.

2.4 FaceIt

FaceIt uses local feature analysis (LFA), which is more robust to facial expression changes or other extrinsic variations such as lighting changes [11]. LFA uses statistical techniques to encode facial features, and all the facial shapes can be well represented. In FaceIt's implementation of LFA, there are two factors used

to determine the identity of the subject. One is the characteristic elements of the face, and the other one is geometric combination of these elements, such as the relative position. Three steps needed to accomplish the recognition process. First, the eye location needs to be provided or computed for all the images. Second, templates need to be generated for all the gallery images. Third, the matching process is conducted by comparing each pair of the probe image and the gallery template. The template has the highest similarity is the recognized subject.

3 Face Image and Eye Location Data Sets

3.1 Image Data

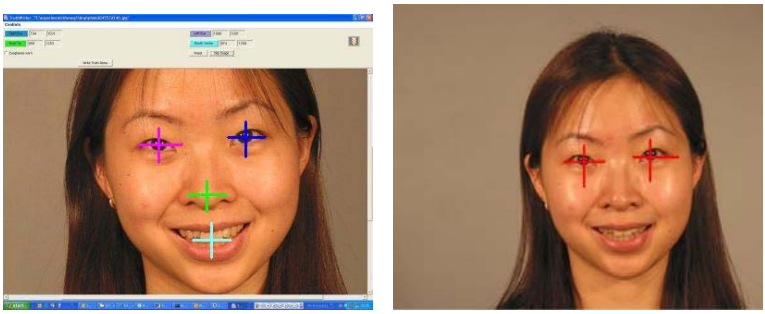
The image data used in this study were collected from Fall 2002 to Spring 2004 at the University of Notre Dame, which form a component of a large database and support a two-year longitudinal (time-lapse) study of face recognition systems' performance [12]. Approximately 200 experimental subjects were photographed weekly with a high-resolution (1704×2272) color digital camera (Canon PowerShot G2 was used) under four different controlled lighting and expression combination conditions, FA|LF, FA|LM, FB|LF, FB|LM [12]. FA refers to a neutral expression and FB is a smile. LF denotes FERET-style facial illumination (two side-lights) and LM denotes side plus center lighting.

3.2 Eye Location Data

In this study, two different eye location sets are provided for each facial image. They are generated by different methods and of different accuracy. The first method is called ground truth writing, which is to locate eye positions manually by using trained human annotators. In ground truth writing, a human operator manually marks the eye pupil center along with the nose tip and the center of the mouth. The eye coordinates obtained in this way are called truth writing (TW) coordinates. Figure 1(a) depicts those marked locations, while the nose tip and mouth center are not used in this study. The second method is called automatic eye locating, which is to locate eyes automatically by using the eye finder function provided by FaceIt system. Representative output from FaceIt appears in Figure 1(b). The eye coordinates generated by this way are called automatic coordinates.

3.3 Eye Location Accuracy Measurements

Comparing these two different eye location sets, the truth writing coordinates are more accurate and treated as the accurate reference locations. For the automatic coordinates, Figure 2 shows some eye location examples and Figure 3 shows the scatter plot for the deviations of the automatic eye coordinates from truth writing coordinates. From the plot, we can tell that there is no bias of the



(a) Eye locations from truth writing. (b)Eye locations from FaceIt.

Fig. 1. Two ways to generate eye locations

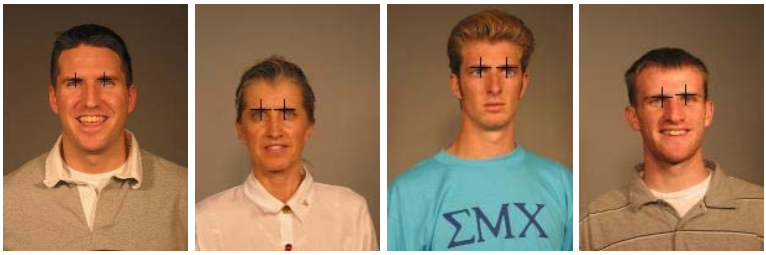


Fig. 2. Examples of inaccurate eye locations generated by FaceIt

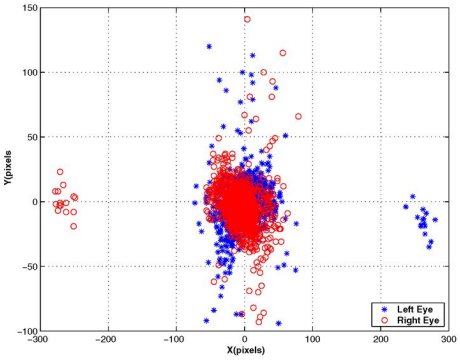


Fig. 3. Scatter plot for deviations of automatic eye coordinates from TW eye coordinates

deviation for either the right eye or the left eye. Also three metrics, as shown below, are used for the quantitative estimation of the accuracy of the automatic eye location.

Metric 1. E_{RMS}, the average of two *Root Mean Square* (RMS) values, $E_{RMS;L}$ and $E_{RMS;R}$, is used to indicate the disparities between automatic coordinates and TW coordinates. $E_{RMS;L}$ is the disparity for the left eye, and $E_{RMS;R}$ is for the right eye. E_{RMS} is given by

$$E_{RMS} = \frac{1}{2}(E_{RMS;L} + E_{RMS;R}) \ ,$$

where

$$E_{RMS;L} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_{iL;T} - x_{iL;A})^2 + (y_{iL;T} - y_{iL;A})^2)} \ ,$$
$$E_{RMS;R} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_{iR;T} - x_{iR;A})^2 + (y_{iR;T} - y_{iR;A})^2)} \ ,$$

$(x_{iL;T}, y_{iL;T})$ and $(x_{iR;T}, y_{iR;T})$ are TW coordinates of the left and right eyes, and $(x_{iL;A}, y_{iL;A})$ and $(x_{iR;A}, y_{iR;A})$ are the corresponding coordinates.

Metric 2. $E_{RMS;max}$ is the RMS error for the larger disparity value comparing two eye coordinates on a facial image.

$$E_{RMS;max} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} max(Distance_L, Distance_R)} \ ,$$

where

$$Distance_L = (x_{iL;T} - x_{iL;A})^2 + (y_{iL;T} - y_{iL;A})^2 \ ,$$
$$Distance_R = (x_{iR;T} - x_{iR;A})^2 + (y_{iR;T} - y_{iR;A})^2 \ .$$

Metric 3. $E_{RMS;mid}$ is the RMS deviation of the interocular midpoint, which is determined by the distance between the TW coordinates and automatic coordinates.

$$E_{RMS;mid} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_{iM;T} - x_{iM;A})^2 + ((y_{iM;T} - y_{iM;A})^2)} \ ,$$

where

$$x_{iM:T} = \frac{x_{iL;T} + x_{iR;T}}{2} \ ,$$
$$y_{iM:T} = \frac{y_{iL;T} + y_{iR;T}}{2} \ .$$

A summary of the accuracy measurement for the automatic eye coordinates appears in Table 1. These values show the average disparities from FaceIt eye coordinates to TW eye coordinates by using three different metrics. This measurement is done on the original images.

Table 1. Automatic Eye Location Accuracy relative to TW coordinates in pixels

	E_{RMS}	$E_{RMS;max}$	$E_{RMS;mid}$
DISPARITIES	14	16	11

4 Face Recognition Experiments

In this section, we will examine the effect of eye location accuracy on face recognition systems' performance as well as the time-lapse effect. The performance of four face recognition systems, PCA, PCA+LDA, EBGM and FaceIt, are compared.

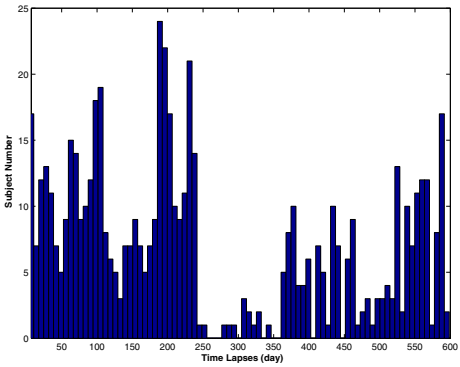


Fig. 4. Time lapse distribution for all subjects' participation

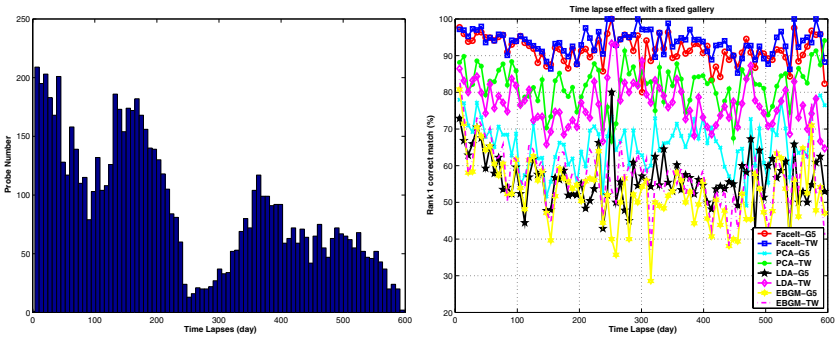
4.1 Experiment Data

All the experiment data we used in this experiment are the images taken from Fall 2002 to Spring 2004. 691 subjects are involved. Figure 4 shows the time-lapse distribution for all subjects' participation. In this study the expression changing factor is excluded, so all the images we use here are taken from subjects with regular expression (FA|LF or FA|LM). To avoid the performance bias, none of the subjects involved in training are used in testing. We divided the whole dataset into two subsets, one set is used as training data and the other one is used as testing data. We chose the 250 subjects who participated in image acquisition over the longest time period as testing subjects, and the remaining 441 subjects were used for the training dataset.

Gallery Data. All the gallery images are FA|LF images. In order to extensively explore the time lapse effect, the experiment is conducted in two ways. One is called the fixed-gallery experiment and the other one is called the running-gallery experiment. The primary difference between them is how the gallery dataset is configured.

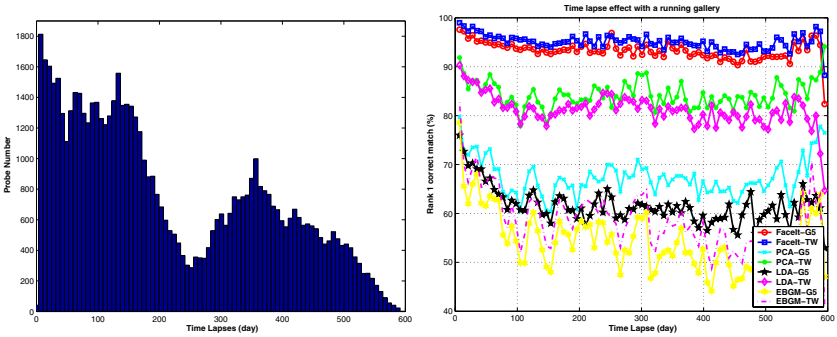
1.Fixed Gallery: For fixed gallery, the gallery contains the earliest image of each of the 250 testing subjects, and the probe contains all the remaining images of these 250 subjects. Each subject has one image in the gallery set. Each probe is an image of a subject which is acquired five or more days after the gallery image.

2.Running Gallery: In the running gallery test, the gallery contains multiple sub-gallery sets and the experiment result for each sub-gallery set will be collected to make the overall performance result. A total of 10 sub-gallery sets are included. The first sub-gallery set contains the earliest images of 250 subjects, and the corresponding probe set contains all the remaining images of these subjects. Each probe is an image of a subject, which is acquired five or more days after the gallery images take. The second sub-gallery set contains the second earliest images of these 250 subjects, and in the same manner, the corresponding



(a) Time lapse distribution. (b)FR system performance

Fig. 5. Fixed-gallery test with 7555 probes



(a) Time lapse distribution. (b)FR system performance

Fig. 6. Running-gallery test with 64300 probes

probe set contains all the subsequent images of subjects taken after images in the sub-gallery set. The rest of the sub-gallery and sub-probe sets are configured in the same way as the first two. The performance results for the running gallery test, shown in the later figure, are the averaged results of the 10 sub-experiments.

Probe Data. All the probe images are FA|LF images. In fixed probe set, a total of 7555 images for the 250 testing subjects are included. For the running probe set, a total of 64300 probe images for the 250 testing subjects are included. Each subject has more than one image in the probe set. The time lapse distribution of the fixed probe set and running probe set are shown in Figure 5(a) and Figure 6(a).

Training Data. All face recognition systems used here except FaceIt need training data. Table 2 shows the configurations for the training datasets for different FR systems. For PCA and LDA, all the training data are taken from the images acquired at the University of Notre Dame, as we described above. For EBGM, the training data are the FERET images, and along with each image

Table 2. The Training Data Configurations

FR Systems	Subj. Num.	Image Type	Num of Images per Subj.	Size
PCA	441	FA|LF	1	441
LDA	200	FA|LF or FA|LM	15	3000
EBGM	70	FERET Data	1	70

25 feature points are manually extracted as well as Gabor jets. Based on the previous study [16], a training dataset size of 70 was chosen for EBGM, and this training dataset is accessible at [10].

4.2 Experiment Results

Experimental results are shown in Figure 5(b) and Figure 6(b). First, the results suggest that eye location accuracy is significant to face recognition engines and the algorithms show different sensitivities to eye location accuracy. FaceIt system has the least sensitivity to eye location accuracy with the best performance. EBGM also shows the relatively low sensitivity. However its performance is the worst one in this experiment. We suspect the performance of EBGM can be improved by using an optimized configuration. For PCA and LDA, both systems can be strongly affected by the eye location accuracy. The experimental results suggest that the local feature analysis, which is used by EBGM and FaceIt, is much less sensitive to eye location accuracy. Second, the result of the running-gallery test shows a relatively clear trend of time-lapse effect. All the algorithms used here indicate that with time-lapse increasing the face recognition system performance decreases.

5 Conclusion

In this study, we examined the effect of eye location accuracy on face recognition systems. Experimental results showed that the eye location accuracy is one of the factors to affect the systems' performance. The sensitivity of different systems can vary, and our experimental results suggest that local feature analysis is less sensitive to eye location accuracy than global feature analysis. We can also conclude that the time-lapse does affect the face recognition performance.

References

1. E. Hjelmas and J. Wroldsen, "Recognizing face from the eyes only," in *Proceedings of the 11th Scandinavian Conference on Image Analysis*, 1999, p. Pattern Recognition.
2. Y. Ryu and S. Oh, "Automatic extraction of eye and mouth fields from a face image using eigenfeatures and multilayer perceptrons," in *Pattern Recognition*, vol. 34, 2001, pp. 2459–2466.

3. J. Min, P. Flynn, and K. Bowyer, "Assessment of time dependency in face recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, under review 2004.
4. A. Yuille, D. Cohen, and P. Hallina, "Feature extraction from faces using deformable templates," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1989, pp. 104–109.
5. C. Chen and C. Huang, "Human face recognition from a single front view," in *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 6, 1992, pp. 571–593.
6. G. Chow and X. Li, "Towards a system for automatic facial feature detection," in *Pattern Recognition*, vol. 26, 1993, pp. 1739–1755.
7. K. Lam and H. Yan, "Locating and extracting the eye in human face images," in *Pattern Recognition*, vol. 29, 1996, pp. 771–779.
8. J. Huang and H. Wechsler, "Eye detection using optimal wavelet packets and radial basis functions (rbfs)," in *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 7, November 1999, p. 1009.
9. G. Feng and P. Yuen, "Multi cues eye detection on gray intensity image," in *Pattern Recognition*, vol. 34, 2001, pp. 1033–1046.
10. R. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The CSU face identification evaluation system user's guide: Version 5.0," May 2003. [Online]. Available: `http://www.cs.colostate.edu/evalfacerec/algorithms/version5/`
11. [Online]. Available: `http://www.identix.com`
12. P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *International Conference on Audio- and Video-Based Biometric Person Authentication*, June 2003, pp. 44–51.
13. G. H. Givens, J. R. Beveridge, B. A. Draper and D. Bolme, "Using a generalized linear mixed model to study the configuration space of a PCA+LDA human face recognition algorithm," in , April 2003.
14. J. Marques, N. Orlans, and A. Piszcz, "Effects of eye position on eigenface-based face recognition scoring," in *Technical Paper of Mitre Corp.*, October 2003.
15. T. Riopka, T. Boult, "The eye have it," in *Proc. of ACM SIGMM MultimediaBiometrics Methods and Applications Workshop*, pp.9–16, Berkeley, CA, 2003.
16. D. Bolme, "Elastic bunch graph matching," in *Master Thesis*, Fort Collins, Colorado, Summer 2003.

Experiments in Mental Face Retrieval

Yuchun Fang[1] and Donald Geman[1,2]

[1] IMEDIA Project, INRIA Rocquencourt
yuchun.fang@inria.fr
[2] Dept. Applied Math. and Stat., Johns Hopkins University
geman@jhu.edu

Abstract. We propose a relevance feedback system for retrieving a mental face picture from a large image database. This scenario differs from standard image retrieval since the target image exists only in the mind of the user, who responds to a sequence of machine-generated queries designed to display the person in mind as quickly as possible. At each iteration the user declares which of several displayed faces is "closest" to his target. The central limiting factor is the "semantic gap" between the standard intensity-based features which index the images in the database and the higher-level representation in the mind of the user which drives his answers. We explore a Bayesian, information-theoretic framework for choosing which images to display and for modeling the response of the user. The challenge is to account for psycho-visual factors and sources of variability in human decision-making. We present experiments with real users which illustrate and validate the proposed algorithms.

Keywords: relevance feedback, mental image retrieval, Bayesian inference

1 Introduction

Traditional image retrieval is based on "query-by-example": starting from an actual image, the objective is to find the images in the database which are visually similar to the query image. Striking results are obtained in special domains, e.g., in comparing paintings, plants and landscapes using the IKONA system [1].

However, in many cases of interest there is no physical example to serve as the query image [2]. Instead, knowledge about the target is based entirely on the subjective impressions and opinions of the user. In other words, the standard query image is replaced by a "mental image". To be concrete, we shall concentrate throughout on face images, although all the algorithms we develop could be applied in other domains, for instance to images of clothes, houses, funitures or paintings. Mental face retrieval has extensive applications in security, e-business, web-based browsing and other areas. Here, as the realization of a study conducted jointly with the SAGEM group, we propose a system for retrieving a mental face image using Bayesian inference and relevance feedback. It is based on an interactive process designed to incrementally obtain knowledge about the target from the responses of the user to a series of multiple choice

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 637–646, 2005.
© Springer-Verlag Berlin Heidelberg 2005

questions. The objective is to minimize the number of iterations until a face is displayed whose identity corresponds to the mental image.

Thus relevance feedback refers to a series of queries and answers. The query is simply a set of displayed images from the database. The answer is the feedback provided by the user. Usually, the opinions or impressions of the user concerning both his target and the displayed images are of high-level, semantic nature, and hence "mental matching" involves human memory, perception and opinions [3, 4]. On the other hand, the representation of the images in the database is generally based on low-level signatures rather than semantic content. This "semantic gap" greatly complicates the task. Indeed the face recognition problem, which is arguably easier, remains largely unsolved, at least with large databases.

Still, if the display and answer models are constructed to explicitly address the issue of coherence, it is possible to incrementally obtain knowledge about the target image. The accumulation of information is represented by an evolving probability distribution over the database, whose entropy is hopefully diminishing (although not monotonically) as information is acquired from the answers. This process of alternating between query and answer is iterated until the user recognizes one of the displayed images as his target, at which point the search terminates. The two primary challenges in mental picture retrieval are then deciding which images to display at each iteration (the "display model") and accounting for the difference between mental matching and signature-based matching (i.e., between mental and feature-based metrics) in designing the conditional probability distribution for the answers given the target (the "answer model").

In our framework, both the target and answers to queries are treated as random variables; the probability distribution of the target evolves over time based on the accumulated evidence from the user's responses. A natural choice for the images to display at each iteration is then the set which maximizes the mutual information between the target and response given all previous answers. As this optimization problem is intractable, a heuristic solution is proposed based on an "ideal" answer model which puts the user and system in synchrony. In addition, in order to find image representations which cohere as much as possible with human decision making, we compare several traditional face recognition signatures. Based on this analysis as well as data collected from human responses, in particular declaring which among a set of displayed images is "closest" to a given target, an answer model is designed for a comparative response. The feasibility of the whole system is demonstrated by estimating mean search times and other summary statistics from mental retrieval experiments with real users.

Whereas there has been considerable work done on face retrieval in the standard setting of query-by-example [5, 6], little has been reported in the case of mental images. Navarret et al. proposed an algorithm based on self- organizing maps [7]; see also the work on "retrieval of ambiguous target" in [8]. Of course, there are many articles on relevance feedback [9], however, most of them involve "category search", which is different from "target search" in the case of mental face retrieval. In our view, the benchmark work on "target search" for mental images is Cox et al [10]; see also the model proposed by Geman and Moquet

[11] for the toy application of mental polygon retrieval. By concentrating on the interactive process and specializing to target search and pairwise comparison tests, the authors in these studies were able to develop ties with Bayesian inference and information theory. However, the answer model in [10], basically a blurring of the actual metric used by the system in comparing two images, is not sufficiently powerful to deal with face retrieval. Moreover, pairwise comparison search is not practical with large image databases. We believe our work constitutes the first comprehensive study of mental face retrieval, both in terms of mathematical foundations and experiments with real users.

The remainder of the paper is organized as follows. The formulation of the problem in terms of Bayesian relevance feedback is described in Section 2. The answer model and display model are explained in detail in Sections 3 and 4 respectively. In Section 5, we discuss signature extraction and analyze the coherence issue. Experimental results are presented in Section 6.

2 Bayesian Relevance Feedback Model

In the framework we propose, mental image retrieval will depend on solving two difficult tasks:

- **A Modeling Problem:** Discovering answer models which match human behavior;
- **An Optimization Problem:** Discovering approximations to the optimal query.

Suppose there are N images in the database $\mathcal{S}$, say $I_1, ..., I_N$. For simplicity, we will identify $\mathcal{S}$ with the index set $\{1, 2, ..., N\}$. One image in the database, Y, is the "target", i.e., the variation on the mental picture assumed to belong $\mathcal{S}$. In the stochastic formulation, Y is a random variable with some initial distribution

$$p_0(k) = P(Y = k), \quad k \in \mathcal{S}.$$

Information about Y is collected from a series of queries. Each query involves two quantities: a subset $\mathcal{D} \subset \mathcal{S}$ of n displayed images and the response of the user, denoted by $X_{\mathcal{D}}$ and taking values in a set $\mathcal{A}$. Obviously $n \ll N$; the choices for n and $\mathcal{A}$ are important issues which will be discussed in the following sections.

The feedback from the user up to time (or iteration) $t = 1, 2, ...,$ is then

$$B_t = \bigcap_{i=1}^{t} \{X_{\mathcal{D}_i} = x_i\}$$

where $\mathcal{D}_i$ is the display at time i and x_i is the user's response. This is the history of queries and answers during the first t iterations.

We wish to compute and update the posterior distribution,

$$p_t(k) = P(Y = k|B_t), \quad k \in \mathcal{S},$$

the probability that image k is the target after t iterations. First, however, we must specify the joint distribution of Y and the observations $\{X_{\mathcal{D}_1}, ..., X_{\mathcal{D}_t}\}$. The posterior p_t is then computed in the usual way. As in previous work, we are going to assume the answers to the queries are conditionally independent given the target Y. This is not an unreasonable assumption in practice. It follows that

$$P(B_t|Y = k) = \prod_{i=1}^{t} P(X_{D_i} = x_i|Y = k).$$

The conditional response distribution, $P(X_{\mathcal{D}} = x|Y = k)$ is what we call the "answer model."

Updating the posterior is now easy:

$$\begin{aligned} p_{t+1}(k) &= P(Y = k|B_{t+1}) \\ &= P(Y = k|B_t, X_{D_{t+1}} = x_{t+1}) \\ &\propto p_t(k)P(X_{\mathcal{D}_{t+1}} = x_{t+1}|Y = k) \end{aligned}$$

In other words, updating $p_t(k)$ merely involves multiplying by the new evidence $P(X_{\mathcal{D}_{t+1}} = x_{t+1}|Y = k)$ and re-normalizing.

3 Answer Model

Designing $P(X_{\mathcal{D}} = x|Y = k)$ involves two primary decisions: determining the set of possible responses $x \in \mathcal{A}$ and capturing the behavior of a user who has image k in mind and is presented with the images in D and asked to respond. This specification inevitably relies on the metric in the signature space, denoted by d. More details about this metric is introduced in Section 5.

There are many possible choices for $\mathcal{A}$. In all cases, the target is identified if present, so let us assume that $Y \notin \mathcal{D}$. One could ask the user to supply a rather precise measure of the degree of similarity between each displayed image and his target. Somewhat less demanding, one could solicit a rough label for each displayed image, such as "relevant" or "not relevant". We have adopted a still simpler scheme in which the user is simply asked to select the image which is "closest" to his target. The price for simplicity is of course a decrease in the amount of information conveyed, and hence in the reduction of uncertainty about Y. Nonetheless, in our experiments, this model proved to be the most practical, both mathematically and in terms of user psychology. It does not unduly burden the user with complex decision-making, nor require any specific knowledge of image representation, and it provides a natural way of bringing the stored metrics into play. To make this precise, assume $\mathcal{D} = \{s_1, ..., s_n\}$ and set

$$\mathcal{A} = \{1, ...n, n + 1, ..., 2n\} \tag{1}$$

For $i \in \{1, ..., n\}$, the response $X_{\mathcal{D}} = i$ signifies that image s_i is not the target but, *in the opinion of the user*, is the one closest to his target. Response $i \in \{n + 1, ..., 2n\}$ signifies that image s_{i-n} is the target.

By definition of such comparative answer, if $k \in \mathcal{D}$, we have

$$P(X_{\mathcal{D}} = i|Y = k) = \begin{cases} 1 \text{ if } k = s_{i-n} \\ 0 \text{ otherwise} \end{cases}$$

Otherwise, i.e., if $k \notin \mathcal{D}$, then for $i \in \{1, ..., n\}$:

$$P(X_{\mathcal{D}} = i|Y = k) = \frac{\phi(d(s_i, k))}{\sum\limits_{s_j \in \mathcal{D}} \phi(d(s_j, k))} \tag{2}$$

Ideally, the closer the image $s_i \in \mathcal{D}$ is to k in the stored metric, the more likely the user is to choose it. Hence, we take ϕ to be monotonically decreasing. In our experiments, we adopt a parametric form for ϕ and learn the parameters from real data (collected user responses) by maximum likelihood estimation.

4 Display Model

One straightforward solution to determine $\mathcal{D}_t$, the n images displayed at iteration t, is to pick the n images which are most likely under the posterior distribution p_t. However, this simple strategy is far from optimal in terms of minimizing the average search time (our ultimate goal) except near the end of efficient searches, when p_t is highly concentrated. Instead, as in other work, we adopt the powerful (and time-independent) strategy of choosing $\mathcal{D}_{t+1}$ to minimize the uncertainty of the target given the search history B_t and new answer $X_{\mathcal{D}_{t+1}}$, measuring uncertainty by entropy:

$$\mathcal{D}_{t+1} = \arg\min_{\mathcal{D} \subset \mathcal{S}} H(Y|B_t, X_{\mathcal{D}}) \tag{3}$$

Since the entropy $H(Y|B_t)$ is independent of $\mathcal{D}$, Eqn.(3) is equivalent to maximizing the conditional mutual information between Y and $X_{\mathcal{D}}$ given B_t:

$$\mathcal{D}_{t+1} = \arg\max_{\mathcal{D} \subset \mathcal{S}} I(Y; X_{\mathcal{D}}|B_t) \tag{4}$$

The mutual information is then computed relative to the joint distribution determined by the answer model and the current posterior on the target.

4.1 Heuristic Solution

The minimization in Eqn.(3) is, unfortunately, a virtually intractable combinatorial optimization problem since there are $\binom{N}{n}$ choices for $\mathcal{D} \subset \mathcal{S}$. (Discarding images already displayed makes little difference.) The algorithm we use is based on an approximation to the corresponding optimization problem resulting from the choice of an ideal answer model under which the user selects the displayed image actually closest to his target using the system metric (or of course selects

the target itself if present). Since Y determines X_{D+1}, it is easy to see that Eqn. (3) is equivalent

$$\mathcal{D}_{t+1} = \arg\max_{\mathcal{D}\subset\mathcal{S}} H(X_{\mathcal{D}}|B_t) \tag{5}$$

However, there is a natural heuristic for this combinatorial optimization problem.

Roughly speaking, since entropy is maximized at the uniform distribution, and ignoring the case in which the target belongs to $\mathcal{D}$, we want to choose n images, call them $\{s_1, ..., s_n\}$, such that the Voronoi partition has cells of almost equal mass under the posterior. A sequential, heuristic solution is then given by the following algorithm:

1. The candidate set $\mathcal{C}_1$ for s_1 consists of all images not previously displayed through iteration t.
2. Select s_1 to be the image $k \in \mathcal{C}_1$ which maximizes $p_t(k)$.
3. Order the images in $\mathcal{C}_1$ according to their distance to s_1. Add these one-by-one to a cluster initialized by $\{s_1\}$ until the mass of the cluster under p_t reaches $\frac{1}{n}$.
4. Define the candidate set $\mathcal{C}_2$ for choosing s_2, by removing the cluster from $\mathcal{C}_1$.
5. Select s_2 to be the image $k \in \mathcal{C}_2$ which maximizes $p_t(k)$.
6. Divide all the images in $\mathcal{C}_1$ into two groups: those closest to s_1 and those closest to s_2. Order the distances in the first group (respectively, second group) according to their distance to s_1 (resp. s_2) and repeat the agglomeration procedure in step 3 relative to both s_1 and s_2. This results two clusters "centered around" s_1 and s_2, each with mass approximately $\frac{1}{n}$.
7. Continue in this way until $\{s_1, ..., s_n\}$ are chosen.

Although there is no guarantee to maximize entropy in Eqn (5), this heuristic solution is fast, simple and achieves good performance in practice.

5 Signatures, Metrics and Coherence

Given our emphasis on retrieving mental images of faces, it would seem natural to use signatures developed for face recognition and face retrieval with query-by-example. As a result, we have analyzed several subspace-based signatures applied in these areas, such as principle component analysis (PCA) [12], the kernel versions of Fisher's discriminant (KFDA) [13]. It should be emphasized however, that in face recognition and retrieval, the target image is available and hence its signature can be computed and directly compared with the signatures of other stored images. In particular, there is no guarantee that effective signatures for face recognition will also prove useful in mental retrieval.

We adopt the L_1 distance (Performance with L_2 is roughly the same) with normalization by size of database and order of value in database as our metric. One reason for the normalization is that standard signatures of the images in $\mathcal{S}$ are sparsely scattered in a high-dimensional Euclidean space and there is enormous variability among the distances between image pairs. Normalizing the distance using the order statistics ameliorates this problem.

Table 1. Face databases used in experiments

NAME	#Subjects	#Images	Composition
FERET(A)	1199	1199	All FERET images
FERET(C)	808	808	Caucasian subset
FERET(SB)	512	512	Semantically balanced subset
FERET(W)	327	327	Caucasian female subset
FERET(SB+F)	531	531	FERET(SB)+ 19 extra (familiar) faces

We investigated the coherence between mental matching and metric-based matching by collecting responses from various individuals. All the experiments in this paper utilize subsets of the FERET database. Since the majority of people in the FERET database are Caucasian, and since the response of most people is heavily biased by semantics, we used the FERET(SB) (see Table 1) in the coherence experiment. In FERET(SB), the distribution of ethnic (Asian, Black and Caucasian) and gender categories (female and male) is roughly uniform. Each data item consists of a triple $(Y, \mathcal{D}, X_{\mathcal{D}})$ corresponding to a target, set of displayed images and user response. The targets were sampled at random from S and the number of displayed images is set to $n = 8$. (Using many fewer or many more has adverse consequences with real users.) The answers are comparative, as described in Section 3. Nine individuals (in the INRIA labs) produced a total of 989 data items (records). Statistics were collected on the rank of the user's choice in terms of the L_1 distance between each display and the target. An example experiment is shown in Fig.1, which compares PCA and KFDA under the L_1 metric; both the density of rank and its cumulative distribution are shown. These two signatures perform about the same. Neither can be said to be highly coherent with mental matching as the probability that the user selects the closest image is only roughly 0.2. Nonetheless, reasonable search times are obtained; see Section 6. Similar results are observed in other signature spaces. Henceforth, we fix our distance to be the L_1 metric on the KFDA image representation.

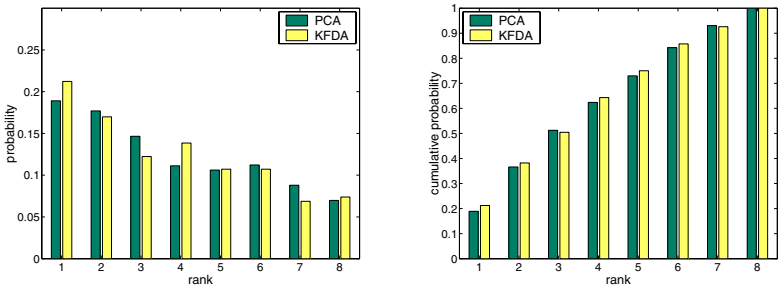


Fig. 1. Results comparing PCA and KFDA. Left: The estimated probability that the user selects the m'th closest image to his target according to the distance in signature space; Right: The cumulative distribution function

6 Experiments in Relevance Feedback

The web interface in the experiments is shown in Fig.2. Let T denote the number of iterations (query/response) until the target appears among the displayed images. Given M tests (full searches), we estimate $E(T)$, the mean of T, and $P(T \leq t)$, the (cumulative) distribution of T by their empirical statistics. That is, if the M tests results in search times $T_1, ..., T_M$, then $E(T) = \frac{1}{M} \sum_{m=1}^{M} T_m$ and $P(T \leq t) = \frac{\#\{1 \leq m \leq M | T_m \leq t\}}{M}$. Evidently, we seek small values of $E(T)$ and cumulative distributions $P(T \leq t)$ which climb as fast as possible.

Experiment I: Influence of the Answer Model
We designed answer models with varying degrees of "noise" in the sense of how well decisions cohere with the actual metric on signatures. For answer model defined in Eqn.(2), synchronization is controlled by the function $\phi(d)$ where $d = d(s_i, Y)$, the distance between the "i'th" displayed image and the actual target Y. The more rapidly $\phi(d)$ decreases (as d increases) the more likely is the user's answer to cohere with the signature metric. We did simulations with four answer models, meaning the answers are generated by sampling from the model. The response of the "ideal user" is always perfectly coherent with metric, i.e., $P(X_{\mathcal{D}} = i | Y = k) = 1$ if $d(s_i, k) < d(s_j, k)$ for all $s_i, s_j \in \mathcal{D}, i \neq j$. This represents the optimal performance obtainable. The other extreme is a random response ($\phi(d) \equiv const$); every displayed image is equally likely to be chosen regardless of its distance to the target. Two cases in between, and far more realistic, are $\phi(d) = \frac{1}{d}$ and $\phi(d) = 1 - d$; the former is evidently more coherent than the latter. One simulation on FERET(A) (see Table 1) with $M = 100$ is shown in Fig.3. In addition to the four (estimated) distribution function, the (estimated) mean search time is listed in the legend box. Clearly the degree of coherence with the metric on signatures characterizes the performance.

Experiment II: Sensitivity to the Size of the Database
To measure the effect of $N = |\mathcal{S}|$, we used databases of increasing size: FERET(W) ($N = 327$), FERET(SB) ($N = 512$), FERET(C) ($N = 808$) and FERET(A) ($N = 1199$)(see Table 1). The curve in Fig.4 shows the variation of $E(T)$ with N. The average search time grows slowly with N, roughly logarithmically.

Experiment III: Performance with Real Users
Tests with real users and a standard research database such as FERET is problematic since the user is not familiar with the people represented in the database. Of course one can select an image at random and ask the user to "memorize it" for few seconds, but this does not provide for a realistic scenario. Instead, we add images of the faces of familiar people to the database and select these as the targets for our experiments with mental image retrieval. The results shown in Fig.5 are based on $M = 78$ complete searches conducted by 22 INRIA researchers using the FERET(SB+F) database (see Table 1). For comparison, we show a simulation under the same experimental setting (i.e., same answer and display models) as well as the distribution corresponding to random display. In this case,

it is easy to see that the cumulative distribution is linear, $P(T \leq t) = \frac{t \times n}{N}$, and the $E(T) = \frac{T_{max}(1+T_{max})n}{2N}$, where T_{max} is the maximum number of iterations possible. The answer model uses a ϕ-function with the free parameters estimated by maximum likelihood. Obviously, the proposed model far out-performs a random response. More importantly, the absolute performance is quite reasonable, with a mean search time of $E(T) \approx 14.7$ iterations and target recovery in fewer than ten iterations in approximately one-half the searches. Fine-tuning the model, such as finding metrics and signatures more coherent with mental matching, would likely result in further improvements.

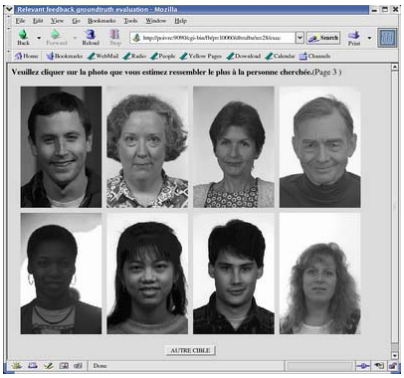


Fig. 2. The interface for experiments

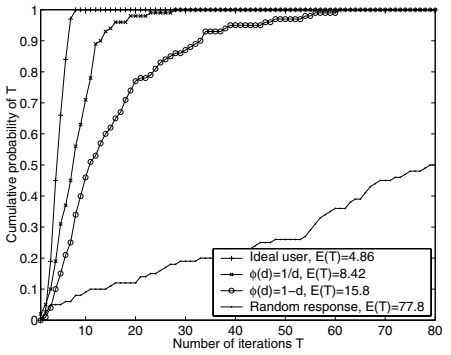


Fig. 3. The influence of the answer model

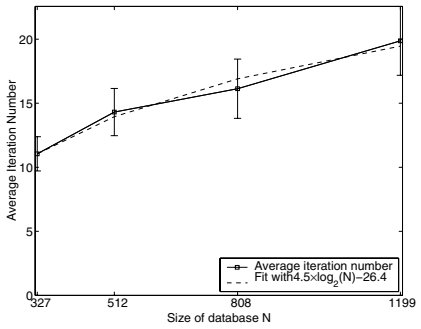


Fig. 4. The influence of N. The box plots are 95% confidence intervals

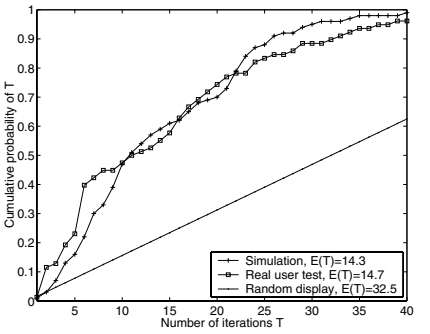


Fig. 5. Experiments in mental image retrieval with familiar faces as targets

7 Conclusions and Future Work

We have constructed a Bayesian model for mental face retrieval within the framework of relevance feedback. In deciding which faces to display to the user to match to the mental picture, a heuristic solution has been proposed based on the maximization of mutual information. The design of the answer model is

motivated by the need to account for the variability in the responses of actual users and the lack of a strong correlation between the basis for mental matching and how images are compared using standard metrics on standard image features. The performance of the system is validated in both simulations and in experiments with real user tests, which demonstrate the feasibility of the proposed model. Improvements are likely to result from metrics and features more adapted to human decision making. Some degree of semantic annotation would also increase efficiency, especially with much larger databases.

Acknowledgment

We are grateful to Ms. N. Boujemaa for helping us to develop the feedback model. We also acknowledge many useful suggestions from Mr. P. Welti of the SAGEM group and the aide of Mr. J-P Chieze in the design of the user inteface.

References

1. Boujemaa, N., Fauqeur, J., Ferecatu, M., and et al. Ikona: Interactive generic and specific image retrieval. In *Intern. Workshop MMCBIR'2001* (2001).
2. Boujemaa, N., Fauqeur, J., and Gouet, V. *What's beyond query by example?* Springer Verlag, 2004.
3. Bigun, J., Choy, K., and Olsson, H. Evidence on skill differences of women and men concerning face recognition. In *Proc. AVPBA* (2001), vol. 1, pp. 44–51.
4. O'Toole, A. J., Phillips, P. J., Cheng, and et al. Face Recognition Algorithms as Models of Human Face Processing. In *Proc. IEEE FG* (2000).
5. Liu, C., and Wechsler, H. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Trans. Image Processing 9* (2000), 132–137.
6. Swets, D. L., and Weng, J. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. PAMI 18*, 8 (1996), 831–836.
7. Navarrete, P., and Ruiz-del Solar, J. Interactive face retrieval using self-organizing maps. In *Proc. Int. Joint Conf. on Neural Networks* (2002).
8. Oda, M. Interactive search method for ambiguous target image. In *Proc. IEEE Intern. Workshop IDB-MMS'96* (1996), pp. 194–201.
9. Zhou, X. S., and Huang, T. S. Relevance feedback for image retrieval: a comprehensive review. *ACM Multimedia Systems Journal 8*, 6 (2003), 536–544.
10. Cox, I., Miller, M., Minka, T., Papathomas, T., and Yianilos, P. The bayesian image retrieval system, pichunter: theory, implementation and psychological experiments. *IEEE Trans. Image Processing 9* (2000), 20–37.
11. Geman, D., and Moquet, R. Q & a models for interactive search. Tech. rep., Dept. of Mathematics and Statistics, University of Massachusetts, 2001.
12. Turk, M., and Pentland, A. Eigenfaces for recognition. *Journal of Cognitive Neuroscience 3*, 1 (1991), 71–86.
13. Liu, Q., Huang, R., Lu, H., and Ma, S. Face recognition using kernel based fisher discriminant analysis. In *Proc. IEEE FG2002* (2002), pp. 197–201.

Fingerprint Image Segmentation Based on Quadric Surface Model*

Yilong Yin[1], Yanrong Wang[1], and Xiukun Yang[2]

[1] Computer Department, Shandong University, Jinan, 250100, China
ylyin@sdu.edu.cn
[2] Identix Incorporated, One Exchange Place, Jersey City, NJ 07302, USA
susan.yang@identix.com

Abstract. It is essential to segment fingerprint image from background effectively, which could improve image processing speed and fingerprint recognition accuracy. This paper proposes a novel fingerprint segmentation method at pixel level based on quadric surface model. Three parameters, *Coherence*, *Mean* and *Variance* of each pixel are extracted and spatial distribution model of fingerprint pixels is acquired and analyzed. Our study indicates that the performance of fingerprint image segmentation with a linear classifier is very limited. To deal with this problem, we develop a quadric surface formula for fingerprint image segmentation and acquire coefficients of the quadric surface formula using BP neural network trained on sample images. In order to evaluate the performance of our proposed method in comparison to linear classifiers, experiments are performed on public database "FVC2000 DB2". Experimental result indicates that the proposed model can reduce pixel misclassification rate to 0.53%, which is significantly better than the linear classifier's misclassification rate of 6.8%.

1 Introduction

In recent years, automatic fingerprint identification has always been a research focus in academic research and industry field of the world. With the efforts of many researchers, the main technical system of automatic fingerprint identification has been built already, and it has been applied in many fields. However, to meet the needs of applications, it's necessary to improve the performance of key algorithms, such as fingerprint image preprocessing, feature extraction and fingerprint matching algorithms, etc. We can say that the researches on fingerprint identification algorithms have developed into a procedure full of competitive.

Quality of fingerprint itself and image capture conditions directly influence the performance of the system, which makes fingerprint preprocessing be a necessary step. In addition, the separation of fingerprint image from background is the first step in preprocessing. A valid and effective segmentation could not only reduce time consumed on image preprocessing and minutiae extraction time but also improve the reliability of identification. So research on fingerprint segmentation has important meaning to the whole fingerprint identification system.

* Supported by the National Natural Science Foundation of China under Grant No. 60403010 and Shandong Province Science Foundation of China under Grant No. Z2004G05

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 647–655, 2005.
© Springer-Verlag Berlin Heidelberg 2005

There are many literatures about fingerprint image segmentation. Overall, the current approaches for fingerprint image segmentation can be classified into the following two categories. One method is based on the block level. B.M.Methre classifies each block into foreground or background according to the distribution of the gradients and gray-scale variance in that block [1,2]. X.Chen et al use linear classifier to classify the block [3]. L.R.Tang et al propose a fingerprint segmentation method based on D-S evidence theory [4]. It is carried out by the direction and contrast of block. S.Wang et al make use of contrast and main energy ratio to complete the separation of valid fingerprint part from background [5]. Q.Ren et al give the method based on feature statistics [6]. Obviously, the resolution of this method only reaches the block level, and the border of foreground and background obtained by this method is rather serrate, which makes it difficult to judge the reliability of features extracted from the border. The other method is based on pixel level. A.M.Bazen [7,8] et al use three features of pixel and establish a linear classifier to implement a pixel-based segmentation. Up to the present, prevailing method of fingerprint segmentation is still based on block. The number of method based on pixel is very little.

In this paper, a method based on quadric surface model for fingerprint image segmentation is presented. According to the spatial distributions of the pixel features in the foreground and background areas of database, it implements the segmentation of fingerprints using quadric surface model. In algorithm, we transfer the quadric discriminant to broad sense discriminant for reducing the computation complexity. Experiments have shown that the performance is excellent.

This paper is organized as follows. Section 2 studies the spatial distribution of pixels based on CMV. Section 3 presents development of the quadric surface model. Section 4 is the experimental results on fingerprint database and section 5 gives conclusions and some discussions about this presented method.

2 Spatial Distribution of Pixels Based on CMV

This section consists of two parts: one is description of pixel features and the other is spatial distribution of pixels based on above features.

2.1 Description of the Pixel Features

In brief, segmentation is the classification of pixels, so the first important problem of the fingerprint segmentation is to define proper parameters to describe the pixel features. The valid and distinguishable features of the pixel should play vital roles in the segmentation. For the pixels in the valid fingerprint foreground and background, the parameters should sufficiently represent their differences on these features. In this paper, *Coherence, Mean and Variance* are selected as the pixel features [8]. The definition of each feature is as following.

2.1.1 Coherence(C). The *Coherence* of the pixel measures how well the gradients are pointing in the same direction around a pixel, which is abbreviated as *Coh* [7,8,9]. Since a fingerprint mainly consists of parallel line structures, the coherence will be considerably higher in the foreground than in the background. In a window W around a pixel, it is defined as follows:

$$Coh = \frac{\left|\sum_W \left(G_{s,x}, G_{s,y}\right)\right|}{\sum_W \left|\left(G_{s,x}, G_{s,y}\right)\right|} = \frac{\sqrt{\left(G_{xx} - G_{yy}\right)^2 + 4G_{xy}^2}}{G_{xx} + G_{yy}}. \quad (1)$$

Where $\left(G_{s,x}, G_{s,y}\right)$ is the squared gradient, $G_{xx} = \sum_W G_x^2$, $G_{yy} = \sum_W G_y^2$, $G_{xy} = \sum_W G_x G_y$ and $\left(G_x, G_y\right)$ is the local gradient. W is the two-dimensional low-pass template sized of 17×17 used for noise reduction. From the formula, we can see that the coherence is among [0,1]. The coherence of pixel in foreground is near 1, while in background it is close to 0, because the directions of noise areas are in chaos.

2.1.2 Mean (M). The second pixel feature is the average gray value. It measures how gray the pixel is. For most fingerprint sensors, the ridge-valley structures can be approximated as black and white lines. So the foreground is composed of black and white lines, while the background, where the finger doesn't touch the sensor, is rather white. This means that the mean gray value in the foreground is lower than it is in the background. The local mean of the pixel is defined as follows:

$$Mean = \sum_W I. \quad (2)$$

Where I is the local intensity of the image. The definition of W is the same as above.

2.1.3 Variance (V). The *Variance* is the third pixel features used in the algorithm, which measures the gray variance around the local area. In general, the variance of the ridge-valley structures in the foreground is higher than the variance of the noise in the background. It is defined as follows:

$$Var = \sum_W \left(I - Mean\right)^2. \quad (3)$$

Where the definitions of I, W are the same as above too.

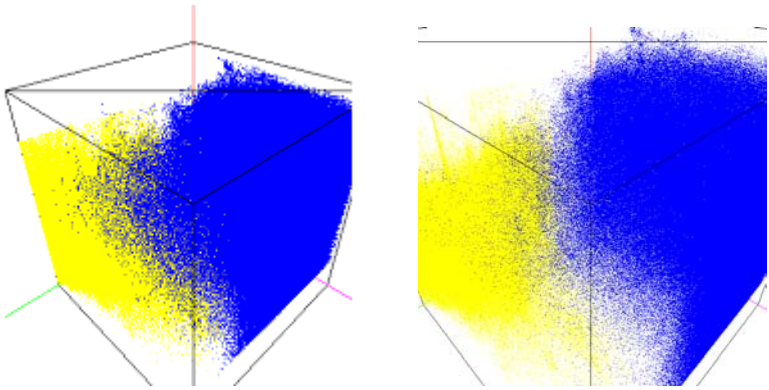
2.1.4 CMV Normalization. Since the value ranges of the three features are different, after extracting the three features of a pixel, we need to normalize them to the range of [0,1].

2.2 Spatial Distribution of Pixels Based on CMV

It is important to make deep insight into the spatial distribution based on the CMV of the pixel. According to this distribution, we may get the appropriate segmentation model by certain technique and achieve the satisfactory segmentation results.

Therefore, 60 representative fingerprint images from FVC2000 Database 1 have been selected as samples. Each pixel's CMV values of all the images are extracted and the spatial distributions of these pixels are shown in Fig. 1(a)(where the yellow section is composed of the pixels in the background, while the blue sections is composed of the pixels in the foreground. Which parts the pixels belongs to is decided

manually and the three coordinates of a point in the figure represented the CMV values of a pixel in the fingerprint image). Fig. 1(b) is obtained by extending the foreground and background pixels in Fig. 1(a) in order to see the overlapping section clearly.



(a) (b)

Fig. 1. Spatial distributions of pixels based on CMV (a)(*the yellow section is composed of the pixels in the background, while the blue section is composed of the pixels in the foreground*), (b)(*obtained by extending (a)*)

From the distributions of pixels, we can see that the pixels of different parts (background and foreground) have different clusters, which indicates that the pixels of the same class have similar features, and that two clusters are not apart completely.

Obviously, based on this spatial distribution, it is difficult to get satisfactory results only using the plane. In addition, linear classifier has its limitations and can't solve such nonlinear problems. All these factors make us select the quadric surface model.

3 Development of Quadric Surface Model

3.1 Selection of Segmentation Model

As we can see from the spatial distribution based on CMV of pixels, pixels from foreground and background are hard to separate with plane. if we select the quadric surface model, we may get better results. In general, suppose the quadric surface model is as follows:

$$g(x) = x^T W x + w^T x + w_0 = \sum_{k=1}^{d} w_{kk} x_k^2 + 2\sum_{j=1}^{d-1} \sum_{k=j+1}^{d} w_{jk} x_j x_k + \sum_{j=1}^{d} w_j x_k + w_0 \qquad (4)$$

W is a real symmetric matrix and w is a d -dimension vector. From (4), $L = \frac{1}{2} d(d+3) + 1$ coefficients are needed and the calculation cost is large. So in order to simplify the algorithm, we define W as a matrix that only elements on main

diagonal is non-zero, others are all zero. $x=[C,M,V]^T$, and $d=3$. Then formula (4) can be rewritten as following,

$$g(x)=x^TWx+w^Tx+w_0=\sum_{k=1}^{3}w_{kk}x_k^2+\sum_{j=1}^{3}w_jx_k+w_0. \tag{5}$$

Apparently, this quadric discriminant is nonlinear, which is difficult to handle with. So by increasing the dimension, the above can be translated into a broad sense linear discriminant. Under this theory, formula (5) can be translated to following,

$$\begin{aligned} g(x)=x^TWx+w^Tx+w_0&=\sum_{k=1}^{3}w_{kk}x_k^2+\sum_{j=1}^{3}w_jx_k+w_0 \\ &=\sum_{i=0}^{6}a_iy_i=a^Ty \end{aligned}. \tag{6}$$

Where,

$$y=[y_0,y_1,y_2,y_3,y_4,y_5,y_6]^T=[C^2,M^2,V^2,C,M,V,1]^T$$
$$a=[a_0,a_1,a_2,a_3,a_4,a_5,a_6]^T.$$

Then the rest problem is how to get the coefficients $a_i\ (0\le i\le 6)$ from a lot of samples.

3.2 Acquisition of Relative Coefficients

BP network algorithm is a relatively good feed-forward neural network algorithm in theory, and it is also widely used in practice. It is a supervised leaning algorithm used in many fields to simulate some very complicated nonlinear relationships when the conventional methods do not work. In this paper, we use the perception of BP learning algorithm to obtain the relative coefficients $a_i\ (0\le i\le 6)$.

Different fingerprint database need to establish different model coefficients [8], in this paper, 30 representative fingerprint images respectively from FVC2000 DB1, FVC2000 DB2 and FVC2000 DB3 are selected to get according model coefficients.

This neural network has 7 inputs and 1 output.

Inputs of the network: $y=[y_0,y_1,y_2,y_3,y_4,y_5,y_6]^T$,

Weight vector: $a=[a_0,a_1,a_2,a_3,a_4,a_5,a_6]^T$.

Transfer function:

$$f(g)=\frac{1}{1+e^{-g}}. \tag{7}$$

The structure of the network is shown in Fig. 2, where the real directional arrow denotes the direction of working signals, while the dashed directional arrow denotes the back error signals.

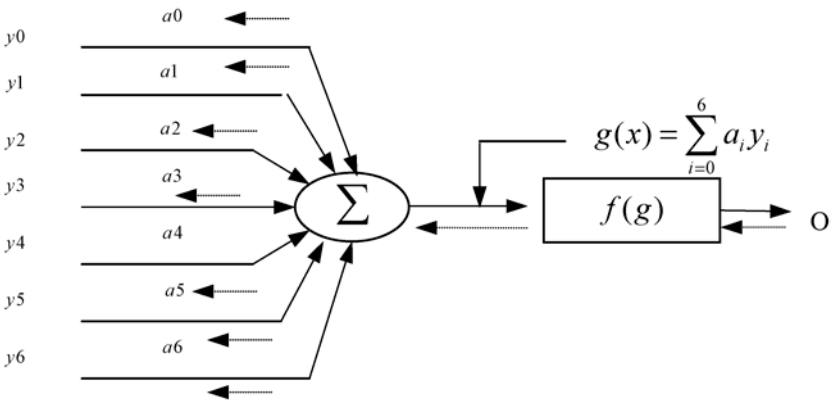


Fig. 2. The structure of network

The segmentation method presented in this paper uses supervised training. The classifications of the pixels (background or foreground) that are involved in the network training are selected manually.

The fingerprint image samples are the same as described above, which are representative enough.

3.3 Implementation of Segmentation

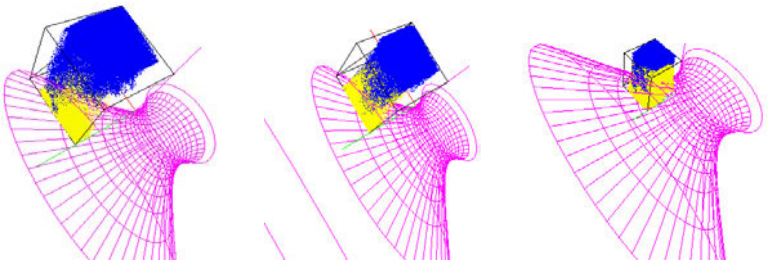
Since we have got the coefficients of the models, segmentation can be implemented using quadric surface model. Suppose $\hat{\omega}_0$ denotes the pixel that belongs to foreground, $\hat{\omega}_1$ denotes the pixel that belongs to background and $\hat{\omega}$ denotes the pixel's classification. The decision function used in the segmentation is:

$$\hat{\omega} = \begin{cases} \hat{\omega}_0 & if \quad g > 0 \\ \hat{\omega}_1 & if \quad g <= 0 \end{cases} . \tag{8}$$

Where the definition of g is in formula (6).

3.4 Visualization of Quadric Surface Model

Using the above neural network model to train three different sample databases, we obtain three different sets of coefficients, indicating that the optimal quadric surface models for fingerprint images acquired by different sensors are different. It also proves the theory proposed by A. M. Bazen regarding the necessity of training different fingerprint databases independently. The quadric surface models obtained from fingerprint training samples in database FVC2000 DB1, DB2 and DB3 are illustrated in Fig.3 (a), (b) and (c), respectively. Blue dots stand for foreground pixels, yellow dots represent background, and red surfaces denote segmentation surfaces.



(a) (b) (c)

Fig. 3. The quadric surface models (a)(*obtained from FVC2000 DB1*), (b)(*obtained from FVC2000 DB2*), (c)(*obtained from FVC2000 DB3*)

4 Experimental Results

Experiment 1: Evaluating the validity of the model for each fingerprint database. From each database (FVC DB1, DB2, DB3), select 300 images respectively and form testing sets (TestDB1, TestDB1, TestDB3) different from training sets. Now we are intended to evaluate if the model is validate for the training databases.

For each pixel, calculate CMV values and put them to formula (6), then calculate (7) to get $f(g)$, and make it among [0,1]. The probability density of Pixels from foreground and background of each database is given in Fig. 4. Make (0,1) to (0,10), and disperse it to 10 equal parts.

Define $Fingerf(x)\, x \in \{0,1,2,3,4,5,6,7,8,9\}$ as the probability of foreground pixels in $(x, x+1)$ and $Fingerb(x)\, x \in \{0,1,2,3,4,5,6,7,8,9\}$ as the probability of background pixels in $(x, x+1)$.

Obviously, $\sum_{x=0}^{9} Fingerf(x) = 1$ and $\sum_{x=0}^{9} Fingerb(x) = 1$. The point $x = 5$ is the threshold of differentiating pixels from foreground and background.

Experiment 2: Doing comparison experiments with A.M.Bazen's method on the second database of FVC2000. The parameters that measure the performance of the segmentation are defined as follows:

$p(\hat{\omega}_1 \mid \omega_0)$: The probability that a foreground pixel is classified as background.

$p(\hat{\omega}_0 \mid \omega_1)$: The probability that a background pixel is classified as foreground.

$\bar{p}_{error}$: The average of $p(\hat{\omega}_1 \mid \omega_0)$ and $p(\hat{\omega}_0 \mid \omega_1)$.

The experimental results for the Database 2 of FVC2000, using our method and the method presented in [8], are shown in Table 1.

Table 1. The segmentation results of our method and the method in[8] on FVC2000 DB2

Method	$p(\hat{\omega}_1 \mid \omega_0)$	$p(\hat{\omega}_0 \mid \omega_1)$	$\bar{p}_{error}$
Our method	0.21%	0.85%	0.53%
Method in [6]	6.2%	7.4%	6.8%

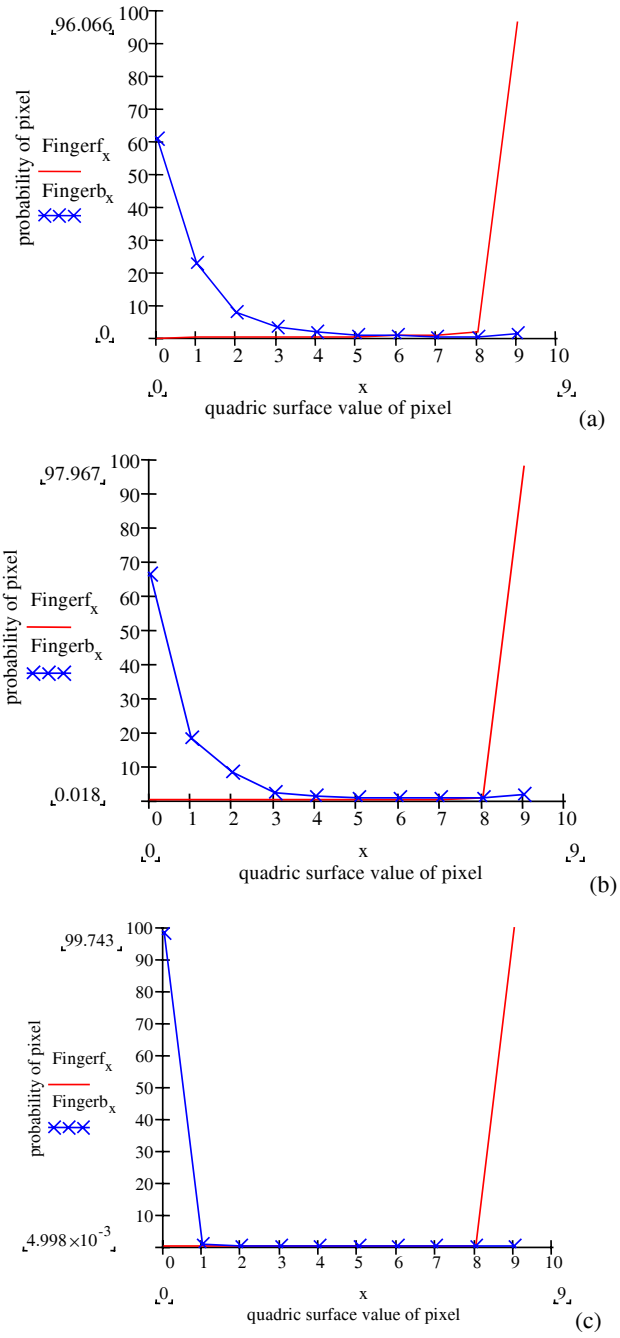


Fig. 4. The probability of pixels from foreground and background in (a)(*TestDB1*), (b)(*TestDB2*) and (c)(*TestDB3*)

5 Conclusion and Discussions

In this paper, a fingerprint segmentation algorithm using quadric surface model is proposed, which is based on the spatial distribution of the pixels derived from coherence, local mean and local variance. The spatial distribution of the pixels has special meaning for the segmentation. Experimental results on fingerprint images of FCV2000 database have demonstrated that the proposed segmentation method has excellent performance and the segmentation result is significantly better than that of the linear classifier.

The selection of curve surface model, and finding the optimal model to achieve more satisfactory segmentation result would be the next important step in fingerprint image segmentation study. Moreover, how to prove that the distribution of pixels based on CMV from foreground and background is a non-linear problem in theory and the reasonable feature selection of pixel are another worthwhile research.

References

1. B.M. Mehtre, N.N. Murthy, S.Kapoor, and B.Chatterjee. Segmentation of fingerprint images using the directional images, Pattern Recognition, 20(4): 429-435, 1987
2. B.M. Mehtre and B.Chatterjee. Segmentation of fingerprint images-a composite method", Pattern Recognition, 22(4): 381-385, 1989
3. Xinjian Chen, Jie Tian, Jiangang Cheng, Xin Yang. Segmentation of fingerprint images using linear classifier, EURASIP Journal on Applied Signal Processing, .2004(4): 480-494, 2004
4. L.R.Tang, X.H.Xie, A.N.Cai and J.A.Sun. Fingerprint Image Segmentation Based on D-S Evidence Theory, Chinese Journal of Computers, 26(7): 887-892, 2003(in Chinese)
5. S.Wang, W.W.Zhang and Y.S.Wang. New Features Extraction and Application in Fingerprint Segmentation. ACTA AUTOMATICA SINICA, 29(4): 622-626, 2003
6. Qun Ren, Jie Tian, Xiaopeng Zhang. Automatic segmentation of fingerprint images, the Third Workshop on Automatic Identification Advanced Technologies (AutoID 2002), Oral Report, Tarrytown, New York, USA, 137-141,2002
7. A.M.Bazen and S.H.Gerez. Directional field computation for fingerprints based on the principal component analysis of local gradients, Proc. ProRISC2000, 215-222, 11th Annual Workshop on Circuits, Systems and Signal Processing, Veldhoven, The Netherlands, Nov. 30-Dec. 1, 2000
8. A.M.Bazen and S.H.Gerez. Segmentation of Fingerprint Images, Proc. ProRISC2001, 276-280, 12th Annual Workshop on Circuits, Systems and Signal Processing, Veldhoven, The Netherlands, Nov. 29-30, 2001
9. A.M.Bazen and S.H.Gerez. Systematic methods for the computation of the directional field and singular points of fingerprints, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no.7, 905-919, 2002

A Fingerprint Matching Algorithm Based on Radial Structure and a Structure-Rewarding Scoring Strategy

Kyung Deok Yu, Sangsin Na, and Tae Young Choi

Department of Electrical and Computer Engineering,
Ajou University, Suwon, Korea
{rokmc100,sangna,taeyoung}@ajou.ac.kr

Abstract. This paper proposes a new fingerprint matching algorithm for locally deformed fingerprints based on a geometric structure of minutiae, called the radial structure, which is a collection of lines from a minutia that connect its Voronoi neighbors. The proposed algorithm consists of local matching followed by global matching, in both of which a new robust scoring strategy is employed. The local matching compares individual radial structures of a query and a template, and the global matching, performed when the local matching fails, utilizes overall radial structures of a query. The algorithm has been tested using the FVC2002 DB1 fingerprint database on a Pentium-4 personal computer with 1.8 GHz clock and 256 Mbyte RAM. The test results show that the average matching time including preprocessing is 0.9 sec, and the equal error rate is 8.22%. It has been observed that the proposed algorithm has a smaller equal error rate by 7.18% than Mital and Teoh's. This is a substantial improvement in the equal error rate on the angle-distance based algorithm of Mital and Teoh. This improvement is attributed to the following features of the proposed algorithm: the radial structure is obtained from Voronoi neighboring minutiae, which results in more robustness to false minutiae; and the scoring strategy rewards similarity in the geometric structure rather than feature types as in Mital and Teoh's algorithm.

1 Introduction

Fingerprint recognition is the most reliable and popular among various biometric recognitions [1]. A typical automatic fingerprint identification system consists of fingerprint acquisition, image preprocessing such as image enhancement and feature extraction, and matching. Fingerprint matching algorithms can be roughly classified into the following three categories: minutia-based [2, 3]; ridge-based [4]; and hybrid methods [5]. Among these categories, the minutia-based algorithms are the most common because they usually perform better in recognition accuracy and the processing time than the others.

Mital and Teoh proposed a minutiae-based matching algorithm in [6], which uses five minutiae of its closest neighbors to form geometric structure along with a scoring method. Their algorithm is simple, very effective, and rotationally

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 656–664, 2005.
© Springer-Verlag Berlin Heidelberg 2005

invariant. However, it has the following two weak points. First, their algorithm becomes sensitive to false minutiae because a false minutia is chosen, if it is closer to true ones, for the structure construction. Second, their scoring method penalizes too heavily discrepancy in the minutiae type between a query and a template even though the geometric structures are the same or similar.

In this regard this paper proposes a matching algorithm that improves on Mital and Teoh's. The proposed algorithm constructs a geometric structure of a minutia, called the radial structure, which consists of lines connecting its Voronoi neighbors. In addition, the algorithm employs a scoring strategy that rewards similarity in the geometric structure of a minutia, even when the feature types are different unlike Mital and Teoh's. Consequently, the combination of the radial structure and the structure-rewarding scoring makes the proposed algorithm robust to false minutiae.

To compare the proposed algorithm with the referenced algorithm of Mital and Teoh's, we have performed the cross experiment of four types using FVC2002 DB1 fingerprint database. The proposed algorithm has a higher recognition rate by 7.18%: in the recognition rate, the proposed matching algorithm outperforms the reference by 2.61% while the proposed scoring method the reference by 5.54%. From the experimental results, it has been found that the proposed matching algorithm and scoring method are more robust to false minutiae than the counterparts of the referenced algorithm. This improvement is attributed to the following features of the proposed algorithm: the radial structure is obtained from Voronoi neighboring minutiae, which results in more robustness to false minutiae; and the scoring strategy rewards similarity in the geometric structure rather than feature types as in Mital and Teoh's.

The rest of the paper is organized as follows: Section 2 deals with Voronoi diagrams and the definition of the radial structure; Section 3 describes the proposed matching algorithm and the scoring method. Finally, Section 4 and 5 present experimental results and conclusions, respectively.

2 Background

2.1 The Definitions of the Voronoi Diagram and the Radial Structure

The Voronoi Diagram. In the plane the Euclidean distance between two points p and q by $dist(p, q)$ is defined:

$$dist(p, q) := \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2}.$$

$P := \{p_1, p_2, p_3, ..., p_n\}$ is a set of n distinct points in the plane, each of these points is called a site. We define the Voronoi diagram of P as the subdivision of the plane into n cells, one for each site in P, with the property that a point q lies in the cell corresponding to a site p_i if and only if $dist(q, p_i) < dist(q, p_j)$ for each $p_j \in P$ with $i \neq j$. We denote the Voronoi diagram of P by $Vor(P)$. The cell that corresponds to a site p_i is denoted $V(p_i)$, which we call the Voronoi cell

of p_i. The follows represent the properties of Voronoi diagram. Let P be a set of n point sites in the plane. First, if all the sites are collinear, then $Vor(P)$ consists of $n-1$ parallel lines and n cells. Second, the number of vertices in $Vor(P)$ is at most $2n-5$ and the number of edges is at most $3n-6$. Third, a point q is a vertex of $Vor(P)$ if and only if its largest empty circle $C_p(q)$ contains three or more sites on its boundary [7]. Figure 1 shows an example of the Voronoi diagram on a fingerprint image after preprocessing.

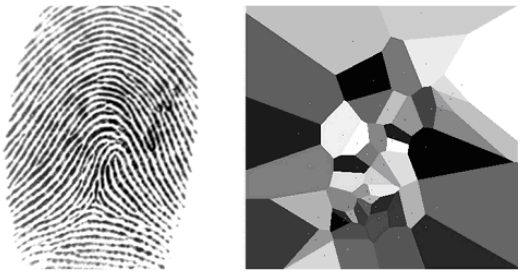


Fig. 1. An example of the Voronoi diagram of a fingerprint

In this paper we use a modified plane sweep algorithm to construct Voronoi diagram–Fortune's algorithm. This algorithm is chosen because it has the $O(n \log n)$ complexity while the half plane intersection algorithm has the complexity of $O(n^2 \log n)$ [7]. To introduce the sweep line algorithm, we consider the several definitions. Parabolas are useful in this sweep line algorithm because for any point p_i, there is a parabola from the sweep line that every point on the parabola is equidistant from both p and the sweep line. Now, we explain the site event and circle event. As the sweep process, a new arc of some parabola is added to wave-front (beach line) only when sweep line touches the some site. This is called a site event. And, the only way that an arc can disappear from the wave-front is when two other adjacent arcs intersect it at a common point. This is called a circle event [8]. Figure 2 shows the breakpoint, wave-front (beach line) and sweep line, and Figure 3 shows the site event and circle event.

The Radial Structure. The radial structure is defined as follows. Figure 4 shows an example of the radial structure. A set $P := \{p_1, ..., p_n\}$ for $n \in Z$ is a fingerprint image consisting of n minutia points. The radial structure of point $p_i \in P$, denoted $R(p_i)$, is defined to the set of all neighborhood minutiae sharing the edge with minutia p_i. We call the center minutia of $R(p_i)$ as c_i, and the neighborhood minutiae of $R(p_i)$ as n_i out of order. So, an arbitrary $R(p_i)$ is defined as follows $R(p_i) := \{c_i, n_1, n_2, ..., n_j\}$ for $2 \leq j \leq n$. Figure 5 shows that a radial structure consisting of center minutia and its neighbors in a fingerprint image after Voronoi diagram is constructed. The minutiae are extracted in the fingerprint image and the radial structures are formed for each $R(p_i)$ is saved

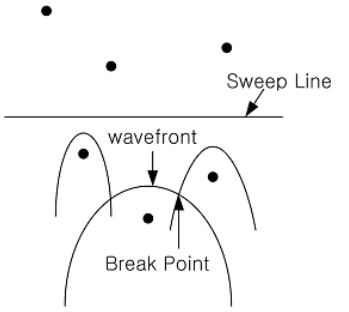


Fig. 2. The sweep line algorithm

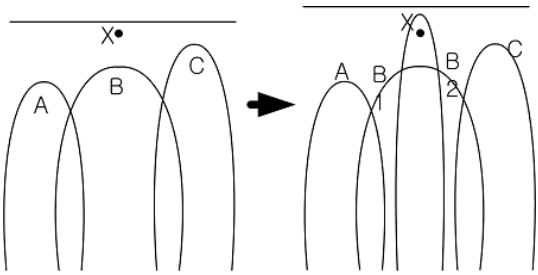


Fig. 3. The site event and circle event

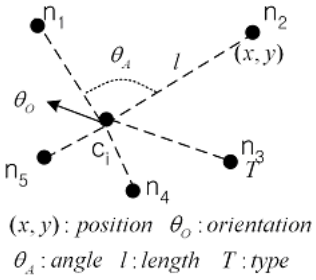


Fig. 4. The correlation factor of the radial structure

with a text file with a form of following set to perform the matching stage.

$$R(p_i) = \{(c_{ix}, c_{iy}, c_{iO}, c_{iT}),\ (n_{1x}, n_{1y}, n_{1\theta}, n_{iO}, n_{1T}), ..., (n_{jx}, n_{jy}, n_{j\theta}, n_{iO}, n_{jT})\}$$

3 The Proposed Matching Algorithm

The proposed matching algorithm proceeds in the following two stages.

Stage 1: Search for the number C_N of the radial structures which have scores higher than a preset threshold T_S by comparing radial structures of a query and a template fingerprint image. If $C_N \geq T_N$, we do not carry out second stage

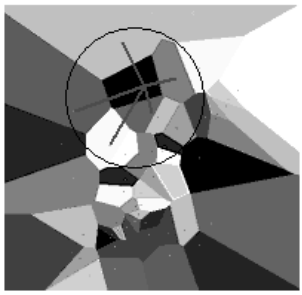


Fig. 5. A radial structure

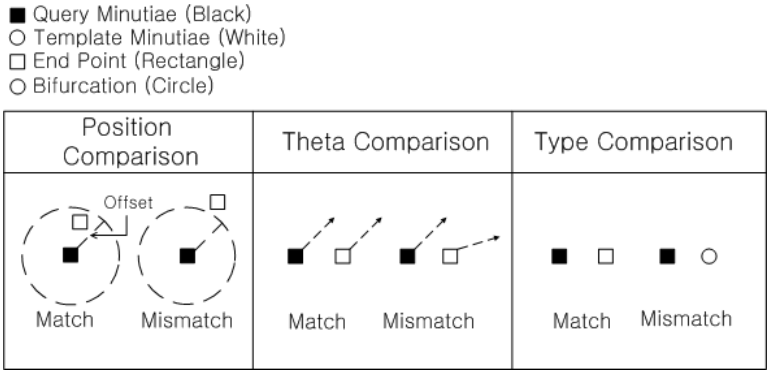


Fig. 6. The comparison criteria of the radial structure

matching and display the same fingerprint image. These matching techniques are useful to locally deformed fingerprint image. If first stage matching is not employed, because total matching score is low, miss-match can occur in spite of same fingerprint image. The basis about similarity of radial structure is shown with Figure 6. Threshold (T_S) is predefined by experimental result to be repeated, and threshold number (T_N) is defined following threshold:

$$T_N = Total\ number\ of\ radial\ structure \times \frac{25}{100}$$

Stage 2: If the number of similar radial structure is less than T_N, the second stage is performed. At the second stage, estimate the transformation parameters by using three radial structures that scores highest at the first stage. The following pseudo code shows a method of extracting the translation and rotation parameters between a query and a template. After extracting parameters, similarity of two fingerprint images is decided on by translating and rotating the query along the extracting parameters. Figure 7 is the flow chart for the process.

```
Algorithm for extracting_parameters(query radial structure Qhi,
template radial structure Thi)
  while i less than 3
     do coincide Qhi's center point and Thi's center point
        calculate the translation by subtracting Qhi from Thi
        while Qhi's neighbor is not empty
           do rotate T as Qhi's angle and score each stage
              each score is stacked the temporary memory
           do search the highest score from the temporary memory
              calculate the angle as sum of rotation angle of
              Qhi's neighborhood
        end of while
  end of while
end of Algorithm
```

In Figure 7, C_N is the number of the radial structures which has a higher score than threshold T_S between a query and a template, and S denotes the total matching score.

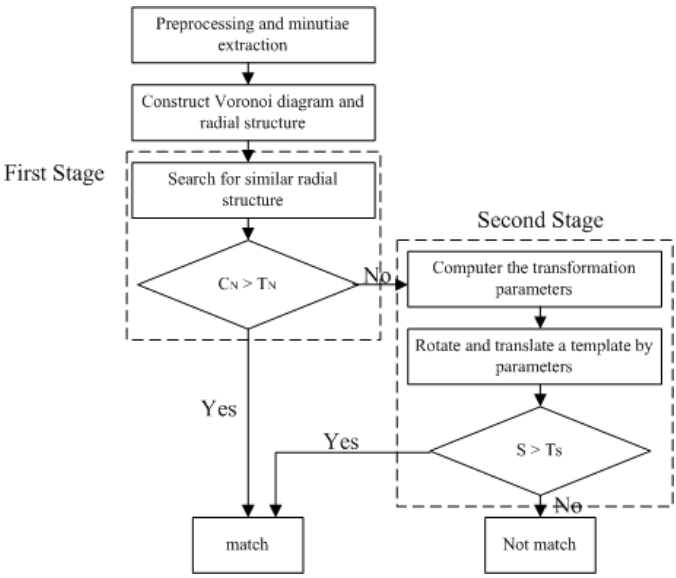


Fig. 7. The proposed matching algorithm

4 Experimental Results

The proposed matching algorithm has been tested with the FVC2002 DB1 on a Pentium-4 personal computer with 1.8 GHz clock and 256 Mbyte RAM. The FVC2002 DB1 contains 800 impressions obtained from an optical sensor, with 8

Table 1. The experimental results of the matching rate (%)

Matching ＼ Scoring	Proposed	Mital and Teoh's
Proposed	91.78	86.24
Mital and Teoh's	89.17	84.60

impressions apiece from 100 fingers [9]. We carried out the experiment categorized in Table 1 in order to compare the proposed algorithm and scoring method with Mital and Teoh's in [6]. Mital and Teoh's algorithm uses a local feature group, in which each of the extracted features is correlated with its five nearest neighboring features to form a local feature group for a first stage matching. Their algorithm is more sensitive than the proposed algorithm to a false minutia, as an example in Figure 8 shows.

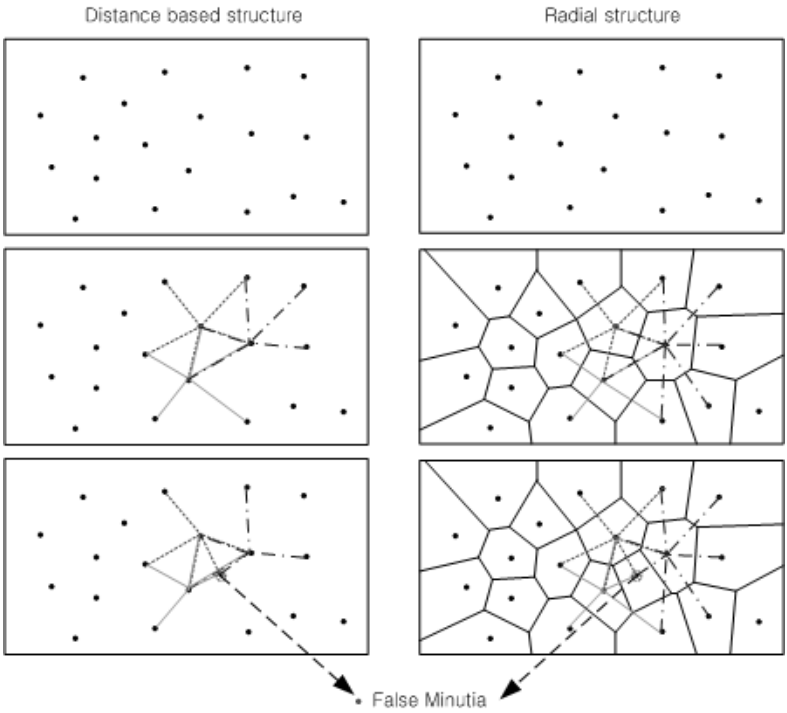


Fig. 8. The deformation of each structure when false minutiae is inserted

Each category in Table 1 was performed 2,800 times for genuineness, and 4,950 times for imposter. Figure 9 shows the distribution of the obtained matching score and Table 1 the matching rate. Table 1 clearly shows that the proposed algorithm outperforms Mital and Teoh's.

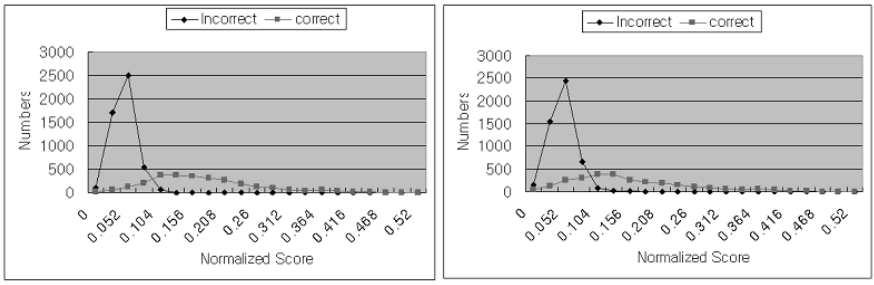


Fig. 9. The distribution of the matching score of the proposed (left) and Mital and Teoh's (right) algorithm

The proposed and Mital and Teoh's algorithms are similar in the sense that they both use a geometric structure of fingerprint minutiae, but they are different in the way of constructing the geometric structures and also in the scoring method. First, the proposed algorithm constructs the radial structure using Voronoi diagrams. Second, in the scoring method, the proposed algorithm gives priority to the geometric structure rather than minutia types such as ridge ending and bifurcation. Third, if a false minutia is inserted in the same geometric structure, the proposed algorithm usually undergoes less deformation than Mital and Teoh's, although they suffer equally in the worst case.

5 Conclusions

This paper has proposed a fingerprint matching algorithm based on the radial structures of minutiae and a scoring method that rewards the geometric structure. Numerical results have shown that the proposed algorithm has a substantially improved equal error rate on an angle-distance based algorithm. The proposed matching algorithm and scoring method are found to be more robust to the false minutiae because of the radial structure and structure-rewarding scoring method. The proposed matching algorithm can be extended for higher recognition accuracy. One extension may involve a combined radial structure, which is a collection of radial structures of neighboring minutiae connected to a principal minutia.

Acknowledgement

This work was supported in part by the Biometrics Engineering Research Center, KOSEF.

References

1. Biometric Market Report (International Biometric Group)
2. A. K. Jain, L. Hong and R. Bolle.: On-Line Fingerprint Verification. IEEE Trans. PAMI, Vol. 19, no. 4, pp. 302–313, 1997.

3. N. K. Ratha, K. Karu, S. Chen, and A. K. Jain.: A Real-Time Matching System for Large Fingerprint Database. IEEE Trans. PAMI, Vol. 18, no. 8, pp. 799–813, 1996.
4. Y. T. Park.: Robust Fingerprint Verification by Selective Ridge Matching. The Institute of Electronics Engineers of Korea, no. 37, pp. 351–358, 2000.
5. A. K. Jain, A. Ross and S. Probhakar.: Fingerprint Matching Using Minutiae and Texture Features. ICIP, pp. 282–285, Thessaloniki, Greece, 2001.
6. D. P. Mital and E. K. Teoh.: An Automated Matching Technique for Fingerprint Identification. In Proceedings of the 1997 First International Conference on Knowledge-Based Intelligent Electronic Systems. KES '97, Vol. 1, pp. 142–147, May 21-23, 1997.
7. M. de Berg, M. van Kreveld, M. Overmars and O. Schwarzkopf.: Computational Geometry: Algorithms and Applications. pp. 145–161, Springer, 1997.
8. M. Horn and J. Weber.: Computational Geometry Lecture Notes: Voronoi Diagrams. April 29, 2004.
9. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman and A. K. Jain.: FVC2002: Second Fingerprint Verification Competition.

A Novel Algorithm for Distorted Fingerprint Matching Based on Fuzzy Features Match*

Xinjian Chen, Jie Tian**, and Xin Yang

Center for Biometrics and Security Research, Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation Chinese Academy of Science, Graduate School of the Chinese Academy of Science, P.O.Box 2728, Beijing 100080, China
tian@doctor.com, jie.tian@mail.ia.ac.cn
http://www.fingerpass.net

Abstract. Coping with non-linear distortions in fingerprint matching is a real challenging task. This paper proposed a novel method, fuzzy features match (FFM), to match the deformed fingerprints. The fingerprint was represented by the fuzzy features: local triangle features set. The similarity between fuzzy features is used to character the similarity between fingerprints. First, a fuzzy similarity measure for two triangles was introduced. Second, the result is extended to construct a similarity vector which includes the triangle-level similarities for all triangles in two fingerprints. Accordingly, a similarity vector pair is defined to illustrate the similarities between two fingerprints. Finally, the FFM measure maps a similarity vector pair to a scalar quantity, within the real interval [0, 1], which quantifies the overall image to image similarity. To validate the method, fingerprints of FVC2004 were evaluated with the proposed algorithm. Experimental results show that FFM is a reliable and effective algorithm for fingerprint matching with non-liner distortions.

1 Introduction

How to cope with these non-linear distortions in the matching process is a very challenging task. According to Fingerprint Verification Competition 2004 (FVC2004) [1], they are particularly insisted on: distortion, dry and wet fingerprints. Distortion of fingerprints seriously affects the accuracy of matching.

Recently, many researchers are dedicated to solve the problem to matching of non-linear distorted fingerprints. Ratha *et al.* [2] proposed a method to measure the forces and torques on the scanner directly. Which prevent capture when excessive force is applied to the scanner. Dorai et al. [3] proposed a method to detect and estimate distortion occurring in fingerprint videos. But the above two methods do not work with the collected fingerprint images. Maio and Maltoni et al. [4] proposed a plastic distor-

* This paper is supported by the Project of National Science Fund for Distinguished Young Scholars of China under Grant No. 60225008, the Key Project of National Natural Science Foundation of China under Grant No. 60332010, the Project for Young Scientists' Fund of National Natural Science Foundation of China under Grant No.60303022, and the Project of Natural Science Foundation of Beijing under Grant No.4052026

** Corresponding author: Jie Tian ; Telephone: 8610-62532105; Fax: 8610-62527995

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 665–673, 2005.
© Springer-Verlag Berlin Heidelberg 2005

tion model to cope with the nonlinear deformations characterizing fingerprint images taken from on-line acquisition sensors. This model helps to understand the distortion process. Dongjae Lee et al. [5] addressed a minutiae-based fingerprints matching algorithm using distance normalization and local alignment to deal with the problem of the non-linear distortion. Senior et al. [6] proposed a method to convert a distorted fingerprint image into an equally ridge spaced fingerprint before matching to improve the matching accuracy. However, the assumption of equal ridge spacing is less likely to be true for fingerprints. C.I.Watson et al. [7] proposed a method to improve performance of fingerprint correlation matching using distortion tolerant filters. Kov´acs-Vajna et al. [8] also proposed a method based on triangular matching to cope with the strong deformation of fingerprint images, demonstrating graphically that the large cumulative effects can result from small local distortions. Bazen et al. [9] employed a thin-plate spline model to describe the non-linear distortions between the two sets of possible matching minutiae pairs. Ross et al [10] used the average deformation computed from fingerprint impressions originating from the same finger based on thin plate spline model to cope with the non-linear distortions.

Different from the above mentioned methods, we propose a novel method, FFM, to match the deformed fingerprints. The fingerprint was represented by the fuzzy features: local triangle features set. The similarity between fingerprints becomes an issue of finding similarities between fuzzy features. A fuzzy similarity measure for two triangles was firstly introduced. The result is then extended to construct a similarity vector which includes the triangle-level similarities for all triangles in two fingerprints. Accordingly, a similarity vector pair is defined to illustrate the similarities between two fingerprints. Finally, the FFM measure maps a similarity vector pair to a scalar quantity, within the real interval [0, 1], which quantifies the overall image to image similarity. Experimental results indicate that the proposed algorithm works well with the non-linear distortions. For deformed fingerprints, the algorithm gives considerably higher matching scores compared to conventional matching algorithms. In fingerprints database DB1 and DB3 of FVC2004, the distortion between some fingerprints from the same finger is large. But our algorithm performs well. The equal error rates (EER) are 4.06% and 1.35% on DB1 and DB3 respectively.

This paper is organized as follows. Section 2 proposes the details of the novel method, FFM, which is applied to match the deformed fingerprints. The performance of the proposed algorithm is shown by experiments in Section 3. Section 4 concludes our work.

2 Fuzzy Features Representation and Fuzzy Match

Matching of non-linear distorted fingerprints is a difficult issue. In some traditional methods, most of algorithms increase the size of the bounding boxes in order to tolerate further apart of matched minutiae pairs because of plastic distortions, and therefore to decrease the false rejection rate (FRR). However, as a side effect, this method may lead higher false acceptance rate (FAR) by making non-matching minutiae get paired. Figure 1 shows a pair of fingerprints of large distortion from FVC2004 DB1 (102_3.tif and 102_5.tif). While the corresponding minutiae in blue rectangle region are approximately overlapped, the maximal vertical difference of corresponding minutiae in red ellipsesregion is above 100 pixels.



(a) (b) (c)

Fig. 1. The example of large distortion from FVC2004 DB1. (a) 102_3.tif, (b) 102_5.tif, (c) the image is fingerprint 102_5 (after registration) added to 102_3. In blue rectangle region, the corresponding minutiae are approximately overlapped. While in red ellipse region, the maximal vertical difference of corresponding minutiae is above 100 pixels

Obviously, it is not feasible to compute the similarity between the deformed fingerprints by only using the global structure of the fingerprints. We proposed a novel method, FFM based on local triangle features set, to match the deformed fingerprints. The fingerprint was represented by the fuzzy features set: local triangle features set. The similarity between fuzzy features set is used to character the similarity between fingerprints.

2.1 Defining of Local Triangle Features and Feature Representation

In the proposed algorithm, the local triangle structure of the fingerprints was constructed the basis of the matching. The feature vector of a local triangle structure *Tk* is given by: $FTk = \{dij, dik, djk, \psi i, \psi j, \psi k, OZi, OZj, OZk, \alpha_i, \alpha_j, \alpha_k\}$. Where, *dij* denotes the distance between minutiae *i* and *j;* ψi denotes the angle between the direction from minutiae *i* to *j* and the direction from minutiae *i* to *k; OZi* denotes the orientation differences within the region of minutiae *i;* α_i denotes the angle between the orientation of minutiae *i* with the direction of the interior angle bisector of corner *i.* The pixels in the square centered around minutiae i within the radius *r* form the region of the minutiae i. It is clear that the local triangle structure feature vector *FTk* is independent from the rotation and translation of the fingerprint. Figure 2 shows a local triangle structure of the fingerprint.

In the process of construct the local triangles, as the fingerprint is deformed, the distance between two minutiae (vertexes of the triangle) should not be too long. It can have larger deformation as the accumulation of deformation from all the regions between minutiae [8].

The feature vector set F, which consists of feature vectors FTk, k=1, 2, …N, of all local triangles detected from a fingerprint, is used to represent the fingerprint. Fingerprint matching is to find a similarity between two feature vector set, one from the template fingerprint and another from the input fingerprint.

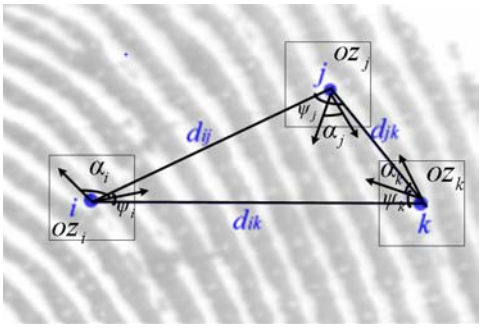


Fig. 2. The local triangle structure of the fingerprint

2.2 Learning Genuine Distorted Pattern Parameters Space

Before measuring the similarity between fuzzy features, we give out the defining of genuine distorted pattern parameters space. The genuine distorted pattern parameters space is derived from a set of genuine matching attempts.

Suppose $FTk = \{dij, dik, djk, \psi i, \psi j, \psi k, OZi, OZj, OZk, \alpha_i, \alpha_j, \alpha_k\}$ is a local triangle feature in template fingerprint and $FTk_I = \{dij', dik', djk', \psi i', \psi j', \psi k', OZi', OZj', OZk', \alpha_i', \alpha_j', \alpha_k'\}$ is also a local triangle feature in input fingerprint. Four distorted pattern parameters len_{diff}, ψ_{diff}, OZ_{diff}, α_{diff} are calculated as following:

$$len_{diff} = \frac{\left\|(dij - dij')\right| + \left|(dik - dik')\right| + \left|(djk - djk')\right\|}{3} \tag{1}$$

$$\psi_{diff} = \frac{\left\|\psi i - \psi i'\right| + \left|\psi j - \psi j'\right| + \left|\psi k - \psi k'\right\|}{3} \tag{2}$$

$$OZ_{diff} = \frac{\left\|OZi - OZi'\right| + \left|OZj - OZj'\right| + \left|OZk - OZk'\right\|}{3} \tag{3}$$

$$\alpha_{diff} = \frac{\left\|\alpha_i - \alpha_i'\right| + \left|\alpha_j - \alpha_j'\right| + \left|\alpha_k - \alpha_k'\right\|}{3} \tag{4}$$

These distorted pattern parameters formed the deformed pattern feature vector $f(len_{diff}, \psi_{diff}, OZ_{diff}, \alpha_{diff})$. To learn the genuine distorted pattern parameters, we used a set of distorted fingerprint images to derive a genuine distorted pattern parameter space. This set of images was extracted from database FVC2004 DB1 set B. The fingerprint database set B contains 80 fingerprint images captured from 10 different fingers, 8 images for each finger.

We matched those images from same finger one to one and computed the distorted pattern parameters. These distorted patterns parameters formed our genuine distorted pattern parameters space. Figure 3 shows a genuine distorted pattern derived from two images in FVC2004 DB1 Set B.



(a) template image (b) input image (c) registration pattern

Fig. 3. A genuine distorted pattern derived from a true matching attempt

2.3 Fuzzy Feature Match

Based upon fuzzy feature representation of fingerprints, the similarities between fuzzy features character the similarity between fingerprints.

2.3.1 Fuzzy Similarity Between Triangles

All of elements in genuine distorted pattern parameters space construct fuzzy feature set D. The center (d) of the fuzzy feature set D is used to represent feature set D with d defined as

$$d = \frac{\sum_{f \in \widetilde{D}} f}{V(\widetilde{D})} \tag{5}$$

which is essentially the mean of all elements of feature set, and may not be an element of feature set. While averaging over all features in a feature set increases the robustness of fuzzy feature, at the same time, lots of useful information is also submerged in the smoothing process because a set of feature vectors are mapped to a single feature vector.

Suppose $FTk = \{dij, dik, djk, \psi i, \psi j, \psi k, OZi, OZj, OZk, \alpha_i, \alpha_j, \alpha_k\}$ is a local triangle in template fingerprint and $FTk_I = \{dij^{'}, dik^{'}, djk^{'}, \psi i^{'}, \psi j^{'}, \psi k^{'}, OZi^{'}, OZj^{'}, OZk^{'}, \alpha_i^{'}, \alpha_j^{'}, \alpha_k^{'}\}$ is a local triangle in input fingerprint. The following method is used to measure the similarity between FTK_T and FTK_I. First, calculate the deformed pattern feature vector $f(len_{diff}, \psi_{diff}, OZ_{diff}, \alpha_{diff})$; then measure the degree of membership of f to the fuzzy feature set D.

In the proposed algorithm, the modified form of Cauchy function is chosen as membership function due to its good expressiveness and high-computational efficiency [11] [12].

The membership function: $C : \widetilde{D} \rightarrow [0, 1]$, is defined as:

$$C(f)=\begin{cases}1 & \text{if g(f, d)} = True \\ \dfrac{1}{1+(\dfrac{\|f-d\|}{m})^{a}} & otherwise\end{cases} \tag{6}$$

Where $f \in \widetilde{D}$, m and a$\in \Re$, m>0, a$\geq$0. g(f, d) = $True$, if and only if the value of each entry in feature vector f is less than the valve of corresponding entry in feature vector d. d is the center location of the function, *m* represents the width ($\|f-d\|$ *for* $C(f)=0.5$) of the function, and *a* determines the shape of the function. Generally, *m* and *a* portray the grade of fuzziness of the corresponding fuzzy feature.

It is clear that the farther a feature vector moves away from the cluster center, the lower its degree of membership to the fuzzy feature.

2.3.2 Fuzzy Feature Matching: Similarity Between Fingerprints

It is clear that it is needed to construct the image-level similarity from triangle-level similarities.

Let $T = \{FTt : 1 <= t <= n$, n is the number of all triangles detected from the template fingerprint } represent the template fingerprint, and $I = \{FTi : 1 <= i <= m$, m is the number of all triangles detected from the input fingerprint } represent the template fingerprint. First, for every $FTt \in$T, we define the similarity measure for it and **I** as

$$l_t^I = \max\{C(FTt \text{ - } FTi) \,|\, i = 1..m\} \tag{7}$$

Combining l_t^I together, we get a vector $l^I = [l_1^I, l_2^I, \quad l_n^I]^T$.

Similarly, for every $FTi \in I$, we define the similarity measure for it and ***T*** as

$$l_i^T = \max\{C(FTi \text{ - } FTt) \,|\, t = 1..n\} \tag{8}$$

Combining l_i^T together, we get a vector $l^T = [l_1^T, l_2^T, \quad l_m^T]^T$.

It is clear that l^I describes the similarity between individual fuzzy features in ***T*** and all fuzzy features in ***I***. Likewise, l^T illustrates the similarity between individual fuzzy features in ***I*** and all fuzzy features in ***T***. Thus, we define a similarity vector for ***T*** and ***I,*** denoted by $L^{(T,I)}$, as $L^{(T,I)} = \begin{bmatrix} l^I \\ l^T \end{bmatrix}$, which is a *n*+*m* dimensional vector with values of all entries within the real interval [0,1].

FFM measure is proposed to provide an overall image to image similarity by summation all the weighted entries of similarity vectors $L^{(T,I)}$. In the FFM measure, both area and center favored schemes are used. The weight vectors w are defined as

$$w = (1 - \alpha)w_A + \alpha w_B \tag{9}$$

Where w_A contains the normalized area percentages of the template and input fingerprints, w_B contains normalized weights which favor triangles neat the image center, $\alpha \in [0,1]$ adjusts the significance of w_A and w_B. Consequently, the FFM measure for template and input fingerprints is defined as

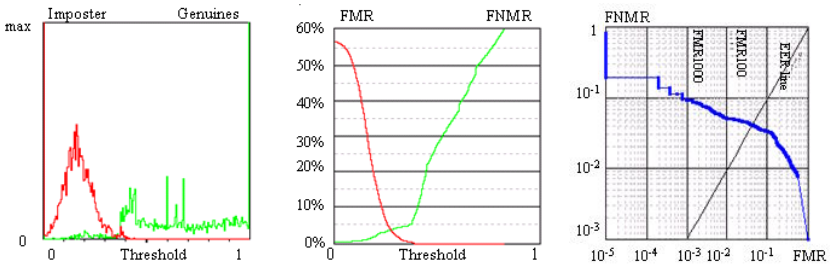
$$Sim = [\,(1 - \alpha)w_A + \alpha w_B\,]\, L^{(T,I)} \tag{10}$$

3 Experimental Results

The proposed algorithm has been evaluated on the fingerprint database of FVC2004. In FVC2004, the organizers have particularly insisted on: distortion, dry and wet fingerprints. Especially in fingerprints database DB1 and DB3 of FVC2004, the distortion between some fingerprints from the same finger is large. Hence, the evaluation of the proposed algorithm is mainly focused on DB1 and DB3 of FVC2004. The proposed algorithm is also compared with the algorithm described by Xiping Luo et al. [13] and the algorithm proposed by Bazen et al. [9].

3.1 Performance on FVC2004 DB1

The fingerprints of FVC2004 DB1 were acquired through optical sensor "CrossMatch V300". The fingerprint database set A contains 800 fingerprint images captured from 100 different fingers, 8 images for each finger. Figure 1 shows an example of large distortion from FVC2004 DB1 (102_3.tif and 102_5.tif). Using the proposed algorithm, the similarity between these two fingerprints is 0.43012. The performance of the proposed algorithm on FVC2004 DB1 is shown in Figure 4. From Figure 4, we find that the similarity threshold at EER point is about 0. 265, so we can judge that the above fingerprint pairs come from the same finger. The EER of the proposed algorithm on FVC2004 DB1 is about 4.06%. The average time for matching two minutiae sets is about 1.12 second on PC AMD Athlon 1600+ (1.41 GHz).



Score distributions FMR(t) and FNMR(t) ROC curve

Fig. 4. Experimental results of the proposed algorithm on FVC2004 DB1_A. The fingerprint images were acquired through optical sensor

3.2 Performance on FVC2004 DB3

The fingerprints of FVC2004 DB3 were acquired through thermal sweeping sensor "FingerChip FCD4B14CB" by Atmel. The size of the image is 300*480 pixels with the resolution 512 dpi. In this fingerprints DB, the distortion between some fingerprints from the same finger is also large. The performance of the proposed algorithm on FVC2004 DB3 is shown in Figure 5. The equal error rate of the proposed algorithm on FVC2004 DB3 is about 1.35%. The average time for matching two minutiae sets is about 1.08 second on PC AMD Athlon 1600+ (1.41 GHz).

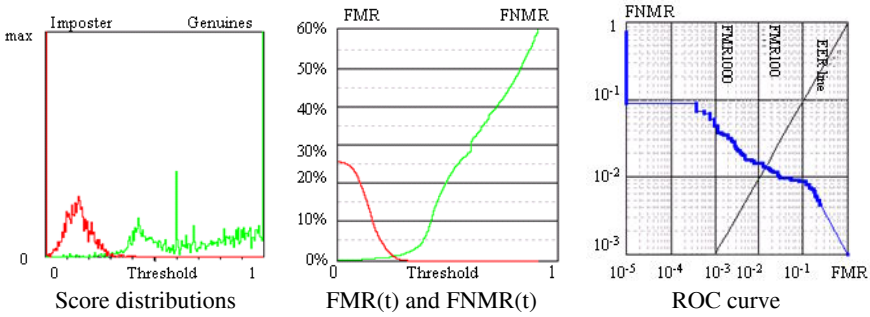


Fig. 5. Experimental results of the proposed algorithm on FVC2004 DB3_A. The fingerprint images were acquired through thermal sweeping sensor

3.3 Comparison of FFM with Other Algorithm

The proposed algorithm is compared with the algorithm described by Xiping Luo et al.'s [13]. The Comparison is performed on FVC2004 DB1. Table 1 lists the comparison of matching score of the proposed algorithm to Xiping Luo's algorithm. It is clearly that case 1 and case 2 have been successfully matched in our FFM algorithm. While for Xiping Luo's method, when size of the bounding boxes is 15, case 1 and case 2 are false rejected. And when size of the bounding boxes is increased to 25, case 2 is accepted but case 1 is false rejected. However, EER is increased from 9.13% to 9.92% at the same time. And EER of the proposed algorithm (4.06%) is far lower than EER of the algorithm proposed by Xiping Luo's (9.13%).

Table 1. Comparison of matching score in the FFM algorithm to Xiping Luo's method

Performance comparison	The proposed algorithm	Xiping Luo's method: size of the bounding boxes = 15	Xiping Luo's method:size of the bounding boxes = 25
Score of case 1 in figure 9	0.43012	0.050000	0.194000
Score of case 2 in figure 10	0.52300	0.061800	0.273600
Score threshold at EER point	0. 26100	0.152000	0.213100
EER(on FVC2004 DB1_A)	4.06%	9.13%	9.92%

The proposed algorithm is also compared with the algorithm proposed by Bazen et al. [9]. In their training database of FVC2002 DB1, for Bazen et al. algorithm, the

EER of the ROC turned out to be 1.8% with r0 = 5 for elastic matching. While in our algorithm, the EER on FVC2002 DB1 is only about 0.26%. The results indicate that the performance of the proposed algorithm surpasses Bazen et al.'s algorithm.

4 Conclusion and Discussion

Coping with non-linear distortions in fingerprint matching is a real challenging task. This paper proposed a novel method for matching the deformed fingerprints. The fingerprint was represented by the fuzzy features: local triangle features set. The similarity between fuzzy features is used to character the similarity between fingerprints. The proposed algorithm was evaluated with fingerprint databases of FVC2004. In DB1 and DB3 of FVC2004, the distortion between some fingerprints from the same finger is large. However, EER are only 4.06% and 1.35% on DB1 and DB3 respectively in our algorithm. And the proposed algorithm has good performance on processing time. The average time for matching two minutiae sets is about 1.1 second. Further investigation is going on to improve the performance of the proposed algorithm.

References

1. Biometric Systems Lab, Pattern Recognition and Image Processing Laboratory, Biometric Test Center, http://bias.csr.unibo.it/fvc2004/
2. N. K. Ratha and R. M. Bolle, "Effect of controlled acquisition on fingerprint matching", 14th ICPR, 1998, Proc., vol. 2, pp. 1659–1661.
3. Chitra Dorai, Nalini Ratha, and Ruud Bolle, "Detecting dynamic behavior in compressed fingerprint videos: Distortion", in Proc. CVPR2000, Hilton Head, SC., Jun. 2000.
4. R. Cappelli, D. Maio, and D. Maltoni, "Modelling plastic distortion in fingerprint images", in Proc. ICAPR2001, Rio de Janeiro, Mar. 2001.
5. Dongjae Lee; Kyoungtaek Choi; Jaihie Kim, "A robust fingerprint matching algorithm using local alignment", 16th ICPR, 2002. Proc., Vol. 3, pp. 803-806.
6. A. Senior and R. Bolle, "Improved fingerprint matching by distortion removal", IEICE Trans. Inf. and Syst., Special issue on Biometrics, E84-D (7): 825-831, Jul. 2001.
7. Watson, C., Grother, P., Cassasent, D.: "Distortion-tolerant filter for elastic-distorted fingerprint matching". In: Proceedings of SPIE Optical Pattern Recognition. (2000), pp. 166-174
8. Z.M. Kovács-Vajna, "A fingerprint verification system based on triangular matching and dynamic time warping", IEEE Trans. Pattern Anal. Machine Intell., vol. 22, no. 11, pp. 1266-1276, 2000.
9. Asker M. Bazen, Sabih H. Gerez, "Fingerprint matching by thin-plate spline modelling of elastic deformations", Pattern Recognition, Vol. 36, Issue 8, 2003, pp.1859-1867.
10. A. Ross, S. Dass and A. K. Jain, "A Deformable Model for Fingerprint Matching", Pattern Recognition, Vol. 38, No. 1, Jan 2005, pp. 95-103.
11. F.Hoppner, F.Klawonn, R.Kruse, and T.Runkler, "Fuzzy Cluster Analysis: Methods For Classification", Data Analysis and Image Recognition, John Wiley & Sons, 1999.
12. Yixing Chen, James Z. Wang, "A Region-Based Fuzzy Feature Match Approach to Content-Based Image Retrieval", IEEE Trans. Pattern Anal. Machine Intell., vol. 24, no. 9, pp. 1252-1267, Sep. 2002.
13. Xiping Luo, Jie Tian and Yan Wu, "A Minutia Matching algorithm in Fingerprint Verification", 15th ICPR, 2000, Proc., Vol.4, pp.833-836.

Minutiae Quality Scoring and Filtering Using a Neighboring Ridge Structural Analysis on a Thinned Fingerprint Image

Dong-Hun Kim

Department of Image Signal Processing, NITGen Co. Ltd., Seoul, Korea
zava@nitgen.com

Abstract. This paper introduces a novel minutiae quality scoring method that relies on analyzing neighboring ridge structures around a minutia on thinned Fingerprint image. Normal ridges with neighboring a minutia have regular structures in the parts of inter-ridge distance, connectivity, and symmetry etc. This is important features of measuring minutiae quality score. It is meaning of a possibility that the present minutia is a true minutia. For making the score, Two test DB sets is firstly made for TM(True Minutiae sets) and FM(False Minutiae sets) by manually filtering minutiae found from automatic extraction. Then the score function is made by statistical method with Bayesian rule for TM and FM. I should evaluate for its discrimination power to these sets and apply to false minutiae filtering in extraction. Experimental results showed that minutiae for TM class is not nearly filtered, but ones for FM class is filtered about 30%. Therefore, I should confirm that it is useful and compatible for minutiae filtering and have an expectation in some fields.

1 Introduction

As fingerprint recognition has become commonly known, the higher performance of identification and verification is needed in real world such as the banking, insurance supply etc. For more practical usage of fingerprint in these fields, we must deal with low quality fingerprints which have some problems owing to false accepted and false rejected cases. To enhance a quality of fingerprint it needs filtering techniques such as matched filter, gabor filter, fft filter etc. These methods have been developed and have made an improvement of ridge directionality and the deduction of noises in the sense of Image enhancement. Also, post-processing and minutiae filtering on thinned image have researched as enhancement in the focus on minutiae extraction.[1]

In this manner, Fingerprint recognition has been gradually improved by various methods in the aspects of the performance enhancement. Particularly, the methods with respect to reducing the number of false minutiae are noteworthy because they can attenuate both False Acceptance error Rate and False Rejection error Rate. Several factors such as the presence of scars, worn artifacts, variations of the pressure between the finger and acquisition sensor, the environmental conditions during the acquisition process, the season factor and so on, can dramatically make the false minutiae. Many researchers have suggested with image enhancement as a preprocessing[2] and image restoration of thinned image[4][5][6] to delete the false minutiae in a broad sense. Minutiae filtering[7] is also reasonable method with high efficiency for time-cost and memory-cost. It has been used to filter the false minutiae

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 674–682, 2005.
© Springer-Verlag Berlin Heidelberg 2005

using the topological analysis of ridges on skeletonized fingerprint image. This became a motivation for me to use structural analysis around a minutiae.

Regular minutiae consists of linked neighboring ridges that have been wired in particular direction and distance in parallel.[3] This is a useful parameter to measure whether is a true minutia or not. When this process has been completed, it could make the Minutiae Quality Score(MQS) based on statistical data of True Minutiae sets and False Minutiae sets, which are constructed from the manual filtering on minutiae obtained from automatic extraction. Minutiae Quality Score is newly given as values from 0 to 100. This is applicable in some fields when dealing with the minutiae in grade. For example, they could be pre-matched with credible minutiae by selecting a MQS threshold in advance.[8] Also, an alignment with believable minutiae selected by a MQS threshold could exist. It should find more exact pivot that is centered of translation and rotation at matching. etc.

When we look at the noisy image, we must take in a note of the following fact that MQS distribution could have a considerable effect on current preprocessing ability, especially noise reduction as a thinned image is an input of MQS processing. On the other side, we could make the measure of enhancing ability through the MQS distribution.

In this paper, definition of the parameter is stated in the beginning and detailed description of the Minutiae Quality Scoring method is stated. To close, confirmation of the validity is proven through experimental results.

2 Parameters Description of a Neighboring Ridge Structure Around a Minutia

2.1 Inter-ridge Distance (IRD)

IRD is a inter-ridge distance between two contacted points that is obtained by orthogonally adjacent neighboring ridges with a basis on minutia direction $(\overrightarrow{PP_d})$. The parameter can be used with similar method both END and BRANCH cases. According to Fig.1, IRD[0][1] is the same meaning of $\overline{P_{d1}P_{d2}}$ that notes a distance between left second neighboring ridge and left first neighboring ridge about the basic ridge. And interesting minutiae point is laid on basic ridge.

IRD has 4 sub-members which is defined variables with [position][neighboring order] judged by a basic ridge. In addition to, VIRD is defined as the same method with the exception of measuring based on the opposite of minutiae direction $(\overrightarrow{PP_u})$ only.

2.2 Connectivity Between Corresponding Points (CCP) by Connected Ridge Figures (CRF)

CCP is decided by CRF which has two ridge connectivity parameters. CRF consists in two parameters obtained by across traced results as it is shown at Fig.1. And the CCP value can be three indexes such as 0, 1, 5 according to the CRF rule that are meaning of unconnected, connected, and branch connected ridge, respectively. Here, I have defined the CRF rules involved structural information. For example, CCP is equal to 0

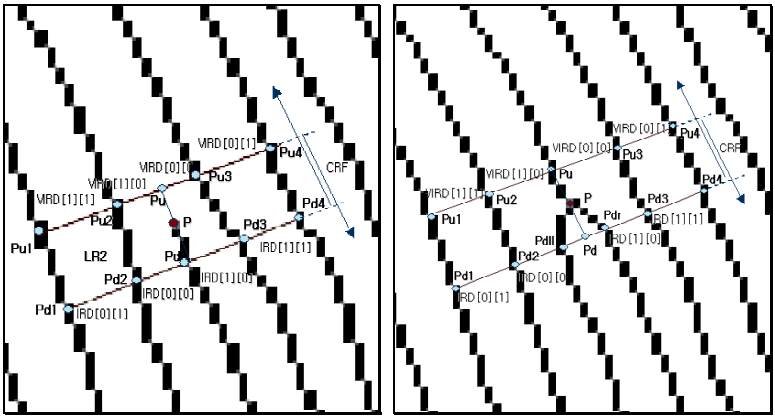


Fig. 1. (a) Ridge Structure Parameters at End, (b) Ridge Structure Parameters at Branch

when CRF is defined (1, 1). This parameter is the meaningful element that provides the structure of neighboring ridges at a interested minutia roughly.

2.3 Global Mean Ridge Width (GMRW) vs. Local Mean Ridge Width (LMRW)

GMRW is very important factor, for it is a standard measure of scoring in this paper. Through the evaluation test about the relation between GMRW and LMRW, I had known the fact that the error between GMRW and LMRW is usually converged to zero in two pixels or so. It makes an allowance for me to adopt GMRW as a standard measure if the margin of two pixels is considered under progressing. Then, GMRW could be regarded as the standard measure instead of the LMRW though MQS is considered on local region.

3 Minutiae Quality Scoring

3.1 Definition of a Score Function

Score function is designed with 3 sub-score parts and 4 weight ratios for optimal solution. Total score is defined as S,

$$\mathrm{S} = w_n \cdot (w_0 \cdot S_0 + w_1 \cdot S_1 + w_2 \cdot S_2) \tag{1}$$

where S_0 is a CCP-based Score, S_1 is a IRD-based Score, S_2 is Symmetry-based Score, w_n is a noise score ratio, and w_i is score weight ratios $(i = 0,1,2)$.

3.2 CCP-Based Score

CCP-based Score is obtained by statistical method with Bayes' Theorem. We make the code (CCP[0][1], CCP[0][0], CCP[1][0],CCP[1][1])=(x_1, x_2, x_3, x_4) for analyzing in the two class TM and FM as stated before. Then, Bayes' Theorem based on Inde-

pendent conditional probability for multiple features is used for scoring because normal neighboring ridges shall be independent on each other. The result is

$$P(x_1, x_2, x_3, x_4) = \sum_{i=1}^{k} P(C_i) P(x_1, x_2, x_3, x_4 | C_i) \tag{2}$$

$$P(x_1, x_2, x_3, x_4 | C_i) = P(x_1 | C_i) \ldots P(x_4 | C_i) \tag{3}$$

$$S_0 = P(C_i | x_1, x_2, x_3, x_4) = \frac{P(C_i) P(x_1 | C_i) \ldots P(x_4 | C_i)}{\sum_{j=1}^{K} P(C_j) P(x_1 | C_j) \ldots P(x_4 | C_j)} \tag{4}$$

CCP-based score is namely the meaning of the probability that class TM, given CCP code, can be estimated. Here, there are some filtering conditions because CCP code is reflected local ridge structures. The parameter selection of the conditions is important for performing the appropriate filtering.

3.3 IRD-Based Score

IRD-based Score is made from the sum of error function results for difference between GMRW and IRD. It is given by

$$S_1 = \sum_{i=1}^{8} f_1(\lambda - RD_i) \tag{5}$$

Where $f_1(\cdot)$ is score transform function obtained from statistical data of TM class, RD_1 is a distance of IRD[0][0], RD_2 is a distance of IRD[1][0], RD_3 is a distance of VIRD[0][0], RD_4 is a distance of VIRD[1][0], RD_5 is a distance of IRD[0][1], RD_6 is a distance of IRD[1][1], RD_7 is a distance of VIRD[0][1], RD_8 is a distance of VIRD[1][1]. Here, the score factor is regarded as the absolute value about distance error between GMRW and RD_i. Here, there is careful about scoring parameters, *[0][0] and *[1][0], at End case.

In Fig.2., the score factor of IRD and VIRD are set to –5 when its value is equal to zero for the sake of convenience with view. This part is ascertained with a suitable filtering condition by comparing distributions of the class TM and FM.

3.4 Symmetry-Based Score

Symmetry-based score is considered with up-down and left-right symmetry. The score is given by

$$S_2 = s_{up-down} + s_{left-right} = \sum_{i=1}^{2} f_2(DE_{ud}{}^{i}) + \sum_{j=1}^{4} f_3(DE_{lr}{}^{j}) \tag{6}$$

Where $f_2(\cdot)$ and $f_3(\cdot)$ is the score transform function obtained by statistical data in TM class, $DE_{ud}{}^{1}$ is a distance error between IRD[0][0] and VIRD[1][0] for the first-left-neighboring ridges. $DE_{ud}{}^{2}$ is a distance error between IRD[0][1] and VIRD[1][1] for the first-right-neighboring ridge, $DE_{lr}{}^{1}$ is a distance error between IRD[0][0] and

IRD[1][0], $DE_{lr}{}^{2}$ is a distance error between IRD[0][1] and IRD[1][1], $DE_{lr}{}^{3}$ is a distance error between VIRD[0][0] and VIRD[1][0], and $DE_{lr}{}^{4}$ is a distance error between VIRD[0][1] and VIRD[1][1].

The PDFs of a score factor, $DE_{ud}{}^{i}$ and $DE_{lr}{}^{j}$, for the class TM and FM is shown at Fig.3.

The distributions for the class TM and FM are different each other, and the error between them can become the basis of making the score.

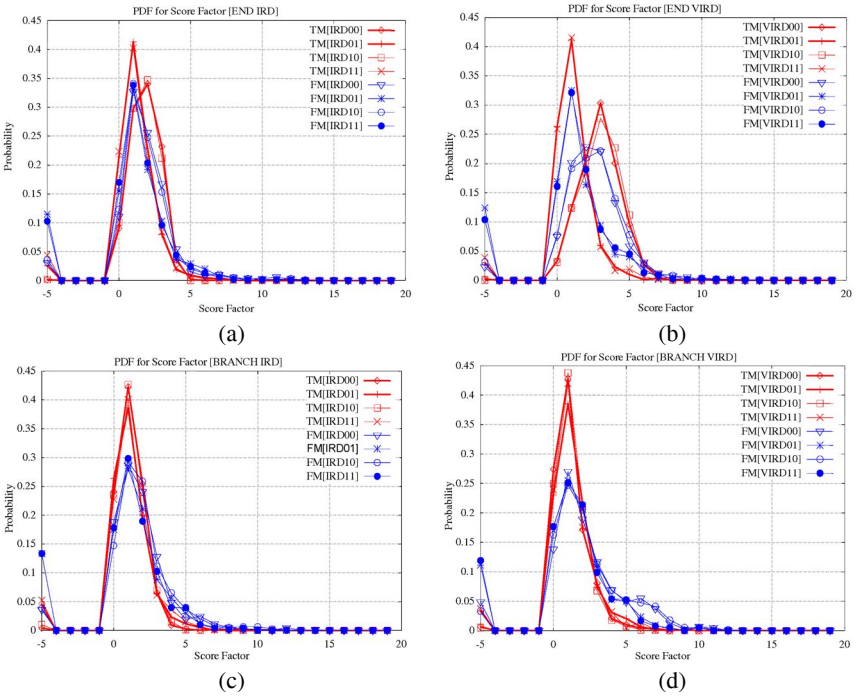


Fig. 2. (a)-(b) PDF for score factor at End, (c)-(d) PDF for score factor at Branch

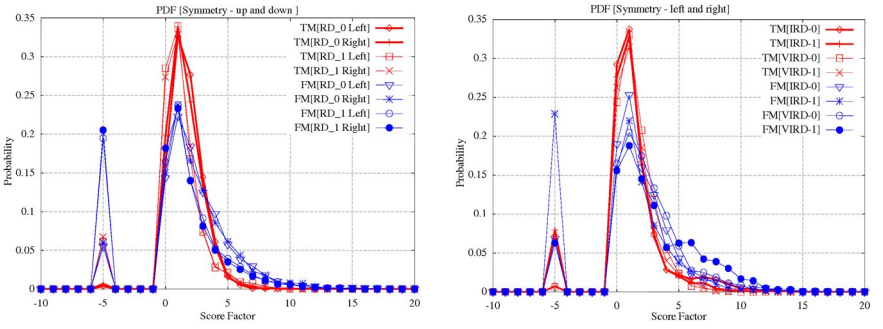


Fig. 3. (a) PDF for Symmetry-based score at End, (b) PDF for symmetry-based score at Branch

3.5 Find Score Transform Function ($f(x)$)

For a given score factor input x, the score transfer function is found by a conditional probability and cumulative distribution of x in the class TM from equation (7) to (10). The probability that the class TM, given the score factor x, can be estimated

$$P(TM|x) = \frac{P(TM)P(x|TM)}{P(x)} = \frac{P(TM)P(x|TM)}{P(TM)P(x|TM) + P(FM)P(x|FM)} \tag{7}$$

At this time, the likelihood ratio can be used for more simple calculation. When there are only two class TM and FM, the likelihood ratio is

$$R = \frac{P(TM)P(x|TM)}{P(FM)P(x|FM)} \tag{8}$$

And

$$P(TM|x) = \frac{R}{1+R} \tag{9}$$

The score transfer function is defined as $f(x)$,

$$f(x) = \kappa \cdot P(TM|x)P_{CDF}(x|TM),\ \kappa \text{ is constant} \tag{10}$$

Where the prior probability $P(TM)$ and $P(FM)$ is experimentally same as 0.5 and $P_{CDF}(x \mid TM)$ is a cumulated distribution of $P(x \mid TM)$. The score transfer function is considered with the probability that the class TM, given score factor x, can be estimated and the probability of the existence of score factor x in the class TM. Here, there are the score transform functions to use for scoring practically at Fig.4.

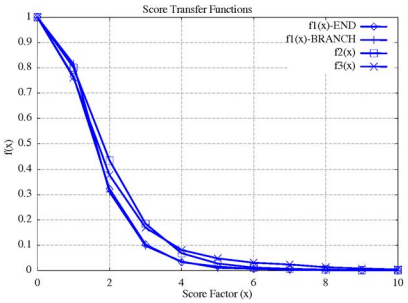


Fig. 4. The score transfer function $f(x)$ for each scoring

The proper design of the score transfer function is a important issue because it makes considerably an influence on the score distribution as a whole.

3.6 Noise Score Ratio

The noise score ratio can be decided with minutiae count numbers on condition that there are many minutiae in noisy region simply. Fig.5. indicate this assumption is rational to use. It shows that minutiae numbers centers round on a interesting minutia

are under 2 in radius of RW distance and under5 in radius of doubled RW distance for the TM class.

When there are minutiae in a fixed distance around a minutia, noise score ratio is given by

$$w_n = \begin{cases} 1 & , if\ N_{RW} < 2\ and\ N_{2RW} < 5 \\ 0.4 & , if\ N_{RW} < 2\ and\ N_{2RW} \geq 5 \\ 0.2 & , if\ N_{RW} \geq 2\ and\ N_{2RW} < 5 \\ 0 & , if\ N_{RW} \geq 2\ and\ N_{2RW} \geq 5 \end{cases} \tag{11}$$

Where N_{RW} is the minutiae count numbers existed within GMRW, N_{2RW} is the minutiae count numbers existed within a section between GMRW and GMRW multiplied with 2. And the thresholds to set the noise score ratio are decided by statistical data from Fig.5.

4 Experimental Results

4.1 Total Score Results

To evaluate the availability of the Minutiae Quality Score, it is executed by comparing the discrimination power through score distribution for the class TM and FM, which is composed of 8729 minutiae and 2086 minutiae in each class.

The result without filtering conditions is shown at Fig.6. The score distribution is reasonable to discriminate the classes as the above.

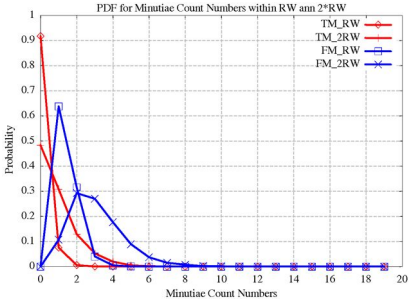
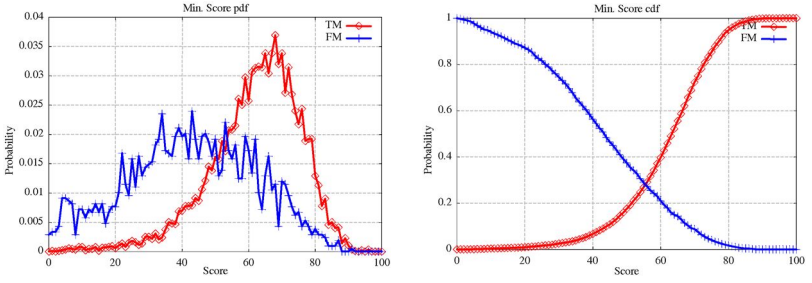


Fig. 5. PDF for Minutiae count numbers existed in a given distance

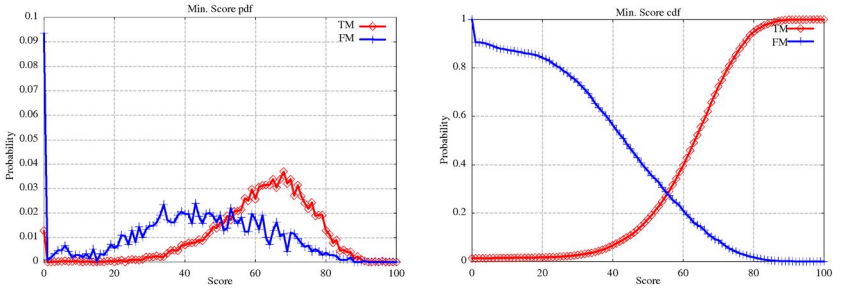


(a) (b)

Fig. 6. (a) PDF for Minutiae Quality Score for the class TM and FM at w=[0.2,0.3,0.5], (b) CDF or Minutiae Quality Score for the class TM and FM at w=[0.2,0.3,0.5]

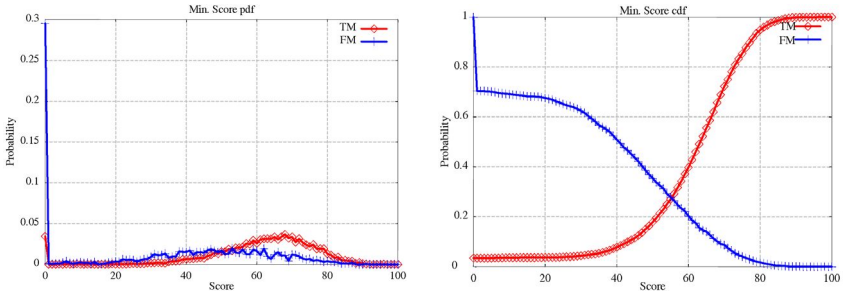
4.2 Filtering Results

Filtering test is executed with beforehand conditions of IRD and CCP. Experimental results is that there are about 30% attenuation in the class FM, but about 3% increment in the class TM at score 0 at Fig.8. It indicates Minutiae Quality Score is fairly useful.



(a) (b)

Fig. 7. (a) PDF for Minutiae Quality Score and IRD filtering for the class TM and FM (b) CDF for Minutiae Quality Score and IRD filtering for the class TM and FM



(a) (b)

Fig. 8. (a) PDF for Minutiae Quality Score and IRD plus CCP filtering for the class TM and FM (b) CDF for Minutiae Quality Score and IRD plus CCP filtering for the class TM and FM

5 Conclusions

In this paper, the novel Minutiae Quality Score(MQS) is defined by statistical methods with the parameters of ridge structures around a minutia. It is evaluated with comparing the distribution in the class TM and FM, which is constructed for the purpose. It shows that MQS is useful information. Also, filtering test is accomplished with conditions obtained under making the MQS. The result has obtained with approximately 30% decrement of false minutiae in the class FM without nearly changing in the class TM. Experimental results show that MQS can be usable for filtering the false minutiae. Later, I will compare MQS with other quality measure algorithms for quality check ability and shows minutiae sets filtered by MQS result in better performance than a current minutiae sets. In current MQS, Parameters and the score

transfer function can be optimized for more discrimination power. MQS shall be applied to a matching part and any other enrollment process after this. Therefore, It may make an expectation for me to upgrade in fingerprint recognition.

References

1. Maltoni M., Maio M., Jain A., Prabhakar, "HandBook of Fingerprint Recognition", 2004.
2. L. Hong, Y. Wan and A. K. Jain, "Fingerprint Image Enhancement: Algorithms and Performance Evaluation", IEEE Transactions on PAMI, Vol. 20, No. 8, pp. 777-789, August 1998.
3. M. Tico, V. Onnia and P. Kuosmanen, "Fingerprint Image Enhancement Based on Second Directional Derivative of the Digital Image", EURASIP journal on Applied Signal Processing, Vol.10, pp.1135-1144, 2002.
4. A. Farina, Zs. M. Kovacs-Vajna, and A. Leone, "Fingerprint Minutiae Extraction from skeletonzed binary images", Pattern recognition, Vol. 32, pp 877-889, 1999.
5. Q. Xiao, and Hazem Raafat, "Fingerprint image postprocessing: a combined statistical and structural approach," Pattern Recognition, Vol. 24, pp. 985-992, 1991.
6. L. Wenxing, W. Zhaoqi, M. Guoguang, "Thinned fingerprint image post-processing using ridge tracing",Proceedings of SPIE, Vol.4552, 2001.
7. F. Zhao and X. Tang, "Duality-based Post-processing for Fingerprint Minutiae Extraction," InfoSecu, 2002.
8. P. Bhowmick, A. Bishnu, B. B. Bhattacharya, C. A. Murthy and T. Acharya, "Determination of minutiae scores for fingerprint image application", Proc. 3rd Indian Conf. On Computer Vision, Graphics and Image Processing, pp. 464-368, 2002.

Hardware-Software Codesign of a Fingerprint Identification Algorithm

Nicolau Canyellas[1], Enrique Cant[1], Giuseppe Forte[1], and Mariano L[2]

[1] Escola T[cnica Superior d]Enginyeria ETSE-URV, Tarragona, Spain
ncanyell@etse.urv.es
http://www.etse.urv.es
[2] Escola Universit[ria Polit]cnica de Vilanova i la Geltr[EUPVG-UPC, Spain
http://www.upc.es

Abstract. Automatic Fingerprint Authentication Systems are rapidly being incorporated in a wide range of applications, satisfying the society demand of accurate identification frameworks in order to prevent unauthorised accesses or fraudulent uses. Most of biometrics based personal identification systems run on high-performance computer based platforms, which execute a set of complex algorithms implemented in software. Those solutions cannot be applied to small, low-cost and low-power embedded systems, based on microprocessors without floating-point arithmetic unit. In this article we present a hardware/software implementation of a fingerprint minutiae extraction algorithm. The proposed system consists of a microprocessor and a coprocessor implemented in an associated FPGA. In order to develop an efficient implementation, fixed-point computations have substituted the floating-point ones. Due to the low feature requirements the whole system is suitable for a SoC with embedded flexible hardware.

1 Introduction

Fingerprints, which have been used for about 100 years in personal identification [6][10], are the oldest biometrics signs of identity, and the most widely used today. Fingerprint biometrics offers a set of key advantages over other biometrics: small and low-cost capacitive or thermal sensors, very low variation of a fingerprint along time, suitability in medium-high security authentication applications, and commodity to users. Any Automatic Fingerprint Authentication System (AFAS) is structured in three different stages: image acquisition, feature extraction and matching, this work covers the minutiae extraction stage, the one with highest computational requirements.

The most common representation used in fingerprint identification is the Galton features or minutiae [10]. Minutiae are the set of points where the ridges of the fingerprint end or split, and minutiae-matching is the usual way to perform automatic fingerprint comparison. It offers a good saliency property (features extracted from fingerprint images contain distinctive information), although not so good suitability (easy feature extraction, low storing requirements and useful matching) due to the high computational capacity required during minutiae extraction

Several approaches to automatic minutiae extraction have been proposed [13][17]. Most of them use orientation field computation, directional filters, binarization and thinning in order to obtain a binary representation of the fingerprint image, where the

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 683 692, 2005.
Springer-Verlag Berlin Heidelberg 2005

thickness of the ridge lines is reduced to one pixel, from where minutiae is easily extracted. In our work we use the Maio and Maltoni ridge line following algorithm [14], because it permits minutiae extraction directly from the gray-scale fingerprint image. This method is computationally less expensive than others, and it can be re-written to be implemented without floating point operations. These facts are very important, bearing in mind that the goal is to implement a high-speed and low-cost embedded system. In order to implement a real-time AFAS we propose a hardware/software design; it consists of a microprocessor with an associated specialised coprocessor, able to perform the most time-consuming functions of the ridge line following algorithm.

The use of fingerprint biometrics coprocessors is still a young field. A great majority of commercial fingerprint OEM modules [1][3][7][9][19] are based on embedded high performance 32-bit processors or DSPs (Digital Signal Processors), but feature extraction times are about 1 second or more. The IKENDI manufacturer offers the IKE-1 ASIC [12] based on an ARM7TDMI processor and a ConvTree-III coprocessor comprising less than 50K gates and it claims to encode 10 to 20 times faster than other DSP solutions at the same clock speed. The UCLA University group has developed the ThumbPod prototype [22], where the fingerprint recognition is based on a 32-bit LEONII microprocessor and a DFT (Discrete Fourier Transform) coprocessor mapped on a high-end Xilinx Virtex II FPGA. The DFT coprocessor is used to determine the dominant ridge flow direction in every block. Experimental results show that the coprocessor permits a 55% execution time reduction for the minutiae extraction, but 4 seconds of execution time is still quite high for many applications.

All of these solutions implement the AFAS system executing a set of software-implemented complex algorithms, and require the use of high feature processors or DSPs in order to perform a real time authentication. In our approach we solve the authentication problem using a low complexity algorithm, and designing a specific coprocessor for it.

A brief description of the ridge line following algorithm is presented in section 2, including our modifications in order to reduce the computational requirements needed to obtain an efficient hardware implementation. In section 3 the results of a software implementation are presented, showing the percentage of overall execution time spent in each one of the algorithm main functions. The critical tasks of the algorithm, located with this information, are mapped on the coprocessor, and section 4 describes its hardware. The obtained experimental results are presented in section 5. Finally section 6 presents the conclusions of our work.

2 The Ridge Line Following Algorithm

For minutia extraction we have analysed the ridge line following algorithm with the parameter values adopted in [14], and then we have modified some processes of the proposed methodology in order to minimise the computational requirements. These modifications are carried out bearing in mind the hardware implementation of critical tasks. We developed Matlab and C versions of the original algorithm, and then we started to simplify the used data types, changing floating point operations with integer versions, and trying to eliminate the computationally expensive operations like products and divisions.

From a mathematical point of view, a ridge line is defined as a set of points which are local maxima along one direction. The algorithm is based on the idea of tracking the fingerprint ridge lines on the gray scale image by "sailing" over these points, according to the fingerprint directional image. At each iteration of Maio's algorithm [14], during the ridge following, a new section Ω is determined on the orthogonal direction to the ridge ($\Phi c = \varphi + \pi/2$ where φ denotes the ridge direction. Then the process locates a local maximum over a regularised silhouette of Ω in order to determine the next iteration point. As depicted in Fig. 1, given a starting point (i_c,j_c) and a starting direction φ_c the algorithm computes a new point (i_t,j_t) by moving μ pixels from the starting point along direction φ_c. At this point the section Ω is computed as the set of $2\sigma+1$ points orthogonal to φ_c and having (i_t,j_t)as median point. A third step involves the computation of an averaged section of Ω, denoted as avg(Ω), as the local average of each of the pixels belonging to section Ω with the two pixels belonging to the two parallel planes at a distance of ± 1 pixel on the φ_c direction, denoted as Ω^{+1} and Ω^{-1}. The next step computes a correlation corr(Ω) of the averaged section avg(Ω) against the gaussian silhouette mask showed in Fig. 3(a), and returns its weak local maximum (i_n,j_n). An auxiliary binary image T of the tracked points and its neighbours is updated in order to prevent a second tracking of the same ridge. Finally (i_n,j_n) and its dominant ridge flow direction φ_n, become the new starting point and direction to start a new iteration. These steps are repeated during a ridge line following until a stop criterion is reached.

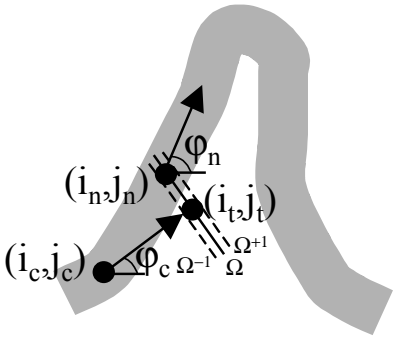


Fig. 1. Ridge line following of a fingerprint ridge, from point (i_c, j_c) to point (i_n, j_n)

The most important change in the algorithm relates to the ridge direction computation, φ_c. This direction represents the ridge line local direction and can be computed as the tangent to the ridge at point (i_c, j_c). The method used by Maio was proposed by Donahue and Rokhlin [8], and is computationally more expensive than the presented in [21] and used by Halici and Ongun in [11]. We have modified the proposed mask in order to obtain 24 different directions with angles varying in steps of 15°, which corresponds with $\mu = 3$ as used by Maio (see Fig 2).

An inspection of the extraction algorithm shows another computational expensive step. In the process of finding the local maximum in a section Ω, the algorithm first calculates the local average of the gray levels of the pixels that belong to Ω, Ω^{+1} and Ω^{-1}, to finally compute a correlation with a gaussian silhouette mask.

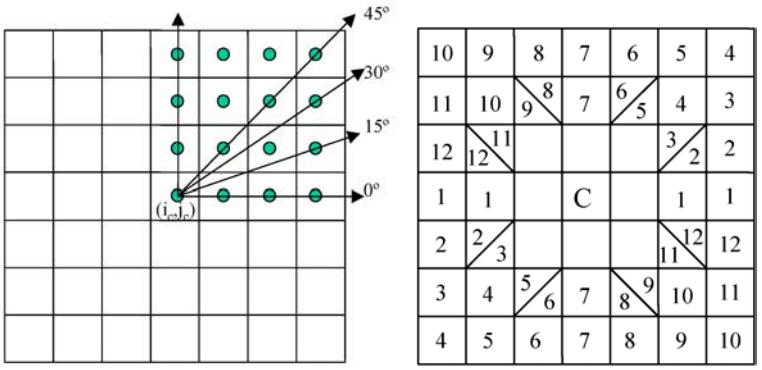
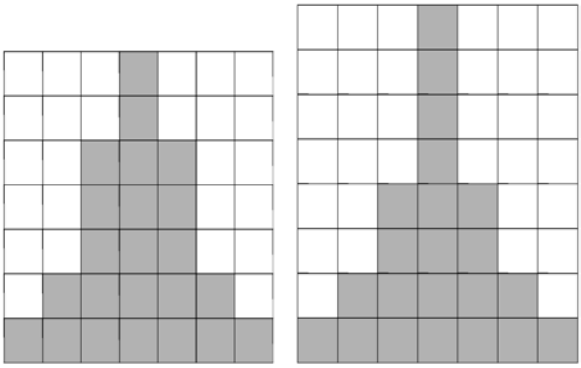


Fig. 2. Selection of angles in 15º steps and the direction computation mask

The local average with two neighbors pixels involves a division by three, and the correlation includes a division by 23; in our algorithm none of these divisions are performed. These changes are equivalent to a change of scale, without loss of accuracy, and do not affect the position of the local maximum.

Moreover, a correlation involves products that can be substituted by bit-shift operations, which can be implemented in hardware in a very efficient way, if performed with values that are powers of 2. Fig. 3 presents the original gaussian mask (a) and our proposed one (b), with canged weighs in order to simplify the arithmetic operations involved in the correlation.



(a) (b)

Fig. 3. The mask has a symmetric Gaussian silhouette. The proposed in [14] (a) with weighs [1/23, 2/23, 5/23, 7/23, 5/23, 2/23, 1/23], and the used in this work (b) with weighs [1, 2, 4, 8, 4, 2, 1]

Our modified version of the algorithm has been validated using the same samples used in [14] and available at [2]. It is composed of seven fingerprints taken from the NIST fingerprint database [20], four fingerprints from an FBI sample set and three fingerprints acquired through an opto-electronic device based on a prism. Table 1 reports the results in terms of undetected minutiae (dropped), non-existent minutiae (false) and type-exchanged minutiae (exchanged).

Table 1. Comparison between algorithms (the second column indicates the number of certain minutiae detected manually)

		original			modified		
fingerprint	minutiae	d	f	x	d	f	x
1	33	0	2	7	0	2	5
2	29	3	1	4	0	2	3
3	28	1	2	4	1	3	4
4	37	3	0	4	4	1	4
5	22	0	0	3	0	3	2
6	23	0	0	4	0	2	2
7	31	2	1	2	1	2	3
8	31	1	0	3	2	0	1
9	21	1	10	1	0	11	2
10	22	1	0	4	1	2	4
11	32	3	5	4	3	6	5
12	33	3	8	2	2	6	5
13	20	0	0	4	2	1	4
14	37	0	5	6	3	3	4

There are minor differences in the comparison between the results reported in Maio's original work with the ones obtained with our modified version of the algorithm. So, the algorithm modifications are validated, and it is demonstrated that is possible to construct a complete identification system based in our modified version of the ridge line following algorithm. In both cases most of errors of the ridge line following algorithm are minutiae exchanges, mainly due to some termination minutiae which are detected as bifurcation minutiae. If a termination minutia is very close to another ridge line the algorithm may skip the termination and intersect the adjacent ridge line. In [14] it is proposed to train a neural network to verify the type of minutiae detected, performing a local analysis of the gray scale image in each minutia neighbourhood. In our approach we are evaluating the method proposed by Methre [15]. We study the dominant direction by computing a histogram of directions in the neighbourhood of the minutia. If there is only one dominant direction the minutia is marked as a termination, but if there are two dominant directions it is marked as a bifurcation.

3 Hardware-Software Partitioning

The software implementation of the algorithm reflects the process of "sailing" over a ridge as described in Section 2, and it consists of a set of loops executing the same iterative process, one time and another.

Code 1. Main functions of our integer version of the algorithm

```
while (!end_condition) {
  next_point(angle_c,μ,&delta_i,&delta_j);
  i_t = i_c+delta_i;
  j_t = j_c+delta_j;
  createsecc(i_t,j_t,angle_c,seccX,seccY);
  filtersecc(seccX,seccY,angle_c,avg);
```

```
  index = corr_weak_max(avg);
  i_n = seccX[index];
  j_n = seccY[index];
  angle_n = tg_dir(j_n,i_n,angle_c);
  ...
  updateT(j_c,i_c,j_n,i_n);
  ...
}
```

The C source lines listed above as Code 1, show the execution of the basic functions used in the inner loop of our version of the ridge-following algorithm. The rest of the minutiae extraction program checks for a stop criterion, and controls the correct processing of the whole fingerprint, but representing a few part of the total execution time.

We analysed the code, detecting the most critical tasks, and then we choose the hardware-software partitioning in order to accelerate the minutiae extraction process. The profiling of the program functions is detailed in Table 2. The function name is in the first column, the second one shows the total number of calls to it, and the last column represents the percentage of the overall time spent in the execution of the function.

We decided to implement a coprocessor that computes the first nine lines of Code 1, after the while sentence. So, the coprocessor will map the first five functions of the Table 2, which represent the 81.0 % of the overall rigde line following algorithm execution time. The updateT function represents 10.4 % of the total execution time but it is not included in the coprocessor because it is not executed in all the executions of the loop. Moreover, the high execution time of this function is due to the elevate number of memory write accesses it performs, but nor for its complexity neither for a great number of total calls.

Table 2. Percent of overall execution time spent in the main functions of the algorithm

	#calls	%
next_point	29608	6.8
createsecc	3471	11.1
filtersecc	3471	16.5
corr_weak_max	3471	30.2
tg_dir	3526	16.4
updateT	1347	10.4
others	-	8.6

On the other hand, the coprocessor includes the which, although it is not a complex function, collects the 6.8 % of the execution time due to its great number of executions. This great number of calls is because it is also called several times during the execution of the createsecc and filtersecc functions.

Table 3 shows experimental results of the program execution in two different prototyping platforms. The first column shows the execution time in a Xilinx MicroBlaze embedded soft core, a reduced instruction set computer (RISC) optimised for implementation in Xilinx FPGAs [16]. This soft core has been implemented in a Xilinx Virtex II XC2V1000-4FG456C device (part of a Celoxica RC200 prototyping plat-

form). The second column corresponds to an implementation in a Memec design Inc. Virtex-II Pro FF672 development board. It incorporates a Xilinx XC2VP4 device, with a PPC405, a 32-bit implementation of the PowerPC embedded-environment architecture [17].

Table 3. Algorithm execution experimental results

	Execution time (s)	
	Mblaze (50Mhz)	PowerPC (100MHz)
total	5.58	2.59
9 first lines of Code 1	5.01	2.37

These execution times validate the profiler results, and show that the first nine lines of code represent about 90 % of .the overall execution time. The implementation of these functions in hardware can improve significantly the identification process.

4 The Coprocessor

The coprocessor implements a hardware version of the critical functions, corresponding to the first 9 lines of the Code 1 after the while sentence. As depicted in Fig. 4, it has been divided in six stages plus a control unit (CU). The coprocessor works with a 256*256 pixels image of 8-bit gray levels stored on a SRAM, and with fixed parameters μ=3 and σ=7 as used by the Maio-Maltoni algorithm [14]. The starting point (i_c,j_c) and the discretised angle φ_c are read from the x0,y0 and w0 ports, then the Nrst signal is asserted and the coprocessor negates Nstop until the computation is completed, returning the weak local maximum point (i_t,j_t) and the discretised angle φ_t of the dominant ridge flow direction in the ports x1,y1 and w1.

The first stage, named xynxt, computes the point (i_t,j_t) from a starting point (i_c,j_c) and angle φ_c. It corresponds to the hardware implementation of the first three lines of Code 1, after the while sentence. The second stage, named secc, implements the fourth line of Code 1 to estimate the $2\sigma+1=15$ points of the section orthogonal to the starting angle φ_c, read from port w0. The computed points of the section are stored in an internal RAM of the coprocessor, denoted as Ω.

The next stage, named filt, implements the fifth line of the Code 1 and computes the average gray values of the pixels that belong to the section Ω with the pixels that belong to the two parallel planes at a distance of ±1 pixel on the φ_c direction, denoted as Ω^{+1}, Ω^{-1}. It reads the internal coprocessor RAM that contains the pixels of the section Ω, to determine the address of pixels at Ω, Ω^{+1} and Ω^{-1}, in the external RAM, in order to compute the avg(Ω) of the $2\sigma+1=15$ points to store it into another internal coprocessor RAM.

The fourth and fifth stages, named corr and x1y1 respectively, implement the lines six to eight of the code. First the correlation of the avg(Ω) is computed as described in Section 2. And then, the fifth stage computes the pixel (i_n,j_n) and writes it on the x1,y1 ports of the coprocessor.

Finally, the last stage, named tgdr, implements the ninth line of Code1. Estimating the dominant ridge flow direction φ_n in the point (x_n,y_n), as described on Section 2.

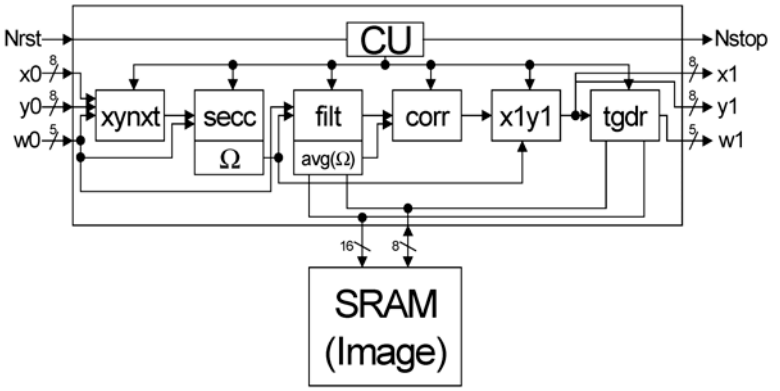


Fig. 4. Overview of the coprocessor linked with an external SRAM that stores a fingerprint image

The number of clock cycles needed to complete all six stages is the sum of clock cycles of every stages plus a few due to the CU needs, resulting a total time of $720*T_{CLK}$. The details of the implementation, as well as the estimate of the number of clock cycles needed by the coprocessor to compute, can be found in some of our previous works [4][5].

The coprocessor has been described in VHDL, and synthesized and mapped on several low cost FPGAs and an ASIC technology.

In a XILINX SPARTAN-II 2S30 FPGA, of about 30Kgates equivalent, the synthesis reports 749 CLBs (configurable logic blocks) used, 87% of all available, and a maximum clock frequency of 76MHz. The synthesis circuit for an Altera ACEX EP1K30, also of about 30Kgates equivalent, occupies 1179 LCs (logic cells) (68%) and a maximum clock frequency of 83MHz. The synthesis results for the Atmel FPSLIC AT94K shows 1.200 LUTs (look up table) occupied and a maximum clock frequency of 23MHz. The synthesis result for an ASIC SCL05U technology of 0.5µm is 15.527 equivalent gates and 97MHz of maximum clock frequency.

In the used prototyping platforms, with a XILINX VIRTEXII XC2V1000fg456 device, the synthesis reports the use of 542 CLBs, and a maximum clock frequency of 111 MHz. The execution of the steps performed by the coprocessor with a 100 MHz system clock, is of 7.2 µs, from a starting point (x0,y0) and indexed angle w0, until the next point (x1,y1) and indexed angle w1 is completed. The execution of the software version of the code takes an average time of 784 µs to execute in the PowerPC version at 100 MHz and 1.69 ms if executed in a microBlaze processor running at 50MHz. Finally, the same code running in an ARM7TDMI processor, with a 100 MHz clock, takes an average time of 211 µs to execute. The coprocessor greatly improves the speed performance when compared with the 32-bit general-purpose microprocessor, due to the pipeline scheme of the coprocessor architecture and the parallel execution of computations into its stages. The time devoted to computations mapped on the coprocessor is reduced 92% when compared with the software running in the PowerPC processor at the same clock speed.

The overall execution time of the algorithm running in the PowerPC with the coprocessor is reduced from 2.59 s to 380 ms, that is a reduction of about 85 %.

5 Conclusions and Future Work

The trend, in the prevention of unauthorised accesses or fraudulent uses, is the progressive replacement of current PIN or password-based identification methods by biometrics-based ones. Low-cost and low-response-time are the key parameters in an embedded fingerprint authentication system, as they will facilitate their integration in a wide range of new applications. Up to now, the biometrics algorithms have been developed using high level languages and complex arithmetic operations, and without thinking in the required computational resources. As a result, the biometrics identification must be performed in a high capacity platform, as the involved algorithms need a high capacity processor, not suitable to be embedded in low cost equipment.

Our work proofs that fingerprint biometrics algorithms can be implemented in low capacity microprocessors without performance loss. In a first stage, the algorithms must be optimised in order to reduce the number of floating point operations. It is demonstrated that the operations can be performed with integer numbers without significant changes in the obtained minutiae. In a second stage, a profiling of the software version locates the most critical tasks of the algorithm. These functions are implemented in hardware, with a significant improvement of the biometrics identification time (about 70 – 90 % of reduction).

In conclusion, we have demonstrated that, if the algorithms are optimised thinking in terms of hardware, an automatic fingerprint authentication system can be embedded in a low-performance-processor based system.

In the near future we will implement a complete embedded fingerprint-based-identification-system. The objective is the use of a 8 bit microprocessor with an associated dynamically reconfigurable FPGA. In this way, the associated hardware dimensions will be reduced, reconfiguring the FPGA to perform the concrete task needed at each moment.

References

1. AuthenTec FingerLoc TMS320C5509 DSP-based EDK. http://www.authentec.com
2. Biometric Information Autonomous Systems Research Group. http://bias.csr.unibo.it/ Universit di Bolognia.
3. Bioscrypt MV1200 OEM Module. http://www.bioscrypt.com
4. Canto,E.et al. "FPGA Implementation of the Ridge Line Following Fingerprint Algorithm". Proceedings of the XIX Conference on Design of Circuits and Integrated Systems, DCIS 2004.
5. Canto,E.et al. "Coprocessor of the Ridge Line Following Fingerprint Algorithm". Proceedings Field-Programmable Logic and Applications. 14th International Conference, FPL 2004. ISBN 3-540-22989-2
6. Chapel, C.E. Fingerprinting a manual of identification Coward McCann. New York 1971.
7. Cogent Systems SecurARM OEM Identification Module. http://www.cogentsystems.com
8. Donahue, M.J.; Rokhlin, S.I. On the use of Level Curves in Image Analysis Image Undestanding, vol. 57, no. 2, 1993.
9. Fingerprint Cards FPC2000 Fingerprint Processor. http://fingerprints.com
10. Galton, F. Finger Prints Macmillan. London 1892.
11. Halici, U.; Ongun G. Fingerprint Classification Through Self-Organized Feature Maps Modified to Treat Uncertainities Proceedings of the IEEE, vol.84, no. 10, October 1996.

12. IKE-1 Multifunction Controller for Image Processing Applications. http://www.ikendi.com
13. Jain, L.C. et al. "Intelligent Biometric Techniques in Fingerprint and Face recognition" CRC press. New York 2000. ISBN 0-8493-2055-0
14. Maio, D.; Maltoni, D. "Direct Gray-Scale Minutiae Detection In Fingerprints" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 19, No. 1. January 1997
15. Mehtre, B.M. "Fingerprint Image Analysis for Authomatic Identification" Machine Vision abd Applications, Springer-Verlag, Vol. 6, 1993, pp. 124-139.
16. Microblaze Processor Reference Guide. Xilinx Inc. www.xilinx.com
17. PowerPC Processor Reference Guide. Xilinx Inc. www.xilinx.com
18. Ratha, N.; Bolle, R. "Automatic Fingerprint Recognition Systems". Springer-Verlag New York Inc. 2004. ISBN: 0-387-95593-3
19. Suprema UniFinger Stand-alone Fingerprint OEM Module. http://www.suprema.co.kr
20. Watson, C.I.; Wilson, G.L. Fingerprint Database. National Institute of Standards and Technology, Special Database 4, April 18, 1992.
21. Wilson, G.L. et al. "Massively parallel neural network fingerprint classification system" NIST Tech. Rep., Advanced Syst. Div., Image Recognition Group, 1993
22. Yang, S. Et al., "A Compact and Efficient Fingerprint Verification System for Secure Embedded Devices", Proc. Of the 37th Asilomar Conference on Signals, Systems and Computers, Novermber 2003.

Fingerprint Matching Using the Distribution of the Pairwise Distances Between Minutiae

Chul-Hyun Park[1], Mark J.T. Smith[1], Mireille Boutin[1], and Joon-Jae Lee[2]

[1] School of Electrical and Computer Engineering
Purdue University, West Lafayette, IN 47907, USA
{park95,mjts,mboutin}@purdue.edu
[2] Division of Computer and Information Engineering
Dongseo University, Busan, Korea
jjlee@dongseo.ac.kr

Abstract. This paper presents an efficient minutiae-based fingerprint representation and matching method using the distribution of distances between points. The proposed method uses the distribution of pairwise distances between minutiae as fingerprint features. The fingerprint matching between the input and the template fingerprints is performed by considering the Euclidean distance between the distributions. Most conventional minutiae matching methods require intensive comparing in order to align the two fingerprints translationally and rotationally, whereas the proposed method does not need such an intensive comparison procedure for the alignment. In addition, the feature vector generated by the proposed method has a small and fixed length, which is more advantageous in some applications such as smart cards. The experiments using the randomly generated 800 minutiae sets and our database consisting of 800 fingerprints show that the proposed method can be used effectively in applications that have limited memory and require high speed.

1 Introduction

Fingerprints have been widely and successfully used as a means for human identification for more than a century. Among the various fingerprint features that have been considered, minutia points have been shown to be one of the most effective and discriminatory fingerprint features [1]. In most fingerprint-based biometric systems using minutiae information, minutiae are extracted from an input fingerprint, and then matching is performed by comparing the various attributes of the minutiae between the input and the templates. Even though the minutiae information is very compact and distinctive, minutiae matching algorithms generally involve translational and rotational alignment or minutiae pairing procedures because fingerprints are recored at different positions and angles during their capture. Perhaps the simplest approach is to perform an exhaustive search to find the optimal alignment between the input and template minutiae [2]. Other authors carry out the alignment using the Hough transform [3]. Although there are many approaches to reduce computational complexity

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 693–701, 2005.
© Springer-Verlag Berlin Heidelberg 2005

[4], most alignment procedures still require a fairly extensive search of the transform space. Further complicating matters is that often there are differences in the number of input and template fingerprint minutiae, which must be reconciled. In addition, since coordinates and their attributes (such as the ridge directions on the minutiae) need to be stored for each template, methods of this type implicitly require a high storage capacity. In this paper we propose a new minutiae-based fingerprint matching method that is based on a compact and fixed-size feature vector and also employs an efficient alignment procedure. The proposed method is based on Boutin and Kemper's reconstruction theorems [5], [6] which show that a geometrical configuration of points can be uniquely represented (accepting a small number of exceptions) by the distribution of the pairwise distances between the points. This representation has the attractive property that it is translation, rotation, and scale invariant. Motivated by these properties, we focus on representing minutiae configurations using a distribution or histogram of distances and analyzing its characteristics. The proposed method first extracts an absolute reference point, and then establishes the region within a certain distance from the reference point as a region of interest (ROI). Here, only the minutiae in the ROI are considered for the later feature extraction. Given the minutiae, the distribution of the quantized distances between all the possible minutiae in the ROI is calculated. Thereafter, the distribution is smoothed to reduce the effect of noise or nonlinear deformation, and treated as a normalized feature vector. In addition, we use the number of the minutiae in the ROI as a feature element in this vector to improve the accuracy. The rest of this paper is organized as follows. The next section describes the proposed minutiae feature representation in detail. In Section 3, the fingerprint matching method is explained. Section 4 presents the experimental results, followed by the conclusions and future work in Section 5.

2 Fingerprint Minutiae Representation Based on the Distribution of Distances

In this section, we briefly introduce Boutin and Kemper's reconstruction theorem, which underlies of the proposed method, and then describe the fingerprint feature extraction algorithm in detail.

2.1 Reconstructing Point Configurations from Distances

An n-point configuration is a tuple of points $P_1,\ldots,P_n \in \mathbf{R}^m$. To an n-point configuration we associate the Euclidean distances $d_{i,j}$ between each pair of points P_i and P_j, and then consider the distribution of distances, i.e. the relative frequencies of the value of the distances. For n fixed, the distribution of distances is given by the set of numbers $d_{i,j}$ possibly with multiplicities if some distances occur several times. So considering the distribution of an n-point configuration is equivalent to considering the polynomial

$$F_{P_i,\ldots,P_n(X)} = \prod_{1\le i,j\le n} (X - d_{i,j}) \tag{1}$$

Clearly the distribution of distances of a point configuration is invariant under permutations of the points and under rigid motion (such as rotation, translation, and reflection) of the point configuration. The question is whether an n-point configuration is uniquely determined, up to a rigid motion, by the distribution of its distances. In other words, if two point configurations have the same distribution of distances, does it mean that they are the same, up to a rigid motion? Boutin and Kemper [5], [6] have proved that the answer to this question is yes for all but a very small set of exceptional point configurations. More precisely, they have shown that the following holds.

Theorem 1 (Reconstruction Theorem). *There exists a polynomial function f in $2m$ variables such that if a point configuration $P_1,\ldots,P_n \in \boldsymbol{R}^m$ satisfies $f(P_1,\ldots,P_n) \neq 0$, then this point configuration is uniquely determined, up to a rigid motion, by the distribution of its distances.*

This theorem implies that, except for a set of measure zero of point configurations, all point configurations are uniquely determined by the distribution of their distances. So, at least in theory, almost all point configurations are uniquely reconstructable from the distribution of their distances. Even though there are a few exceptions that cannot be reconstructed, the distribution of distances has the attractive property that it is, for the most part, information preserving as well as translation and rotation invariant. Generally there is no guarantee that any feature vectors for fingerprint matching is a unique representation of the pattern, and practically, one may not be able to reconstruct the original pattern from the feature vectors. But, to the extent that the features are reasonably distinct, such an approach is potentially very attractive [7]. Fig. 1 shows two examples of n-point configurations in a plane together with their associated distributions. The work on which we report explores the application of pairwise distance distributions as feature factors for minutiae matching, motivated by the attracted invariance properties and the Boutin and Kemper's reconstruction theorem.

In this paper, we assume that minutiae information is already known and focus on how to effectively represent and match the minutiae configurations. Since there is typically great variety in the quality and position of acquired input fingerprints, we employ a reference point around which we compute the distribution. This variation in translational position can result in minutiae points falling outside of the seen region. Because of this, we detect a reference point such as core points, and compute the distributions relative to this point. Feature vectors calculated in this way have a compact form, a fixed length, and are rotationally invariant.

Feature vectors consist of 1) the number of the minutiae in the ROI and 2) the smoothed distribution of the quantized distances between all possible minutia points in the ROI.

2.2 ROI (Region of Interest) Detection

If the same minutiae sets are always obtained for the same finger, ROIs for fingerprints do not have to be detected. However, it is difficult to expect such

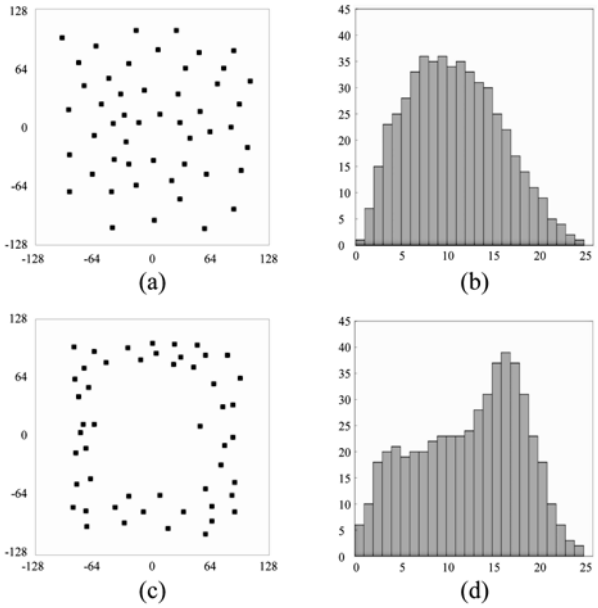


Fig. 1. (a), (c) Examples of a 50 point configuration and (b), (d) their associated distributions of pairwise distances with bin size=10

situation in real fingerprint biometric systems. Therefore, in order to solve this problem, the proposed method finds the ROI of a fingerprint using the reference point information, and uses only the minutiae in the ROI for calculation of the distribution. This approach has some disadvantages that it is not easy to locate the reference point consistently and accurately and the minutiae outside the ROI are not used. But it has the advantages that the procedure is simple and the registration can be performed by storing the compact and fixed length feature vector instead of all the minutiae coordinates.

In order to extract the reference point, the proposed method uses the Poincare index based method [8]. Then the minutiae of interest are obtained by finding the minutiae within a certain distance from the detected reference point as shown in Fig. 2(a).

2.3 Feature Vector Generation

Once the minutiae in the ROI of an input fingerprint are found, the proposed method calculates the histogram of the quantized distances between all the possible minutiae in the ROI. Let f_k be the number of the distances in the range corresponding to the k-th bin and let be the k-th value of the smoothed histogram. The k-th feature value of the feature vector, v_k, is obtained as follows:

$$v_k = M \cdot \frac{f'_k}{\sum_{l=1}^{N} f'_l}, \; k = 1, \ldots, N \tag{2}$$

$$f'_k = \begin{cases} \frac{f_k + f_{k+1}}{2}, & \text{if } k = 1 \\ \frac{f_{k-1} + f_k + f_{k+1}}{3}, & \text{else if } 2 \le k \le N-1 \\ \frac{f_{k-1} + f_k}{2}, & \text{otherwise} \end{cases} \quad (3)$$

where M and N are the normalization constant and the number of bins, respectively. Due to nonlinear deformation of fingerprints, the distances between the same minutiae pair are a bit different and the distances can be counted in a different bin, usually one of the neighboring bins. To reduce the effect of this problem, the calculated histogram is smoothed using Eq. 3. The full feature vector $V = (v_0, v_1, v_2, \ldots, v_N)$ is completed by filling the first feature value, v_0, with the number of the minutiae in the ROI. The number of minutiae is added to a feature vector in order to improve accuracy. In the experiment each bin size is empirically set to 25 pixels that are at least more than the smallest distance between minutiae. A sample normalized histogram is graphically displayed in Fig. 2(b).

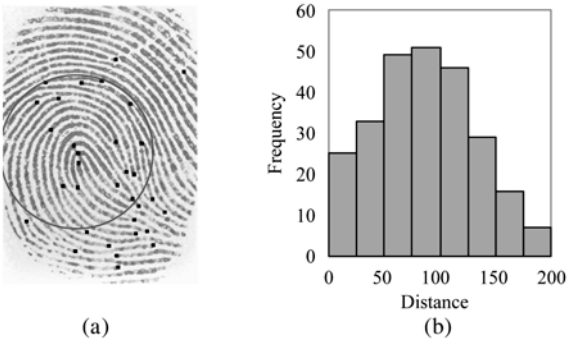


Fig. 2. The minutiae in the ROI and the normalized histogram (distribution) of a sample fingerprint. (a) Minutiae in the ROI (radius=100), (b) normalized histogram (M=255, N=8)

3 Fingerprint Matching Using Histogram-Based Feature

Since fingerprint feature vectors are obtained using the minutiae in the regions with the same area for the same finger, the number of the minutiae can be a distinctive feature. Therefore the proposed method considers the minutiae number as critical information in calculating the matching distance. As mentioned in Section 2.3, the first element (v_0) of a feature vector is the number of the minutiae in the ROI.

If we let v_k^I, v_k^T denote the input and template feature vectors, respectively, the matching distance is calculated as follows:

$$d = w \cdot \left| v_0^I - v_0^T \right| + \frac{1}{N} \sqrt{\sum_{k=1}^{N} (v_k^I - v_k^T)^2}, \; w \ge 0 \quad (4)$$

where w is a weighting coefficient and N is the number of bins. The first term of the distance measure in Eq. 4 is the weighted difference between the minutiae numbers and the second term is the Euclidean distance between the two histogram features. The larger w is used, the more the matching distance is affected by the difference between the minutiae numbers.

4 Experimental Results

To analyze the characteristics of the distributions, we used randomly generated minutiae information before using the real fingerprint minutiae. First, we selected the first image of each class of the FVC2000 2a database [9], and then manually detected the minutiae of 100 selected fingerprints. Allowing a certain amount of tolerance from these manually detected minutiae, we generated additional 700 minutiae sets using a random number generator (see Fig. 3(a)). The tolerance of 5 pixels means that the minutiae locations can be different up to 10 pixels of each other in the same class.

First, we investigated how much the bin size affects the matching result. Fig. 3(b) shows the changes in the equal error rate (EER) according to the bin size when the randomly generated sets of minutiae with the tolerances of 1 pixel and 5 pixels were used. In this experiment, we used not just the minutiae in the ROI but all the minutiae information. We can see that there is the bin size with the smallest error rate for each minutia set and the histogram-based feature tends to be less sensitive to noise if the bin size is larger.

In an actual situation, genuine minutiae might be missed and spurious minutiae might be detected during the automatic minutiae extraction stage. Therefore we need to check out how much the minutiae detection errors affect the matching result. In Table 1, we demonstrated the changes in the EER according to the difference between the minutiae numbers. The difference range of 5 means that the minutiae numbers can be different each other in the same class from 0 to 5. Since the average minutiae number of the 100 sample fingerprints used for generating random minutiae is about 32 and the fingerprints with the minutiae number less than 20 are over 12%, the proposed method is a bit sensitive to missing of minutiae. However, when the number of missed minutiae was small, the accuracy was good. The method ($w = 2$) using the number of minutiae together with the histogram feature outperformed the method ($w = 0$) using only the histogram feature.

Table 1. Changes in the EER (%) according to the difference between the minutiae numbers (bin size = 30, tolerance = 5). w=0 means that only histogram features are used for matching

Difference range	0	1	2	3	4	5
w=0	3.12	6.24	8.64	12.30	16.33	18.04
w=2	0.59	2.51	4.00	6.66	10.30	11.87

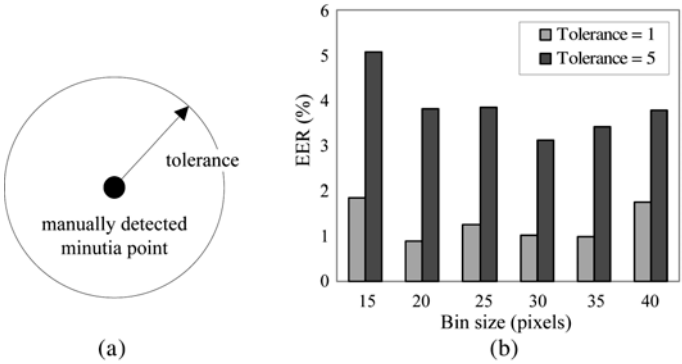


Fig. 3. (a) Tolerance range for random minutiae generation, and (b) changes in the equal error rate according the bin size

For our experiments, we also collected 800 fingerprints (80 classes) using a capacitive type low cost fingerprint sensor. The acquired fingerprint images were 256 gray scale images and 364×256 in size. In the proposed method, if the reference point of a fingerprint image is located near the image border, some minutiae that are supposed to be on the image region might not be available and this results in a bad matching result. Hence, when the subjects provided their fingerprints we guided them to place their fingers in the center of the sensor. In spite of our guidance, the reference points of some fingerprints were located near the image borders. In this work, we manually detected the minutiae of all the fingerprint images. Due to nonlinear deformation and reference point detection error, the numbers of minutiae in the ROI can be different even between the fingerprint images from the same finger. The larger radius value is used to establish the ROI, the more minutiae may be employed for calculation of a feature vector. In case that the radius is too large, the matching result can be worse because the ROIs tend to exceed the image region, which means much different minutiae sets could be compared even though the two fingerprints are from the same finger. Therefore, the radius used to establish a ROI was empirically determined as 100 pixels considering the fingerprint image resolution.

In the experiment, the EER of the proposed method was 4.51%. We plotted the ROC curve as well in Fig. 4. The accuracy of the proposed method might not be considered enough high for the situation that the fingerprint minutiae were extracted manually in the experiment, but the proposed method has several advantages over the most conventional minutiae-based methods in that it requires much less and fixed size memory space and the matching speed is so high. The methods that all the minutiae coordinates should be stored as the fingerprint feature require at least 85 bytes (actually much more memory space is needed because in general various attributes of the minutiae also are used as fingerprint features) if we assume that the fingerprint image size is 364×256 and there are 40 minutiae on the image, whereas the feature vector of the proposed method needs only 9 byte of memory space. The processing speed of the proposed method is much faster than the matching methods performing an exhaustive search for the alignment, because it is based on simple computation of the Euclidean distance.

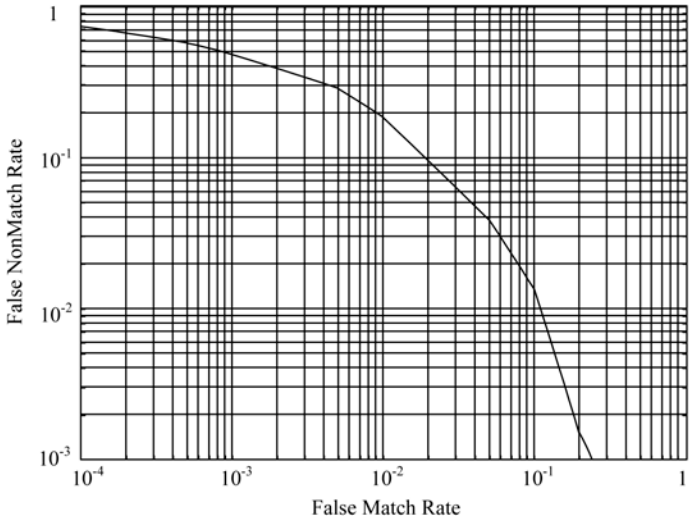


Fig. 4. ROC curve for the proposed method

5 Conclusion and Future Works

We have proposed a new fingerprint minutiae matching method using the distribution of the pairwise distances between minutiae. The minutiae configuration is easily represented as a compact and fixed length feature vector by calculating the distances between all the possible minutiae pairs in the ROI and computing their histogram. Mainly due to the intrinsic nature of the histogram-based representation, the alignment between the input and template is easily solved and the matching is performed by simply calculating the Euclidean distance between the two feature vectors. In our method, the minutiae number information is also used as a discriminatory feature to improve accuracy.

Experimental results conducted using our randomly generated 800 sets of minutiae and 800 fingerprint images show that the proposed method not only has a reasonable accuracy and a high processing speed but the feature vector generated by the proposed method has a compact and fixed length. This result demonstrates that the proposed method has a possibility that the histogram-based feature is very effective for fingerprint minutiae matching and it can be advantageous in some applications with a limited memory space. Since the proposed method exploits the distances between minutia points as a key feature, we need to develop a distance measure less sensitive to nonlinear deformation. Besides, to improve the accuracy and practicability of the proposed method, an algorithm robust to the minutiae detection errors is required to be added to the current method.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2004-042-D00151).

References

1. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of fingerprint recognition. Springer-Verlag (2003)
2. Huvavandana, S., Kim, C., Hwang, J.N.: Reliable and fast fingerprint identification security applications. Proc. Int. Conf. on Image Processing **2** (2000) 503–506
3. Ratha, N.K., Karu, K., Chen, S., Jain, A.K.: A real-time matching system for large fingerprint databases. IEEE Trans. PAMI **18**(8)(1996) 799–813
4. Udupa, R., Garg, G., Sharma, P.: Fast and accurate fingerprint verification. Proc. Int. Conf. on Audio- and Video-Based Person Authentication (3rd)(2001) 192–197
5. Boutin, M., Kemper, G.: Which point configurations are determined by the distribution of their pairwise distances. Submitted to Int. J. Comp. Geom. and Appl. (2004)
6. Boutin, M., Kemper, G.: On reconstructing n-point configurations from the distribution of distances or areas. Adv. Appl. Math. **32**(2004) 709–735
7. Jain, A.K., Pankanti, S., Prabhakar, S., Hong, L., Ross, A., Wayman, J.L.: Biometrics: A grand challenge. Proc. Int. Conf. on Pattern Recognition, Cambridge, UK, **2** (Aug. 2004) 935–942
8. Karu, K., Jain, A. K.: Fingerprint classification. Pattern Recognition. **29** (3) (1996) 389–404
9. Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., Jain, A.K.: FVC2000: Fingerprint verification competition, IEEE Trans. Pattern Anal. Mach. Intell. **24** (3) (2002) 402–411

A Layered Fingerprint Recognition Method

Woong-Sik Kim and Weon-Hee Yoo

Department of Computer Science & Information Engineering Inha University
#253 YongHyun Dong, Nam Ku, Incheon, Korea
ADELAIA@hitel.net, whyoo@inha.ac.kr

Abstract. This thesis is on fingerprint recognition method and system, more in detail, the layered fingerprint recognition method and system to compare not only the minutiae or singular point, but also compare the images of detailed area for making prompt and accurate comparison of fingerprint. Since Edward R. Henry established the modern fingerprint approach in the recent days, the fingerprint has been applied in numerous fields. In particular, fingerprint is a powerful personal verification means that has been widely used for financial transactions, crime investigation and various security systems [1,2]. The contents proposed in this thesis calculate the relativity between the minutiae and extract certain area for the standard on similar patterns in measuring the similarity that it provides effective identification on the consistency on fingerprint with similar minutiae and fingerprint with prompt and rotation in fabric [9].First, the singular point and the minutiae are compared to seek the similar pattern promptly, then compare with the image of certain area based on the singular point (core point or delta point) that more accurate fingerprint can be recognized by performing the matching processes in multi-stages. Through the experiment that applies this theory, we have confirmed that there is a great improvement of verification rate and erroneous recognition rate of other person's fingerprint.

1 Introduction

The process of verification by using fingerprint is largely divided into the classification of several fingerprints into different shapes and the procedure of matching for the subject person. In addition, such an individual verification system of fingerprint is further divided into the identification system of one to multiple number to distinguish the inputted fingerprint from the registered fingerprint and the verification system that contrast and distinguish the inputted fingerprint with the registered fingerprint on one to one basis [3].

Analyzing the fingerprint used for such a fingerprint recognition system, fingerprint exists in many number of singular point area on top of the normal area made up in fingerprint ridge with the accurate direction. From such a singular point, the point that progresses with the ridge but disconnected is called as ending point the point where the ridge is divided is called as bifurcation, and this is collectively referred to as the minutiae. Also, for fingerprint, there is singular point such as the core point, the center point of the flow of fingerprint ridge, and the delta point that the vertical ridge and the horizontal ridge are met and generated [4,6,8].

In general, there are approximately 100~150 minutiae on a finger, and each person has different type, location and direction. Therefore, this type of location and direction of minutia may be used as a way of deciding for fingerprint [5].

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 702–709, 2005.
© Springer-Verlag Berlin Heidelberg 2005

However, the previous fingerprint recognition method that used the singular point or the minutiae was the method using the coefficient relationship between the minutiae adjacent or the position that was dispersed in space between the minutiae that depended completely on the extraction of the minutiae that the rate of error in recognition became larger when the minutiae was mistakenly extracted or having similar patterns. In other words, in most of fingerprint recognition system, the type of information for minutiae and the ending point and bifurcation due to the error arising during the filtering process of the image that, in actuality like Fig. 1, it is frequent cases of erroneous verification of different fingerprint to recognized as the same fingerprint.

This thesis is on the fingerprint comparison method that it not only compares the generally used minutiae but also is a layered fingerprint recognition method that may compare the prompt and accurate comparison of fingerprint under the foundation of image foundation comparison of the detailed area.

The previous fingerprint recognition method that used the singular point or the minutiae was the method using the coefficient relationship between the minutiae adjacent or the position that was dispersed in space between the minutiae that depended completely on the extraction of the minutiae that the rate of error in recognition became larger when the minutiae was mistakenly extracted or having similar patterns

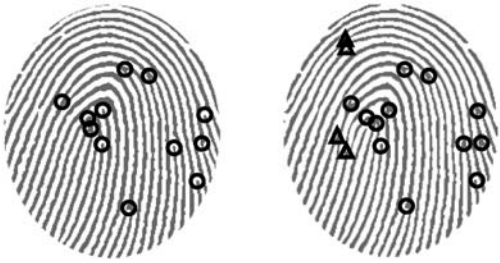


Fig. 1. Actual example of fingerprint image matched in similar pattern

2 Main Context

In most systems, the two fingerprints are deemed to be the same person since the type information of minutiae in the ending point and bifurcation due to the error arising from the filtering process of the image. However, looking into the detailed area, the minutiae patterns of these two images are different, particularly, the ridge pattern of the center area is different.

This thesis calculates the relativity between the minutiae and extracts certain area for the standard on similar patterns in measuring the similarity that it provides effective identification on the consistency on fingerprint with similar minutiae and fingerprint with prompt and rotation in fabric.

The fingerprint matching process of the fingerprint recognition system under this thesis is shown on Fig.2.

First, the fingerprint image is extracted from the fingerprint sensor. For a sensor that captures the image, there are optical method and semiconductor type area sensor and optical method and thermal sensitive type linear sensor.

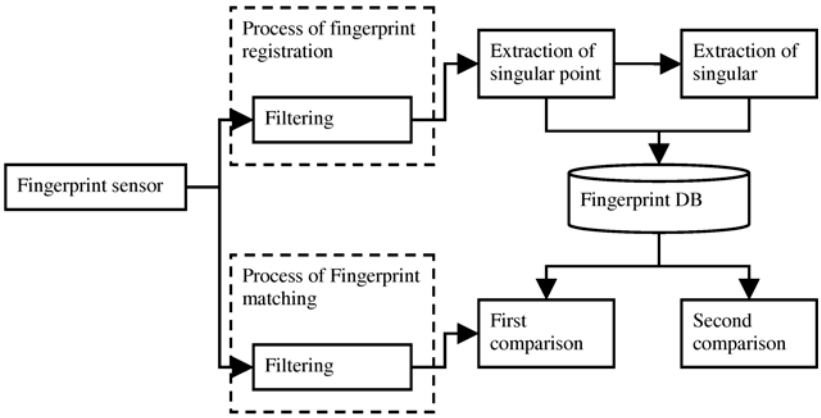


Fig. 2. Introduction Chart for fingerprint recognition system

The quality of fingerprint image captured from the fingerprint sensor may have a significant difference for gender, age and skin composition of individual depending on the sensor. Therefore, there is a need of filtering process to make entire quality of image consistent and restructuring the part with disconnected ridge or crushed part by using the direction of the ridge. The output of this process is the B/W image. The next stage is the minutiae extraction stage that the layered extraction of singular point information is made. This information is used as the information of layer comparison in the process of the fingerprint matching.

In the extraction stage for singular point and minutiae, the core point, the center of flow for fingerprint ridge and the delta point where the vertical ridge and the horizontal ridge are met and generated.

The method of extracting the singular point is to extract the direction map on the fingerprint ridge and then search for the point where the patterns of core and delta are displayed. The pattern for core and delta can be found by the search of the vector circle [5].

As shown on Fig.3, the core point is shown the direction to the vertical side on the arc to a direction from the vector circle. The delta point shows the same direction with the vector circle.

The core point is indicated as S1 (X, Y and A) (X indicates the coordinate of x-axis, Y indicates the coordinate of y-axis, and A is the direction angle of the minutiae), and the delta point is indicated as S2 (X, Y, A1, A2 and A3). (X indicates the coordinate of x-axis, Y indicates the coordinate of y-axis, and A1, A2 and A3 are the direction angle of the point that merges the ridge.)



Fig. 3. Search process of core



Fig. 4. Search process of delta

For the method of extracting the minutiae, there are the method by the ridge tracing and the method of using the 3x3 matrix through thinning, and this patent has applied the method of extracting the 3x3 matrix through thinning. Each minutia is indicated as T (X, Y and A). (X indicates the coordinate of x-axis, Y indicates the coordinate of y-axis, and A is the direction angle of the minutiae.)

The last stage is the minutiae location area extraction stage that certain area is captured on the basis of the singular point, particularly, the core point.

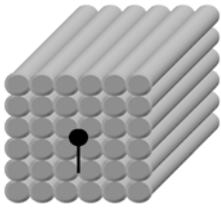


Fig. 5. Extract certain areas based on the core point

In the thesis, B/W image 64x64pixel = 4096pixel is extracted based on the singular point. When this singular point data is indicated for each byte, it has 4096 bytes that it causes the slow down in the speed of the entire system that each 8-pixel is packed for 1 byte to store on the fingerprint DB.

In the fingerprint matching stage, as in the fingerprint registration stage, the singular point and minutiae, minutiae location area data are extracted through the filtering process on the fingerprint image inputted. And, after reading the singular point and minutiae data of fingerprints subject for comparison under Comparison I, arrange each minutiae set based on the location of singular points to compare and then determine the consistency of each. If Score 1 is certain standard or more, then read the minutiae area data of subject compared in Comparison II and reflect the skin deformation rate, the distortion rate between the singular points from the Comparison I to change the area data and inspect for consistency with the minutiae area data inputted. If Score 2 is certain standard or more, it recognizes as the same person.

The detailed flow charts of the above processes for fingerprint registration and the process for fingerprint matching are shown on Fig. 6 and Fig. 7.

With the reference to Fig 6, in the acquisition stage of fingerprint image, the fingerprint sensor is used to acquire the fingerprint image.

The quality of fingerprint image capture from the fingerprint sensor has a significant difference depending on the characteristics of sensor, gender, age, individual skin and others of the user. Therefore, Therefore, in the filtering process, it makes the

entire quality of images consistent and restructures the disconnected ridge or crushed part by using the direction of ridge. The image outputted at the filtering stage is the B/W image.

In the extraction stage of the singular point and the minutiae, it first extracts the core point, the center of the flow for the fingerprint ridge and the delta point where the vertical ridge and the horizontal ridge are met and generated. As described earlier, the method of extracting the singular point is to extract the direction map on the fingerprint ridge and then search for the point where the patterns of core and delta are displayed. The pattern for core and delta can be found by the search of the vector circle [5].

As shown on Fig. 2 and Fig. 3, the core point is shown the direction to the vertical side on the arc to a direction from the vector circle, and the delta point shows the same direction with the vector circle. The minutiae are found by using the 3x3 matrix through thinning.

The minutiae and the singular point are expressed as follow and stored on the fingerprint DB. Another word, the core point is indicated as S1 (X, Y and A) with X indicating the coordinate of x-axis, Y indicating the coordinate of y-axis, and A being the direction angle of the minutiae, and the delta point is indicated as S2 (X, Y, A1, A2 and A3) with X indicating the coordinate of x-axis, Y indicating the coordinate of y-axis, and A1, A2 and A3 being the direction angle of the point that merges the ridge.

Then based on the singular point, the area for "64 pixel x 64 pixel = 4096 pixel" is extracted and it is stored in the fingerprint DB. When this singular point data is indicated for each byte, it has 4096 bytes that it causes the slow down in the speed of the entire system that each 8-pixel is packed for 1 byte to store on the fingerprint DB.

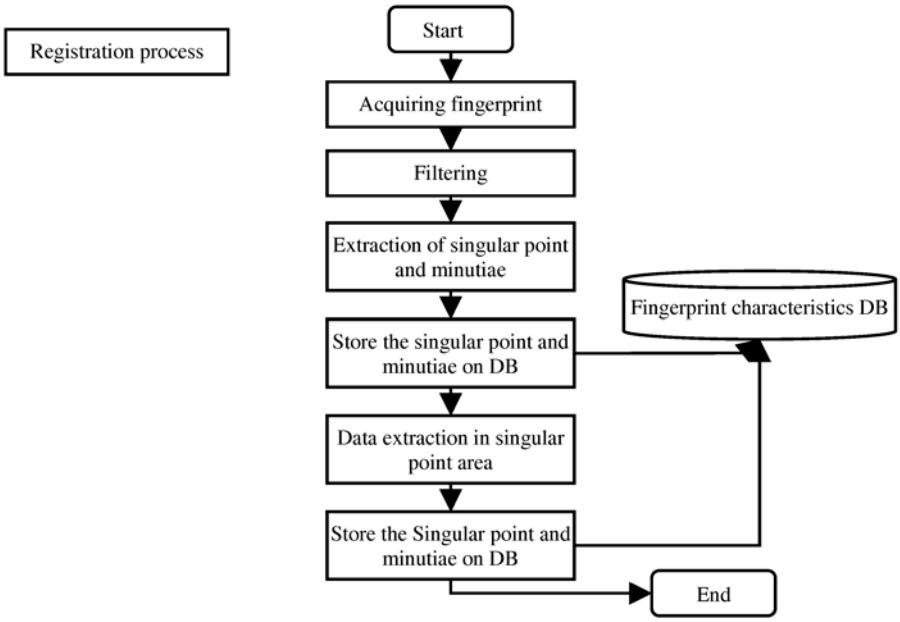


Fig. 6. Process of fingerprint registration

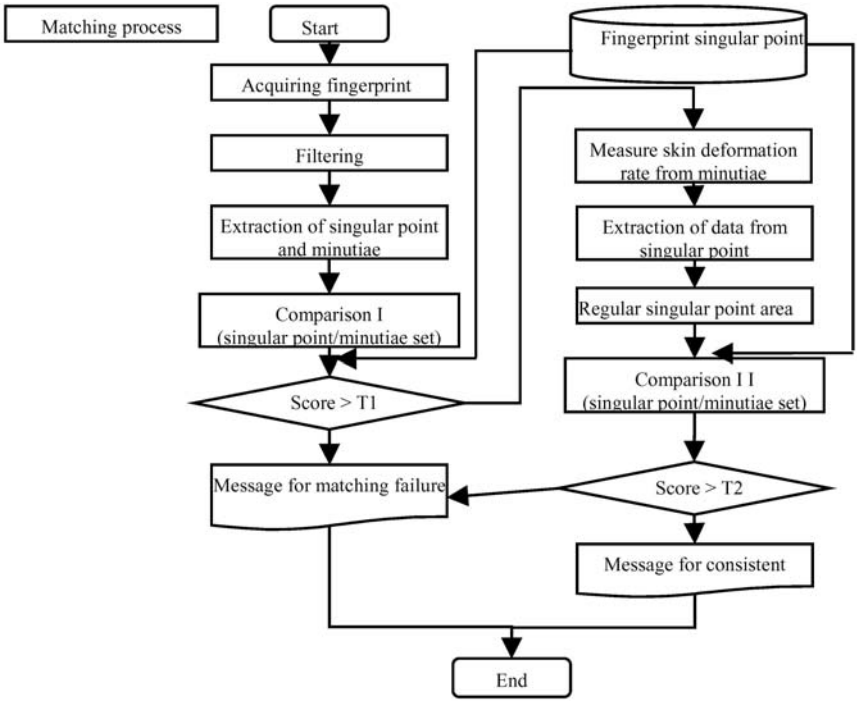


Fig. 7. Process of fingerprint matching

Referring to Fig. 7, the In the fingerprint matching process, as in the fingerprint registration process, the singular point and minutiae, minutiae location area data are extracted through the filtering process on the fingerprint image inputted.

And under the Comparison I, after reading the singular point and minutiae data of fingerprints subject for comparison from the fingerprint DB, arrange each minutiae set based on the location of singular points to compare and then determine the consistency of each.

As the result of reading, if the Score 1 is certain standard (T1) or better, it turns over to the next stage, and if Score 1 is less than T1, it sends the message of matching failure.

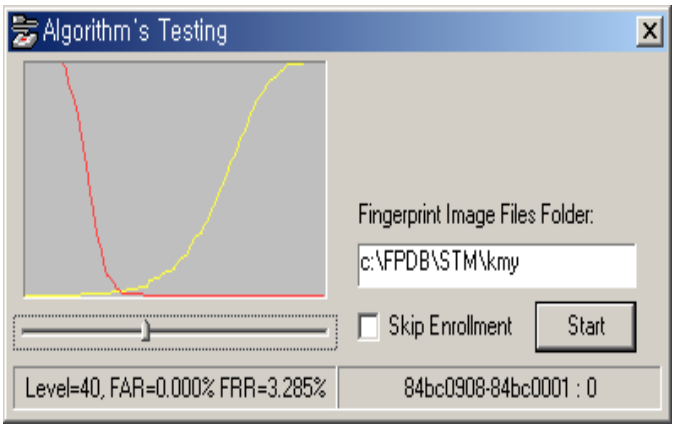
If Score 1 is certain standard or more, then read the minutiae area data of subject compared in Comparison II and reflect the skin change rate, the distortion rate between the singular points from the Comparison I to change the area data and inspect for consistency with the minutiae area data inputted. If Score 2 is certain standard or more, it recognizes as the same person. As the result of examination, if the Score 2 is certain standard (T2) or better, it recognizes as the same person, and if Score 2 is less than T2, it sends the message of matching failure.

3 Experiment

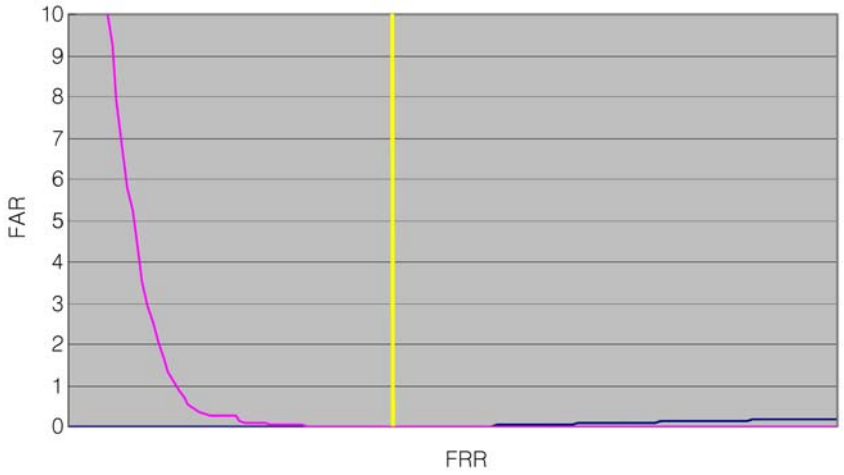
The result of the fingerprint DB test (compare to the before and after the algorithm improvement) is as follows. * EER is the equal error rate with the average figure of

FAR and FRR that it measures the degree of error for algorithm. In the actual system, choose the threshold value of FAR=0, rather than selecting the threshold value with the minimum figure of EER in a way of minimizing the erroneous recognition rate.

1) Previous algorithm
   - EER = 1.216 % (when FAR = 0.608% and FRR = 1.825%)



2) Improved algorithm
   - EER = 0.0039% (when FAR = 0.0% and FRR = 0.0077%)



4 Conclusion

In the comparison of the existing algorithm and the improved algorithm, we can see clearly better result. The test fingerprint DB is 80 images from the thumb to the little finger that have been used for actual verification, and the previous algorithm showed a drop in the verification rate for the little finger, but the improved algorithm show a great result in the fingerprint DB even for the ones that included the little finger.

In conclusion, considering the fact that the capability difference of the two algorithms under the ST Bench Marking Test was low while the EER of algorithm that are highly ranked in test results of companies from the fingerprint verification competition (FVC) is around 1~2%, the improved algorithm can be commercialized. In addition, a marked improvement has been made compared with the previous algorithm.

References

1. Dario Mario, and David Maltoni, “Direct Gray-Scale Minutiae Detection In Fingerprints”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 1,, pp. 27-39, Jan., 1997.
2. Lin Hong, Yifei Wan, and Anil Jain, “Fingerprint Image Enhancement: Algorithm and Performance Evaluation”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20,no. 8,pp. 777-789,Aug., 1998.
3. A.K.Jain, L. Hong, R. Bolle, “On-Line Fingerprint Verification”, IEEE Trans., Pattern Analysis and Machine Intelligence, vol. 19,no. 4,pp. 302-313, April., 1997.
4. A.K. Jain, S.Prabhakar, L. Hong, and S. Pankanti, “Filterbank-based fingerprint matching”, IEEE Trans. On Image Processing. 9(5):846-859, May 2000.
5. Nalini K, and A.K. Jain, “A Real-Time Matching System for Large Fingerprint Database”, IEEE Trans., Pattern Analysis and Machine Intelligence, vol. 18,no. 18,pp. 799-813, Aug., 1996.
6. Xudong Jiang and Wei-Yun Yau, “Fingerprint Minutiae Matching Based on the Local and Global Structures”, International conference on pattern Recognition(ICPR’00)-Vol 2.,no. 18,pp. 1038-1042, Sep., 2000.
7. T. MAEDA, M. Matsushida and K. Sasakawa, “Identification Algorithm Using a Matching Score Matrix”, IEICE Trans. INF. & SYST., Vol.E-84-D,no. 7, pp. 819-824, July 2001.
8. Xudong Luo, Jie Tian and Yan Wu, “A Minutia Matching algorithm in Fingerprint Verification**”**, International conference on pattern Recognition (ICPR’00)-Vol 4., no. 18,pp. 4833-4841, Sep., 2000.
9. J. D. Stoszand L. A. Alyea, "Automated Systems for Fingerprint Authentication Using Pores and Ridge Structure," Proceedings of SPIE, Automatic Systems for the Identification and Inspection of Humans (SPIE Vol. 227 7), San Diego, 1994, p. 210-223.

Super-template Generation Using Successive Bayesian Estimation for Fingerprint Enrollment

Choonwoo Ryu, Youngchan Han, and Hakil Kim

School of Information & Communication Engineering, Inha University
Biometrics Engineering Research Center, Korea
{cwryu,ychan,hikim}@vision.inha.ac.kr

Abstract. This paper proposes an algorithm for generating a super-template from multiple fingerprint impressions in fingerprint enrollment for the purpose of increasing recognition accuracy. The super-template is considered as a single fingerprint template which contains highly likely true minutiae based on multiple fingerprint images. The proposed algorithm creates the super-template, in which the credibility of each minutia is updated by applying a successive Bayesian estimation (SBE) to a sequence of templates obtained from input fingerprint images. Consequently, the SBE assigns a higher credibility to frequently detected minutiae and a lower credibility to minutiae that are rarely found from the input templates. Likewise, the SBE is able to estimate the credibility of the minutia type (ridge ending or bifurcation). Preliminary experiments demonstrate that, as the number of fingerprint images increases, the proposed algorithm can improve the recognition performance, while keeping the processing time and memory storage required for the super-template almost constant.

1 Introduction

In general, a fingerprint verification algorithm includes feature extraction and matching processes. Feature extraction collects a set of features from a fingerprint image, while the matching process makes the decision as to whether the two feature sets are from the same finger or not. Although a lot of research has been done on feature extraction[1–3], the results of this process are susceptible to be affected by various conditions, such as the condition of the surface of the scanner and the human skin, and the pressure imparted in making the impression. For example, extracted minutiae contain not only genuine minutiae, but also dropped minutiae (true minutiae not extracted by the algorithm), spurious minutiae (false minutiae created by the algorithm), or altered minutiae (true minutiae with the wrong type).

Hence, most feature extraction processes entail the removal of spurious and altered minutiae[4–6]. Even though they are able to remove false minutiae, minutia-removing algorithms generally cannot recover any dropped minutiae. Moreover, their incomplete removal rules or prefixed threshold values can eliminate some genuine minutiae. In order to overcome these limitations in false

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 710–719, 2005.
© Springer-Verlag Berlin Heidelberg 2005

minutiae removal and true minutiae recovery based on a single fingerprint image, multi-impression algorithms[7–10] have been proposed for the purpose of creating a template(s) from multiple fingerprint images.

These multi-impression algorithms can be distinguished by the information source that is used, i.e. images or features. Firstly, image-based algorithms utilize minutiae information to compute a transformation matrix that defines the spatial relationship among multiple impressions. Jain et al.[7] and Lee et al.[8] determined the parameters of the matrix using minutiae from a pair of impressions in their image fusion steps. Jain et al. modified an iterative closest point algorithm, while Lee et al. suggested a distance map matching method for fine alignment. These algorithms are expected to provide improved fusion performance by using both image and minutiae information. On the other hand, the computational cost and time are augmented due to the increase in the amount of data.

Secondly, feature-based multi-impression algorithms can be expected to produce fast fusion results due to their requiring less memory and computation. Toh et al.[9] and Jiang et al.[10] improved the credibility of the minutiae by applying fusion techniques only after successful genuine matching, where the minutiae in the enrolled template are updated on the basis of the input template. These works are a special case of the feature-based algorithms, because they consider only two templates, the enrolled template and the template input during the authentication operation.

Meanwhile, multi-impression selection is a widely applied method in fingerprint enrollment, because of its simple logic and easy implementation, which involves selecting K templates from N input fingerprints ($K \leq N$). A case study presented by Jain et al.[11] applies a selection scheme for choosing the K templates out of N inputs and shows better performance than random selection.

In this paper, we propose a feature-level fusion algorithm for enrolling multi-impressed fingerprints, in which a single super-template is generated from a set of multiple templates. Unlike previous works[9, 10] on feature-level fusion, this study focuses on enrollment processes dealing with an arbitrary number of input multiple impressions. The proposed algorithm consists of input image selection and template fusion. By using a similarity measure based on matching scores, the image selection process chooses appropriate input images effectively, which consequently improves the recognition performance. Then, the template fusion process makes use of a successive Bayesian estimation (SBE) method[12, 13], in order to update the credibility of each minutia based on a sequence of input fingerprints, where the number of input fingerprints is unlimited. The detailed processes of the algorithm are described in sections 2 and 3. The simulation results are shown in section 4 and conclusions are drawn in the last section.

2 Fingerprint Image Selection

We assume that K fingerprint images belonging to the same finger are captured by a fingerprint scanner and that all of them are used for the process of

template fusion. Providing that all of the input images are feature extractable, the proposed fingerprint image selection algorithm examines whether the input fingerprint images are acceptable or not.

The evaluation measure of the image selection algorithm is based on the similarity between two input templates. In this study, similarity is measured by a matching algorithm, which was previously used in a conventional authentication system using single impression. Higher similarity implies more redundancy in the feature information, which implies that only a small amount of new information can be obtained by fusing these two input templates. On the other hand, the fusion of two templates having lower similarity produces a large amount of new information. In this case, it should be noticed that some of these pairs of input templates may not be able to be correctly fused or may correspond to different fingerprints that were mistakenly acquired during the input process, due to human error.

The total number of all genuine matching cases among K fingerprint images is ${}_KC_2$. We define the similarity, s_{ij}, as the matching score between the ith and jth template, and the average similarity, S_i, as the average value of all possible s_{ij} for a fixed value of i. The similarity value s_{ij} has a scalar value, in the range of zero to one. The value $s_{ij} = 0$ means that no similarity exists between the templates, while $s_{ij} = 1$ indicates that the fingerprints contain perfectly identical minutiae information.

The ith fingerprint image becomes the removal candidate which may be rejected from the input list, if it satisfies any of the following conditions. Firstly, any of its similarity values s_{ij} is lower than the given threshold value th_{diff}. This means that the templates will be rejected if they do not contain sufficient information for updating the template, even if they are from the same finger. When several removal candidates exist, the rejected fingerprint is determined by comparison of the average similarity value S_i. The fingerprint image which has the minimum average similarity is removed from the input image list.

Secondly, the fingerprint image will also be removed from the input image list, if any of the matching scores s_{ij} is bigger than the threshold, th_{same}, and its average value S_i is the maximum value among the removal candidates. In this case, the fingerprint is assumed that the template do not have enough new information.

The similarity checks refer to the comparison steps between the matching score and the threshold values. If any fingerprints are rejected by the similarity check, new fingerprint images will be acquired from the fingerprint scanner. For each eliminated image, a new image will have to be scanned, and the similarity check process will be repeated until the selection condition is fulfilled.

3 Super-template Generation

A super-template implies a superior feature set which is generated from multiple fingerprint images by combining information from different inputs. The fingerprint feature set used in this study is a set of minutiae information, and the super-template so obtained contains minutiae not only of larger areas, but also

with higher credibility than that of the template obtained from any individual input image. The super-template is created under the condition of an unknown number of input images, by utilizing a sequential updating method.

In Fig. 1, the images on the upper row show the templates obtained from each input image, and the images on the lower row visualize the process of updating the super-template. The information contained in the super-template is improved in the template fusion process, during which the credibility of each minutia in the latest super-template is updated on the basis of the minutiae in the current input image. The darker the area in the super-template, the higher the credibility of the minutiae obtained by the fusion of the information contained in common regions of the input images.

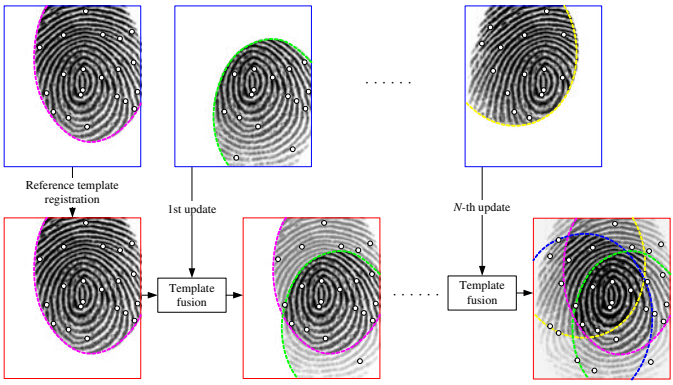


Fig. 1. Sequential updating for super-template generation

The super-template generation algorithm consists of 3 steps, namely the selection of the reference template and updating order, template fusion, and super-template creation. The template having the maximum average similarity is chosen as the first template, and is called the *reference template*. This is the template that has the biggest area in common with the other templates. The updating order is determined by the similarity between the *reference template* and the other templates. Those templates having higher similarity values will be used in the updating process earlier than those templates having smaller values.

The minutiae of the inputs are registered in the *updating minutiae set* $\mathbf{u} = \{u_i\}$. Here, $\mathbf{u}$ is a set of registered minutia u_i, where

$$u_i = (x_i, y_i, \theta_i, \zeta_i, p_{m_i}, p_{\zeta_i}, hit_i) \tag{1}$$

is the minutia vector describing the location (x_i, y_i), the direction θ_i, and the type ζ_i. Further, the probabilities, p_{m_i} and p_{ζ_i} are the credibility of the minutia u_i and its type, respectively, while hit_i is the number of occurrences of the corresponding minutia in the inputs. The segmentation information of the nth input image, seg_n, is acquired using our feature extraction algorithm in the form of a chain code representation [14]. The segmentation information of inputs, $seg = \{seg_n\}$, are also kept and used when the minutia credibility p_{m_i} is estimated.

During the first update, there are no minutiae in the *updating minutiae set* $\mathbf{u}$. Therefore, the minutiae in the *reference template* are registered as initial probabilities, as described in Eq. 2.

$$u_i = (x_i^{(ref)}, y_i^{(ref)}, \theta_i^{(ref)}, \zeta_i^{(ref)}, p_{m_{INIT}}, p_{\zeta_{INIT}}, 1) \tag{2}$$

where $p_{m_{INIT}}$ and $p_{\zeta_{INIT}}$ are the initial probabilities of the true minutia and the true type, respectively. These initial probabilities are highly related to the performance of the feature extraction algorithm. Thus, their values were determined by the evaluation of the feature extraction algorithm. The number of occurrences hit_i is set to 1 when the *reference template* is registered, and the segmentation information of the reference fingerprint image is set to the value of seg_0.

Figure 2 represents the process in which the information of the *reference template set* $\mathbf{u}$ is updated by the nth input template $\mathbf{t_n}$. The pose transformation parameter between the templates is determined by a matching algorithm. The input template $\mathbf{t_n}$ and its segmentation information are transformed by the pose transformation parameter. Supposed that $\mathbf{t_n^*}$ is the transformed template obtained from $\mathbf{t_n}$. There are three cases of correspondence between the minutiae in templates $\mathbf{u}$ and $\mathbf{t_n^*}$, namely minutiae having corresponding minutia, minutiae only existing in template $\mathbf{u}$ and minutiae only existing in template $\mathbf{t_n^*}$.

Our proposed *SBE* method is applied in all three cases with different conditions. The *updating minutiae set* $\mathbf{u}$ includes all of the minutiae obtained after updating all of the inputs $\mathbf{t_n}(n = 1, \ldots, K-1)$. The super-template selectively contains higher credibility minutiae in the *updating minutiae set* $\mathbf{u}$.

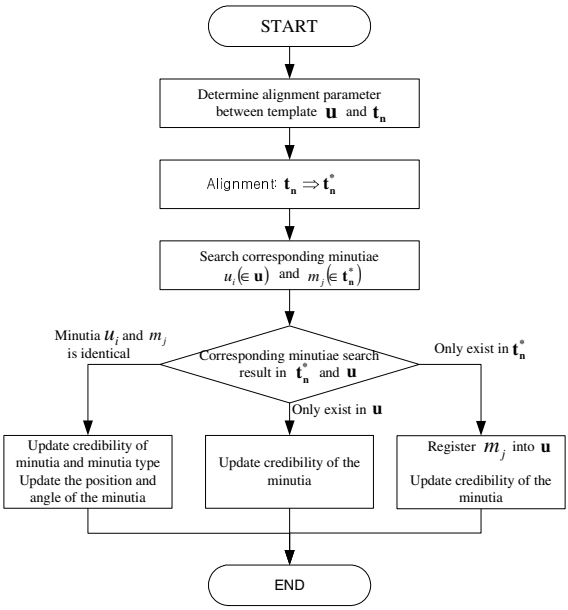


Fig. 2. Flowchart of the template update

In determining the pose transformation parameter between the templates $\mathbf{u}$ and $\mathbf{t_n}$, the matching algorithm produces several candidates and estimates the true parameters by matching their scores, which involves the repetition of the same routine. In this study, a fast parameter estimation algorithm is used to determine the value of the parameter [15]. This algorithm estimates the value of the parameter by using the Parzen density estimation method. By applying the pose transformation parameter, the nth input template $\mathbf{t_n}$ is transformed to $\mathbf{t_n^*}$ and its segmentation information is transformed and registered to seg_n. This segmentation information is applied when the minutiae are updated.

Suppose that $C(\cdot)$ represents a function deciding whether a pair of the minutia $u_i(\in \mathbf{u})$ and $m_j(\in \mathbf{t_n^*})$ are identical or not. We utilize the Bayesian probability modeling algorithm, which has a different distribution according to the type observation [16]. By applying all cases of minutia $u_i(\in \mathbf{u})$ and $m_j(\in \mathbf{t_n^*})$ to function $C(\cdot)$, the minutiae in the templates $\mathbf{u}$ and $\mathbf{t_n^*}$ are assigned to one of the three cases, i.e., minutiae having corresponding minutia, minutiae only existing in template $\mathbf{u}$, and minutiae only existing in template $\mathbf{t_n^*}$. In the first two cases, the credibility of the minutia is updated by the proposed *SBE*, while, in the last case, a new minutia is registered into the *updating minutiae set* $\mathbf{u}$. Based on the nth observation of the minutia u_i, its credibility is updated by the following SBE rule:

$$\begin{aligned} p'_{m_i} &= p\left[g\left(u_i\right) = TRUE | b_n\left(u_i\right)\right] \\ &= \frac{p\left[b_n(u_i)|g(u_i) = TRUE\right] p\left[g(u_i) = TRUE|b_{n-1}(u_i)\right]}{\sum\limits_{g(u_i)} p\left[b_n(u_i)|g(u_i)\right] p\left[g(u_i)|b_{u-1}(u_i)\right]} \end{aligned} \tag{3}$$

where the status of u_i is given as

$$g(u_i) = \begin{cases} TRUE & \text{if } u_i \text{ is true minutia} \\ FALSE & \text{if } u_i \text{ is false minutia} \end{cases} \tag{4}$$

and the existence status in the nth template $\mathbf{t_n^*}$ is defined as

$$b_n(u_i) = \begin{cases} CM & \text{if corresponding minutia of } u_i \text{ is found in this template} \\ FG & \text{if } u_i \text{ is not } CM \text{ but locates in the foreground region} \\ BG & \text{if } u_i \text{ locates in the background region} \end{cases}. \tag{5}$$

The new credibility of minutiae u_i is calculated by updating previous minutiae credibility with nth observation status $b_n(u_i)$. As found in Eq. (5), The observation status has three cases. The probability $p\left[b_n(u_i) = BG|g(u_i)\right] = 0.5$ because no evidence is discovered from the background region. However, the probabilities $p\left[b_n(u_i) = CM|g(u_i)\right]$ and $p\left[b_n(u_i) = FG|g(u_i)\right]$ can be estimated by the evaluation of feature extraction and matching algorithm performance.

By the same token as in Eq. (3), the credibility of the type of u_i is updated as

$$p_{\zeta_{NEW}} = p\left[t(\zeta_i) = CT|k_n(\zeta_i, \zeta_j)\right]$$

$$= \frac{p\left[k_n(\zeta_i,\zeta_j)|t(\zeta_i = CT)\right] p\left[t(\zeta_i) = CT|k_{n-1}(\zeta_i,\zeta_j)\right]}{\sum_{t(\zeta_i)} p\left[k_n(\zeta_i,\zeta_j)|t(\zeta_i)\right] p\left[t(\zeta_i)|k_{n-1}(\zeta_i,\zeta_j)\right]} \tag{6}$$

where the status of the type ζ_i is defined as

$$t(\zeta_i) = \begin{cases} CT \text{ if } \zeta_i \text{ is correct minutia type} \\ IT \text{ if } \zeta_i \text{ is incorrect minutia type} \end{cases} \tag{7}$$

and the correspondence status between the type ζ_i of u_i from the *reference template* and the type ζ_j of m_j from the nth template $\mathbf{t_n^*}$ is defined as

$$k_n(\zeta_i,\zeta_j) = \begin{cases} ST \text{ if the same type} \\ DT \text{ if differenct type} \end{cases}. \tag{8}$$

All the probabilities used in Eq. (6) can be estimated by the evaluation of the fingerprint recognition algorithm. By utilizing Eqs. (3) and (6), the credibility of the minutia and minutia type can be estimated as following:

(Case 1) minutiae having corresponding minutia

The information of minutia u_i is updated by the corresponding minutia m_j. The minutia position and direction are calculated by means of the following equation:

$$(x_i', y_i', \theta_i') = \left(\frac{x_i hit_i + x_j^*}{hit_i + 1}, \frac{y_i hit_i + y_j^*}{hit_i + 1}, \tan^{-1}\left(\frac{\sin(\theta_i) hit_i + \sin(\theta_j^*)}{\cos(\theta_i) hit_i + \cos(\theta_j^*)} \right) \right) \tag{9}$$

Equation (9) calculates the new position and direction by the arithmetic mean. The credibility of the minutia u_i is estimated by Eq. (3) with the condition $b_n(u_i) = CM$. The credibility of the minutia type $p_{\zeta_{NEW}}$ is calculated by Eq. (6) and the new type and its probability are defined as

$$(\zeta_i', p_{\zeta_i}') = \begin{cases} (\zeta_i, p_{\zeta_{NEW}}) & \text{if } p_{\zeta_{NEW}} \geq 0.5 \\ (\zeta_j, 1 - p_{\zeta_{NEW}}) & \text{otherwise} \end{cases}. \tag{10}$$

This implies that the minutia type can be changed to another type when the probability is less than 0.5, because there are only two possible types: ridge ending and bifurcation. In the case where the type is changed, the probability of the new type will be $1 - p_{\zeta_{NEW}}$.

(Case 2) minutiae only existing in the *updating minutiae set* u

The credibility of the minutia u_i can be updated when the minutia only exists in the template u. The credibility is estimated by Eq. (3) with the condition of the segmentation information of the nth input image. The other information is preserved, as shown in Eq. (11).

$$(x_i', y_i', \theta_i', \zeta_i', p_{\zeta_i}', hit_i') = (x_i, y_i, \theta_i, \zeta_i, p_{\zeta_i}, hit_i) \tag{11}$$

(Case 3) minutiae only existing in the template $\mathbf{t}_\mathbf{n}^*$

A minutia only existing in template $\mathbf{t}_\mathbf{n}^*$ is registered in the *updating minutiae set* $\mathbf{u}$. When registering this minutia to the template $\mathbf{u}$, all of the information belonging to the minutia m_j is preserved, except for the credibility of the minutia, as shown in Eq. (12):

$$(x_i', y_i', \theta_i', \zeta_i', p_{\zeta_i}', hit_i') = (x_j^*, y_j^*, \theta_j^*, \zeta_j, p_{\zeta_{INIT}}, 1) \; . \tag{12}$$

The credibility of the minutia type has an initial probability of $p_{\zeta_{INIT}}$ and the number of occurrences is set to 1. However, the credibility of the minutia p_{m_i} is recursively updated until all of the segmentation information $seg_0(n = 0, \ldots, n)$ has been applied to Eq. (3). The minutia credibility has an initial probability of 0.5, and is updated by these segmentation observations from the position using Eq. (3).

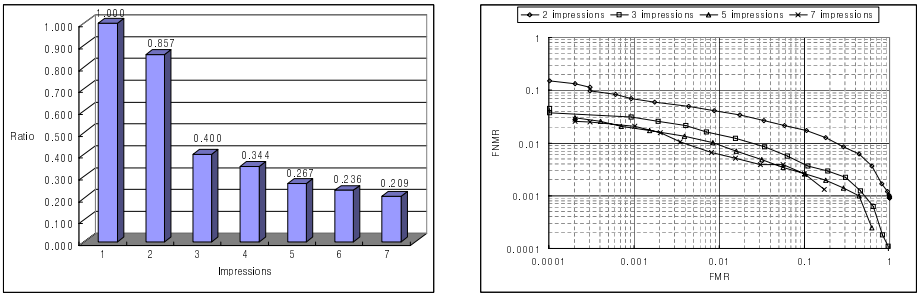
After updating all of the inputs, the *updating minutiae set* $\mathbf{u}$ contains all of the minutiae of the inputs. Among these minutiae, those that are found several times in other templates have higher credibility p_{m_i}', while those minutiae that are rarely found in other templates have lower credibility. However, those minutiae that are posed in the background area of other inputs may have relatively higher credibility.

In this paper, the super-template includes those minutiae having a higher credibility than the given threshold. The information contained in the super-template can be the same as that of the single input. Therefore, in our experiments, an ordinary one-to-one matching algorithm can use the super-template without modification.

4 Experimental Results

For the performance evaluation, this study utilizes the FVC2002 (The second Finger-print Verification Competition) [17] DB2 SetA database for our experiments. Each set consists of 100 fingers and 8 impressions per finger. With this database configuration, if k is the number of templates used for enrollment, then ${}_8C_k$ is the number of possible enrollments from a given finger and $8 - k$ is the number of genuine matchings against an enrolled template. Therefore, the total number of genuine matchings that can be performed is $100 \times (8 - k) \times ({}_8C_k)$. The first enrolled template is matched against the first image of the other fingers in impostor matching. However, in the case of a single impression, we use the FVC2002's FMR (False Match Rate) and FNMR (False Non Match Rate) testing methods [18].

Figure 3(a) shows the ratio of the EER in the case of multi-impression to the EER in the case of single impression. In these experiments, the EER is dramatically de-creased when more input impressions are applied. The DET (Detection Error Trade-off) curves in Fig. 3(b) show this in more detail. For most values of the authentication threshold, better accuracy is obtained when more impressions are provided for fusion.



(a) EER changes (b) DET curves

Fig. 3. Comparisons of recognition errors with various impressions

In addition, the proposed super-template generation algorithm offers the advantage of improved matching time and lower storage requirements. In our experiments, the average matching time is similar to that of a single impression template, regardless of the number of inputs.

5 Conclusions

In this paper, we proposed a super-template generation algorithm which used multiple inputs from the same finger. A fingerprint image selection algorithm was also proposed for the purpose of choosing the appropriate inputs. The template fusion algorithm presented in this paper utilizes the successive Bayesian estimation method. By utilizing this method, the algorithm can handle an unlimited number of inputs by means of the sequential updating scheme. The proposed algorithm not only achieves good recognition accuracy, but also keeps the matching time and storage requirements. In the preliminary experiments, the proposed algorithm shows a 60% decrease in EER as compared to the single impression case for the enrollment of 3 templates. Better error reduction can be obtained by merging a greater number of inputs.

Acknowledgement

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center(BERC) at Yonsei University.

References

1. A. Jain, L. Hong and R. Bolle, "On-Line Fingerprint Verification," *IEEE Trans. on PAMI*, vol. 19, no. 4, pp. 302-314, April 1997.
2. D. Maio and D. Maltoni, "Direct Gray-Scale Minutiae Detection In Fingerprints," *IEEE Trans. PAMI*, vol. 19, no. 1, pp. 27-39, January 1997.
3. L. Hong, Y. Wan and A. Jain, "Fingerprint Image Enhancement: Algorithm and Performance Evaluation," *IEEE Trans. PAMI*, vol. 20, no. 8, pp. 777-789, August 1998.

4. A. Farina, Z.M. Kovacs-Vajna and Alberto Leone, "Fingerprint minutiae extraction from skeletonized binary images," *Pattern Recognition*, vol.32, no. 4, pp. 877-889, 1999.
5. D. Ahn, C. Ryu and H. Kim, "Removal of False Minutiae based on Structural and Directional Attributes in Fingerprint Recognition," *Proc. 3rd International Workshop on Information Security Applications*, pp. 355 - 367, 2002.
6. D. Maio and D. Maltoni, "Neural Network Based Minutiae Filtering in Fingerprints," *Proc. 14th ICPR*, pp. 1654-1658, August 1998.
7. A. K. Jain and A. Ross, "Fingerprint Mosaicking," *Proc. Int'l Conf. on Acoustic Speech and Signal Processing*, vol 4. pp. 4064-4067, 2002.
8. D. Lee K. Choi, S. Lee and J. Kim, "Fingerprint Fusion Based on Minutiae and Ridge for Enrollment," *LNCS 2688*, pp. 478-485, 2003.
9. K. A. Toh, W. Y. Yau, X. D. Jiang, T. P. Chen, J. Lu and E. Lim, "Minutiae Data Synthesis for Fingerprint Identification Application," *Proc. IEEE Int'l Conf. Image Processing*, 2001.
10. X. Jiang and W. Ser, "Online Fingerprint Template Improvement," *IEEE Trans. PAMI*, vol. 24, no. 8, pp. 1121-1126, August 2002.
11. A. Jain, U. Uludag and A. Ross, "Biometric Template Selection: A Case Study in Fingerprints," *LNCS 2688*, pp. 335-342, 2003.
12. K. Fukunaga,*Introduction to Statistical Pattern Recognition*, 2nd Ed., pp. 389-390, 1990.
13. W. Choi, C. Ryu and H Kim, "Navigation of a Mobile Robot using Mono-Vision and Mono-Audition," *Proc. IEEE SMC*, vol. 4, pp. 686-691, 1999.
14. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd Ed., Prentice-Hall, pp. 644-646, 2002.
15. C. Ryu and H. Kim, "A Fast Fingerprint Matching Algorithm Using Parzen Density Estimation," *LNCS 2587*, pp. 525-533, 2003.
16. S. Joun, E. Yi, C. Ryu and H. Kim, "A Computation of Fingerprint Similarity Measures based on Bayesian Probability Modeling," *LNCS 2756*, pp. 512-520, August 2003.
17. D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman and A. K. Jain, "FVC2002: Second Fingerprint Verification Competition," *Proc. 16th Int'l Conf. Pattern Recognition*, vol. 3, pp. 811-814, 2002.
18. FVC2002 Web site: http://bias.csr.unibo.it/fvc2002/

Secure Fingerprint Matching with External Registration

James Reisman[1], Umut Uludag[2], and Arun Ross[3]

[1] Siemens Corporate Research, 755 College Road East, Princeton, NJ, 08540
james.reisman@siemens.com
[2] Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824
uludagum@cse.msu.edu
[3] Lane Department, West Virginia University, Morgantown, WV, 26506
ross@csee.wvu.edu

Abstract. Biometrics-based authentication systems offer enhanced security and user convenience compared to traditional token-based (e.g., ID card) and knowledge-based (e.g., password) systems. However, the increased deployment of biometric systems in several commercial and government applications has raised questions about the security of the biometric system itself. Since the biometric traits of a user cannot be replaced if compromised, it is imperative that these systems are suitably secure in order to protect the privacy of the user as well as the integrity of the overall system. In this paper, we first investigate several methods that have been proposed in the literature to increase the security of the templates residing in a biometric system. We next propose a novel fingerprint matching architecture for resource-constrained devices (e.g., smart cards) that ensures the security of the minutiae templates present in the device. Experimental results describing the impact of several system parameters on the matching performance as well as the computational and storage costs are provided. The proposed architecture is shown to enhance the security of the minutiae templates while maintaining the overall matching performance of the system.

1 Introduction

Fingerprint-based biometric systems are among the most popular personal authentication systems partly because of their long history. Further, fingerprint sensors are relatively small and cheap and, hence, can be readily incorporated into devices such as cellular phones, laptop computers, computer keyboards, PDAs, etc. This has resulted in a proliferation of these devices in the market.

A generic fingerprint system consists of four main modules: (a) the acquisition module which senses the fingerprint of a user and produces a raw image; (b) the feature extraction module which processes the raw image and extracts a compact set of features representing the fingerprint; (c) the matching module which compares the extracted feature set with the templates residing in the database by generating match scores; and (d) the decision module which determines or verifies the identity of the user based on the match scores. The performance of a fingerprint matcher hinges on the accurate derivation of transformation parameters (typically, the translation and rotation values) relating two feature sets. This process, known as registration or alignment, impacts the matching performance of minutiae-based systems[1].

[1] Some algorithms avoid alignment. See [1], for example

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 720–729, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Ratha et al. [2] discuss the various types of attacks that can be launched against a fingerprint system which can undermine the integrity of the system and, subsequently, result in the loss of privacy of the user. Therefore, it is necessary to design techniques that either prevent or neutralize the effect of these attacks. In this paper we focus on template protection schemes in resource-constrained systems such as smart cards where matching can potentially be performed in an external host computer.

Yang and Verbauwhede [3] describe a fingerprint system consisting of a secure and a non-secure part. The biometric template resides in the so-called secure part and the matching operation, which requires access to the stored biometric template, is executed in this part. In their formulation, the authors augment local minutiae information with minutiae neighborhood information. The distances (d) and angles (φ) between a specific minutia M and its nearest N (typically 6) neighbors, and the directions (ϑ) of these neighbors define the local structure for M. A collection of local structures representing all the minutiae in the image constitutes the template. During the matching process, no alignment (registration) information describing the translation and rotation between the template and input feature set is provided. Rather, the matcher compares the local structures of all minutiae in the input set against those in the template and two local structures are deemed to be a match if at least 3 of their 6 neighbors are similar in terms of the (d, φ, ϑ) triplet; the final match score is a function of the total number of matched structures. Experiments conducted on a very small database (of 10 different fingers) show that a 1% FRR (False Reject Rate) can be achieved at a 0% FAR (False Accept Rate).

Pan et al. [4] introduce a match-on-card system for fingerprints. Their system uses (x, y, θ) values for representing the minutiae set. The 3-dimensional transformation space (describing the alignment) is discretized into a set of candidate transformations defined by an accumulator array. The transformation between the input feature set and the template is obtained by comparing all the minutiae in the two sets. The authors employ a hierarchical scheme that evolves from a coarse-to-fine scale to determine the final transformation. First, a coarsely quantized accumulator array is used to compute the approximate transformation. The array cell corresponding to the best transformation is further discretized and the process repeated several times. The iterative nature of the algorithm ensures that only a limited amount of memory is required at any instance. An EER (Equal Error Rate) of ~6% is reported for the experiments carried out with 100 different fingers.

Moon et al. [5] describe another minutiae-based match-on-card system. Again, the minutiae are represented as (x, y, θ) triplets. The registration between the input feature set and the template is accomplished *outside* the smart card in order to reduce its workload. To facilitate this, the smart card stores the average horizontal and vertical coordinate values (μ_X^E, μ_Y^E) and the direction (μ_θ^E) of all minutiae in the enrolled template. When an input minutiae set is presented to the smart card for matching, these average values are calculated ($\mu_X^I, \mu_Y^I, \mu_\theta^I$) by the host computer. The parameters $\mu_X^E, \mu_Y^E, \mu_\theta^E$ are next transferred from the smart card to the host, where the minutiae of the input set is translated and rotated such that

$\mu_X^I = \mu_X^E, \mu_Y^I = \mu_Y^E, \mu_\theta^I = \mu_\theta^E$. The transformed input minutiae are sent to the smart card, where point matching yields the number of matched minutiae. The authors only report the genuine matching results on a very small database (of 10 different fingers) and, hence, it is difficult to assess the overall matching performance of the system. However, in our opinion, the use of average coordinates and angle provides a rather coarse (and possibly, incorrect) registration that will degrade system performance.

Ishida et al. [6] present a matching scheme for smart cards. The smart card stores the binary fingerprint image as well as the core point of a gray-scale fingerprint image. When an input fingerprint is presented for matching, the host computer transmits the location of its core to the smart card which determines the correct translation parameter. Next, the smart card selects some coordinates (typically around the minutiae points of the enrolled template), and after modifying them using the translation parameter, sends these coordinates to the host. The host then transmits a rectangular image patch (called *chip*, which is centered around the received coordinate) of the input image to the smart card, which proceeds to calculate the similarity between the received image chip and the corresponding chip of the enrolled image. To account for small variations in translation, multiple input chip images obtained via small shifts around the initial chip image are sent to the smart card. If, for a particular chip location, no match can be found in the input image, the smart card selects another chip location, and repeats the process. For a database with 576 fingers, the authors report that a 2% FRR is achieved at 0.1% FAR.

Yang et al. [7] discuss a method for fingerprint verification in another resource-constrained device – the Personal Digital Assistant (PDA). The authors implement the minutiae extraction, core point detection and minutiae matching operations in fixed-point arithmetic (32 bit words with [-32768, 32767] range), citing that majority of processors in mobile applications do not support floating-point arithmetic. In their scheme, minutiae (represented via (x, y, θ) triplets) are extracted using a modified version of the ridge line tracing algorithm in [8]. The transformation parameters are recovered using the location of core point in the two images. To optimize the computation time, only a subset of minutiae in the input set is used for matching; in fact, only minutiae closest to the fingerprint core are used during the matching process, which in itself employs a simple bounding box technique. The authors report that on a database of 383 different fingers, an EER of nearly 7% is achieved (the corresponding floating-point EER is reported as 6%).

2 System Architecture

The goal of automatic fingerprint matching systems is to accurately determine or validate an individual's identity. However, these systems are vulnerable to spoofing, registration template theft, and other attack methods [10]. In order to discourage identity theft, the registered fingerprint feature set needs to be held in a secure location. One possibility is to store the registered fingerprint information and accomplish matching on a *smart chip* device (e.g., smart card, USB dongle) kept in the possession of the individual. The chip releases the secure information (e.g., cryptographic keys), or a verification signal, upon being presented with a fingerprint feature representation matching the registered representation stored on the chip. The difficulty in

this application scenario is that the smart chip processor is generally incapable of handling fingerprint matching duties with satisfactory accuracy.

We propose a method where sufficient information is passed to the fingerprint scanning unit's processor (i.e., an external processor) to allow alignment of the template, but which would be insufficient to reconstruct the user's fingerprint representations. This is accomplished by using a multi-step approach in which certain selected verification tasks are moved off chip. The steps needed for fingerprint verification – fingerprint representation extraction, alignment and matching – are decoupled into separate processes.

The method is built upon the previously proposed ridge-texture based matching algorithm of Ross et al. [11]. Matching using this algorithm is fast and accurate once the templates are properly aligned, and can be handled by the smart chip's processor. More computationally intensive operations such as feature extraction are implemented in the scanning unit. Ridge map alignment is achieved by a comparison of minutiae information in the form of minutiae triplet sets. The minutiae triple sets are constructed in such a way that neither the individual's ridge map nor minutia map can be regenerated from it. This allows much of the computationally expensive ridge map alignment to be moved off-chip without loss of security.

2.1 Basic Matching Algorithm

Here, a simple description of the algorithm that forms the basis for the smartcard based fingerprint matching is provided. For more details about this algorithm, please see [11]. A set of Gabor filters, whose spatial frequencies correspond to the average inter-ridge spacing in a fingerprint, is used to capture the ridge strength at equally spaced orientations. The number of these filters is an important system parameter: e.g., if 4 filters are used, the filters are oriented at 0^o, 45^o, 90^o and 135^o; for 8 filters, the orientation difference between filters is 22.5^o, allowing a finer representation of the fingerprint ridges. A square tessellation (based on the block size used for grid generation, e.g., 8x8 blocks) of the filtered images is then used to construct a multi-dimensional *ridge feature map*. This map is expressed as an array of values, where each value can be represented with a fixed number of bits (e.g., a 4-bit representation allows 16 different levels for each feature while an 8-bit representation allows 256 levels, increasing the granularity). A similarity score of two fingerprints can be generated by finding the Euclidean, Hamming or similar vector distance between the two ridge feature maps. Further, multiple templates (e.g., 3) can be used instead of a single template during verification: using multiple templates helps in better capturing the variability of the fingerprint, hence increasing the matching accuracy.

When the ridge feature map is extracted, the 2 images need to be aligned and tessellated in the same manner. This requires the prior determination of the relative transformation between the two images. It is this step that is moved off the smart chip in the proposed algorithm.

2.2 Triplet-Based Fingerprint Registration

First, the minutiae of the fingerprint are extracted. Then, a minutia triplet (Figure 1) is used to describe a set of 3 minutia points.

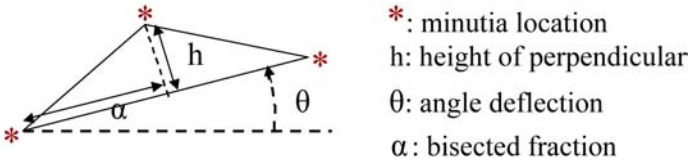


Fig. 1. A minutiae triplet triangle

Four parameters describe the shape of the minutia triplet (h, α, s, θ):

Height of Perpendicular (h): The height in pixels from the longest side to the vertex opposite it.

Bisected Fraction (α): The fraction (from 0.0 to 1.0) of the longest side from the perpendicular to its clockwise vertex (the "Primary Vertex").

Deflection Angle (θ): The angle of deflection that the longest side (running from its clockwise vertex to its counterclockwise vertex) makes with the horizontal ray running left to right in the fingerprint image.

Side Length (s): The length associated with α.

The location of the triplet in the image is defined by the (x,y) location (in pixels) of the primary vertex.

The full triplet list contains all possible combinations of the minutiae in groups of three. The extraction of the parameters depends upon the accurate determination of the longest side of the triplet triangle and the ordering of the vertices. The following qualifying criteria are applied to avoid unsuitable triplets:

- The longest side of a triplet triangle must be greater than the 2nd longest side by a certain threshold.
- The height of the triplet triangle must be greater than a certain threshold.

The selected triplets may still be more than what can be stored in a given smart chip. This number can be reduced by ordering the triplets using a scoring function and retaining the N highest scoring triplets (N is defined by the storage limitations of a given smart chip). The scoring function ranks the triplets according to various factors including:

- Repeatability of the three minutia in a given triplet in subsequent fingerprint samples (less likely if the triplet uses minutia on the periphery of the print, or far away from each other), and
- Repeatability of parameter measurement between samples (less likely if a triplet used minutia that are close together).

The scoring function is maximized when the triangle is centrally located and of "medium size", best satisfying both parameter stability and minutia repeatability. Finally, the pruned triplet list and the ridge feature map for the enrollment template are stored in the smart card.

2.3 Fingerprint Verification

Firstly, the minutiae are extracted from the query fingerprint and a complete triplet list generated in the scanning unit. The smart chip device is then requested to transmit

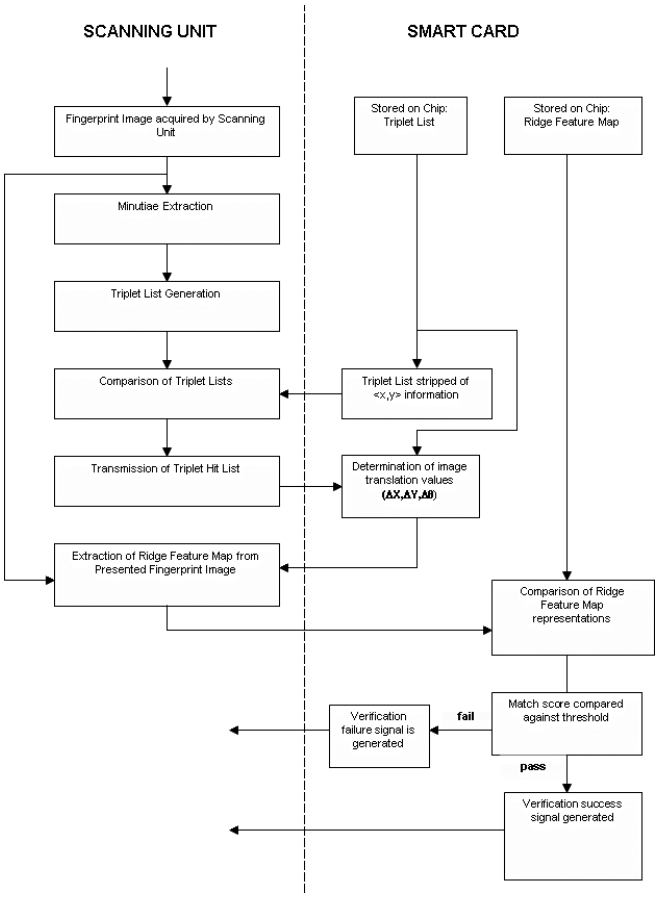


Fig. 2. Exchange of information between the smart card and the scanning unit

the parameters of the pruned triplet list, which is sent withholding the identifying (x,y) location information. Each transmitted triplet is tracked by an index number. The triplet list from the smart chip (list A) is next compared to the complete triplet list extracted by the scanning unit (list B). A triplet *hit* is said to occur when all triplet parameters from a triplet pair are within proscribed distances.

The resultant triplet hit list contains the following information: (index^a, x^b, y^b, $\Delta\theta^{ab}$) where $\Delta\theta^{ab} = \theta^a - \theta^b$. The hit list is transmitted to the smart chip device. Using the index identifier to reference $\mathbf{x}^a$ and $\mathbf{y}^a$, each triplet hit is converted into the perceived $\mathbf{\Delta x}^{ab} = \mathbf{x}^a - \mathbf{x}^b$, $\mathbf{\Delta y}^{ab} = \mathbf{y}^a - \mathbf{y}^b$, and $\mathbf{\Delta\theta}^{ab}$ transformations. Using the Parzen window voting technique, the global $\mathbf{\Delta X}$ and $\mathbf{\Delta Y}$ translations for the image is estimated from the individual $\mathbf{\Delta x}^{ab}$ and $\mathbf{\Delta y}^{ab}$ values. The hit list is then pruned to retain hits whose $\mathbf{\Delta x}^{ab}$ and $\mathbf{\Delta y}^{ab}$ fall within a threshold distance from the global $\mathbf{\Delta X}$ and $\mathbf{\Delta Y}$. Parzen window voting is next applied to the pruned list to determine the global $\mathbf{\Delta\theta}$ value. The global transformation values ($\mathbf{\Delta X}$, $\mathbf{\Delta Y}$, and $\mathbf{\Delta\theta}$) are sent from the smart chip to the scanning device, where they are used to "align" the ridge feature map of

the query fingerprint. The aligned ridge feature map is transmitted to the smart chip for matching, using the method described in Ross *et al*. [10]. If the match results in a distance score within the match tolerance threshold, a verification signal or the secret information (e.g., cryptographic key stored in smart card) is transmitted to the outside host.

3 Experimental Results

In this section, we provide results aiming to show the effect of different system parameters (block size, number of filters, feature bit quantization, and number of templates per user) on the overall performance. These results can be used to set the system parameters for a specific smart-card based fingerprint matching system, considering required accuracy and execution time, and the smart card capacity. Note that when using multiple templates, the conglomerate matching distance between a query and multiple template images is selected to be the minimum value of the associated distance set.

We used the Siemens Fingerprint Database, containing 100 images each of 36 distinct users. Images are 224x288 pixels, with 500 DPI resolution, and a 8 bit/pixel gray level quantization. ROC (Receiver Operating Characteristics) curves are generated by considering 13 images per user as template images, and 87 images per user as query images.

Effect of Block Size: Figure 3 shows the ROC curves for different block sizes, with other parameters fixed (number of filters is 4, bits per feature is 8). We see that 8x8 block size is the optimal one: smaller blocks (4x4) make the system too sensitive to block tessellation boundary artifacts. On the other hand, larger blocks (16x16) decrease the discriminability of templates, hence resulting in inferior performance.

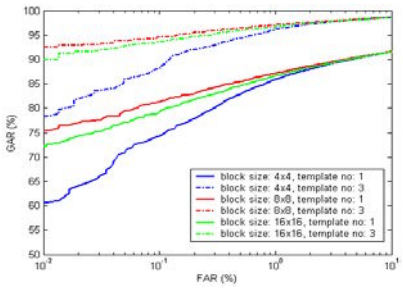


Fig. 3. Effect of block size

Also, using multiple templates per user increases the performance of the system considerably, at the cost of increased storage space and processing time (especially critical for smart-card based applications).

Effect of Number of Filters: Figure 4 shows the ROC curves for different number of filters, with the other parameters fixed (block size is 8x8, bits per feature is 8). It is observed that increasing the number of filters results in increased performance, reaching saturation at 4 filters. Hence, choosing 4 filters is a good tradeoff between performance, and computational and storage costs.

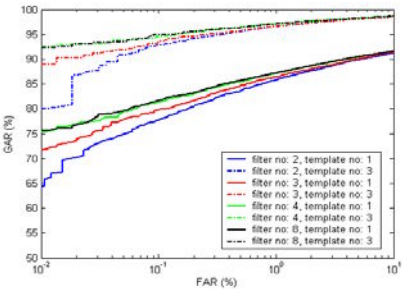


Fig. 4. Effect of number of filters used

Effect of Feature Bit Number: Figure 5 shows the ROC curves for different bit numbers per feature, with other parameters fixed (block size is 8x8, filter number is 4). We observe a performance increase when the number of bits is increased from 4 to 8, especially for small FAR (False Accept Rate) values. On the other hand, increasing the bit number to 16 does not result in any improvements.

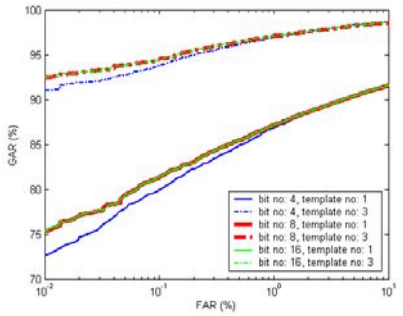


Fig. 5. Effect of bits used on matching

Effect of Multiple Template Number: Figure 6 shows the ROC curves for different template numbers per user, with other parameters fixed (block size is 8x8, filter number is 4, bits per feature is 8). As expected, increasing the number of templates results in increased performance, especially for small FAR values.

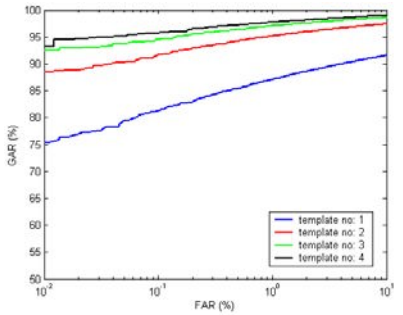


Fig. 6. Effect of number of templates

Accuracy vs. Template Size: Figure 7(a) shows the EER values against associated template sizes (in KBytes) for several different system configurations. Similarly, Figure 7(b) shows the FRR (False Reject Rate) values against associated template sizes, at a FAR equal to 0.1%. Note that in these figures, the configurations that result in values closer to the lower left corner of the graphs can be thought of as *optimal* ones (low error rate and small template size).

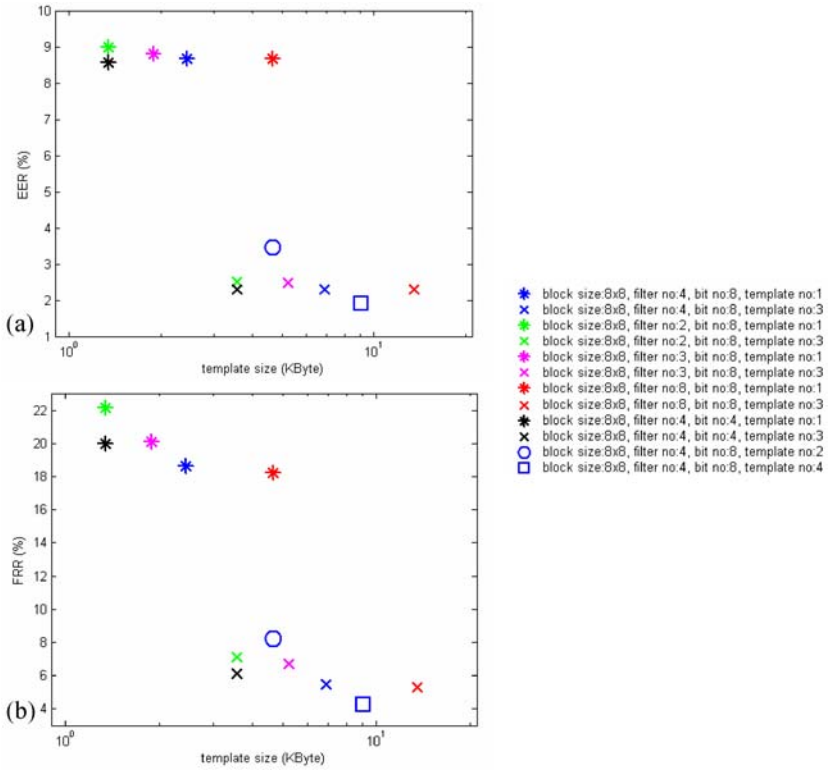


Fig. 7. (a) EER vs. template sizes for different configurations, (b) FRR vs. template sizes for different configurations (FAR = 0.1%)

In light of these experiments, we can conclude that it is better to use multiple templates per user (e.g., 3) compared to using only one template and trying to extract more information from each template (by increasing filter number, decreasing the block size, or increasing the bit numbers per feature). Note that the computational and storage costs are linearly dependent on the number of templates (e.g., for 3 templates, the required storage space is tripled, along with the computational time).

4 Conclusions

A novel fingerprint matching technique suited for execution in resource-constrained devices (e.g., smart cards) is proposed. The registration of fingerprints, which is computationally expensive, takes place in the host computer (i.e., scanning unit),

whereas the matching takes place within the resource-constrained device, without revealing the user template to the outside environment. This accomplishes both template security (since the matching takes place within the smart card, which can be protected against malicious attacks via software or hardware) and high accuracy (since the registration benefits from the availability of abundant computational resources in the host) at the same time. Experimental results showing the effects of several system parameters on the matching accuracy have been provided. These experiments help in evaluating the feasibility of candidate smart-card based fingerprint matching techniques.

References

1. M. Bazen and S. H. Gerez, “An Intrinsic Coordinate System for Fingerprint Matching”, Proc. 3rd International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 223-228, Sweden, June 6-8, 2001.
2. N. Ratha, J. H. Connell, and R. M. Bolle, "An Analysis of Minutiae Matching Strength", Proc. 3rd International Conference on Audio- and Video-Based Biometric Person Authentication, pp. 223-228, Sweden, June 6-8, 2001.
3. S. Yang and I.M. Verbauwhede, “A secure fingerprint matching technique”, *Proc. ACM SIGMM Workshop on Biometrics Methods and Applications*, pp. 89-94, 2003.
4. S.B. Pan, D. Moon, Y. Gil, D. Ahn, and Y. Chung, “An ultra low memory fingerprint matching algorithm and its implementation on a 32-bit smart card”, *IEEE Trans. Consumer Electronics*, vol. 49, no. 2, pp. 453-459, May 2003.
5. Y.S. Moon, H.C. Ho, K.L. Ng, S.F. Wan, and S.T. Wong, “Collaborative fingerprint authentication by smart card and a trusted host”, *Proc. Canadian Conf. Electrical & Computer Engineering*, pp. 108-112, 2000.
6. S. Ishida, M. Mimura, and Y. Seto, “Development of personal authentication techniques using fingerprint matching embedded in smart cards”, *IEICE Trans. Inf. & Syst*., vol. E84-D, no. 7, pp. 812-818, 2001.
7. T.Y. Yang, Y.S. Moon, and K.C. Chan, “Efficient implementation of fingerprint-verification for mobile embedded systems using fixed-point arithmetic”, *Proc. ACM Symp. Applied Computing*, pp. 821-825, 2004.
8. D. Maio and D. Maltoni, “Direct gray-scale minutiae detection in fingerprints”, *IEEE Trans. PAMI*, vol. 19, no. 1, pp. 27-40, 1997.
9. A.K. Jain, L. Hong, S. Pankanti, and R. Bolle, “An identity authentication system using fingerprints”, *Proc. IEEE*, vol. 85, no. 9, pp. 1365-1388, 1997.
10. U. Uludag and A.K. Jain, “Attacks on biometric systems: a case study in fingerprints”, *Proc. SPIE-EI 2004, Security, Steganography and Watermarking of Multimedia Contents VI*, vol. 5306, pp. 622-633, 2004.
11. A. Ross, A. Jain, and J. Reisman, “A hybrid fingerprint matcher”, *Pattern Recognition*, vol. 36, pp. 1661-1673, 2003.

Palmprint Recognition Using Fourier-Mellin Transformation Based Registration Method*

Liang Li, Xin Yang, Yuliang Hi, and Jie Tian**

Center for Biometrics and Security Research, Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Graduate School of the Chinese Academy of Science, P.O.Box 2728 Beijing 100080 China
tian@doctor.com
http://www.fingerpass.net

Abstract. In this paper, we propose a new method for palmprint recognition based on the registration results between two palmprint images. After preprocessing, a unified coordinate system is constructed for each palmprint image. If two palmprints images are captured from the same hand, the translation, rotation and scaling alignment differences between their corresponding image blocks should be small. Therefore the registration parameters obtained by the Fourier-Mellin Transformation are used to measure the similarity between two palmprint patterns. Experimental results demonstrate the effectiveness of our ideas.

1 Introduction

In information and network-based society, a person has to remember lots of passwords, pin numbers, account numbers and other security codes. Nowadays, biometrics system has taken place of this concept to some extent since it is more convenient, reliable and stable. Different techniques have been developed, each of them having its own advantages and disadvantages in terms of user acceptance, cost, performance, etc. [1]. From all these techniques, palmprint is considered as a relatively new biometric feature for personal verification and have several advantages: stability and uniqueness; medium cost as it only needs a platform and a low/medium resolution CCD camera or scanner; it is very difficult to be mimicked; it is very easy and acceptable to users; it is lack of relation to police or justice. It is for these reasons that palmprint recognition has attracted more interests from researchers.

There are many features in a palmprint image that can be extracted for authentication. Principal lines, wrinkles, ridges, minutiae points, singular points, and texture are regarded as useful features for palmprint representation[2]. For palmprint, though, there is no universal method of feature extraction and recognition. In existing research, the majority focused on: points and lines[3][4][5][6]; texture analysis[7][8][9]; statistic features[10][11][12] and hybrid of different types of features[13].

* This paper is supported by the Project of National Science Fund for Distinguished Young Scholars of China under Grant No. 60225008, the Key Project of National Natural Science Foundation of China under Grant No. 60332010, the Project for Young Scientists' Fund of National Natural Science Foundation of China under Grant No.60303022, and the Project of Natural Science Foundation of Beijing under Grant No.4052026

** Corresponding author: Jie Tian; Telephone: 8610-62532105; Fax: 8610-62527995

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 730–738, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper we investigate a novel palmprint recognition method which uses Fourier-Mellin Transformation (FMT) based on registration method. After extracting the palmprint region of interest (ROI), we divide the ROI into 16 blocks and then convert each block palmprint to frequency domain using FMT. Each pair of two corresponding blocks is registered in terms of their FMT features. Information obtained in registration stage is used to measure the similarity between two palmprint images.

The rest of this paper is organized as follows. In the Section 2 is the preprocessing stage. Section 3 presents the background theory of FMT and Section 4 is devoted to the methods of registration-based feature extraction and matching algorithm. The experimental results are listed in Section 5. At last, we discuss our algorithm and future work in Section 6.

2 Preprocessing

Our work is carried on the PolyU Palmprint Database[14]. The image of this database contains the whole palmprint and other parts of a palm and background. Therefore a preprocessing step is needed to extract the ROI. In these images, we find that the background border between middle finger and ring finger is stable because the border is the shadow of pegs equipped at image acquisition device, these pegs were used to separate and locate the fingers. According to this prior, we can locate this border line robustly using line searching algorithm(see Fig.1). 5 steps are included in preprocessing stage:

1. Apply a lowpass filter to smooth left half part of the original image and convert this part to a binary image (see Fig.1b).
2. Use analogically vertical line searching algorithm to find the line segment P_1P_2 along palm's boundary(see Fig.1c).
3. Make a palmprint coordinate system which y-axis is line P_1P_2 and the origin is the midpoint of this line segment(see Fig.1d).
4. Extract a subimage of a fixed size based on the coordinate system. The size of cropped subimage, in our work, is 128×128(see Fig.1e).
5. Normalize the ROI image so that it has a prespecified mean and variance(see Fig.1f).

Two ROI images which come from the same person obtained by our method will have relative offsets inasmuch different placement of palms in acquisition. Small amount of displacement, however, will have small effect on recognition result, because registration method used in our method will counteract the displacement to some extent, see Section 4.2. Compared with the method proposed in [8] which uses bisectors to locate ROI, our method just processes half of the image and needs not to compute the bisectors. It is more simple and fast.

3 Theory of Fourier-Mellin Transformation

The Fourier-Mellin transformation is a useful mathematical tool for the recognition of images because its resulting spectrum is invariant in rotation, translation and scale[15]. The Fourier Transformation itself is invariant in translation in Cartesian

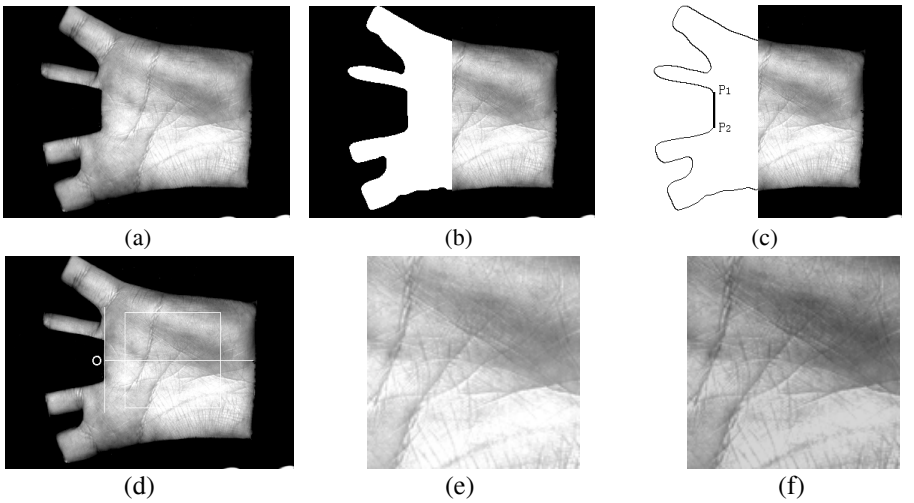


Fig. 1. The main steps of preprocessing. (a)Original database image, (b)Binarizing half of image, (c)Tracking boundary and searching line segment, (d)Building coordinate system, (e)Extracting ROI, (f)Normalizing ROI

coordinate system and in rotation by converting the Cartesian coordinate system to Polar coordinate system; the Mellin Transformation provides the invariant results for scale.

If an image $f_2(x,y)$ is a translated, rotated and scaled replica of $f_1(x,y)$ with translation (x_0, y_0) ,rotation θ_0 and uniform scale factor σ, then

$$f_2(x,y) = f_1(\sigma(x\cos\theta_0 + y\sin\theta_0) - x_0, \sigma(-x\sin\theta_0 + y\cos\theta_0) - y_0) \tag{1}$$

According to the Fourier Transform property, transforms of f_1 and f_2 are related by

$$F_2(u,v) = \exp(-j2\pi(ux_0 + vy_0))\sigma^{-2}(F_1(\sigma^{-1}(u\cos\theta_0 + v\sin\theta_0), \sigma^{-1}(-u\sin\theta_0 + v\cos\theta_0)) \tag{2}$$

With magnitudes F_1 and F_2, rotation can be deduced by representing the rotation with polar coordinates. i.e., in polar representation

$$F_{2p}(\theta, r) = \sigma^{-2} F_{1p}(\theta - \theta_0, r/\sigma) \tag{3}$$

Scaling can be further deduced to a translation by using logarithmic for the radial axis, thus

$$F_{2pl}(\theta, \sigma) = \sigma^{-2} F_{1pl}(\theta - \theta_0, \log(r) - \log(\sigma)) \tag{4}$$

Let M_1 and M_2 be the magnitudes of F_{1pl} and F_{2pl}, their Fourier magnitudes spectra in polar representation are related by (ignoring σ^{-2})

$$M_1(\theta, r) = M_2(\theta - \theta_0, \log r - \log\sigma) \tag{5}$$

Using the equation (5) and the phase correlation technique, scale σ and rotation angle θ_0 can be found out. Once the scale and angle information are obtained, the image is scaled and rotated according to the amounts of σ and θ_0, respectively, and the amount of translational movement is found out using phase correlation technique as well.

4 Registration-Based Feature Extraction and Matching

4.1 FMT Feature Extraction

In order to get the correspondence of one palmprint with the other for the purpose of verification, the ROI subimages is divided equally into 4×4 blocks. By dividing the ROI, more detailed registration information will be obtained. We transform each block to frequency domain using FMT as described above. Before being mapped to log-polar plane, the FMT spectra needs to be multiplied with a highpass filter to reduce the effect of discretization and logarithm resampling. The Fig.2(b) shows the FMT

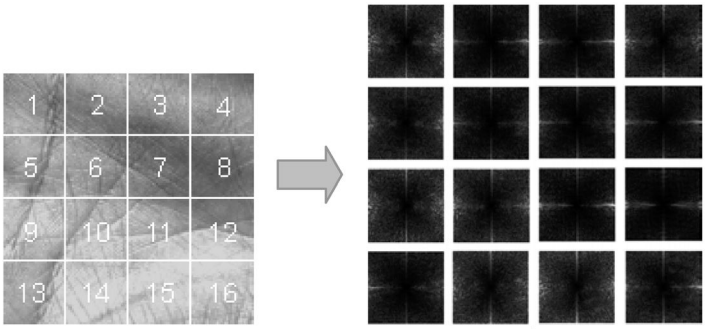


Fig. 2. Fourier-Mellin Transformation of blocks. a)Divided ROI, b)FMT feature images of blocks

4.2 FMT Feature Registration

As mentioned above, each block of one image is considered as a rotated, scaled and translated replica of corresponding block of the other image. Then we can use FMT-based registration technique described in Section 3 to obtain the translation, rotation and scaling information of each pair of blocks. The vector that describes the similarity of two palmprints is given by

$$V = (Tx_1, Ty_1, \theta_1, \sigma_1, Tx_2, Ty_2, \theta_2, \sigma_2, \ldots\ldots, Tx_{16}, Ty_{16}, \theta_{16}, \sigma_{16})$$

where Tx_i, Ty_i, θ_i ($0° \leqslant \theta_i \leqslant 180°$) and σ_i are the vertical translation movement and horizontal translation movement and rotation angle and scaling factor of the i^{th} block (i = 1,2,…,16), respectively.

For simplicity, we define that $d_i = \sqrt{Tx_i^2 + Ty_i^2}$, i.e. d_i is the translation distance. Now the above vector is translated by

$$V' = (d_1, \theta_1, \sigma_1, d_2, \theta_2, \sigma_2, ..., d_{16}, \theta_{16}, \sigma_{16})$$

We call vector V' registration feature vector (RFV). Ideally the values of d_i and θ_i are near to zero, and the values of σ_i are near to 1. In realistic situation where displacements exist in acquisition, the probability that all the values of d_i and θ_i are zeros is very small when genuine match, because even two images captured in the same session will have a amount of offsets in translation, rotation and scale. However the amounts of registration parameters when imposter match are much larger than the one when genuine match. An example of genuine registration and imposter registration of one block can be seen in Fig.3.

4.3 Classification Criteria

The main purpose of this paper is to investigate the discriminability of registration information, therefore, we just define a simple linear similarity function to test classification performance. This function is given by

$$\begin{aligned} &C(d_1, d_2, ..., d_{16}, \theta_1, \theta_2, ..., \theta_{16}, \sigma_1, \sigma_2, ..., \sigma_{16}) \\ &\quad = \alpha f_1(d_1, d_2, ..., d_{16}) + \beta f_2(\theta_1, \theta_2, ..., \theta_{16}) + \gamma f_3(\sigma_1, \sigma_2, ..., \sigma_{16}) \end{aligned} \tag{6}$$

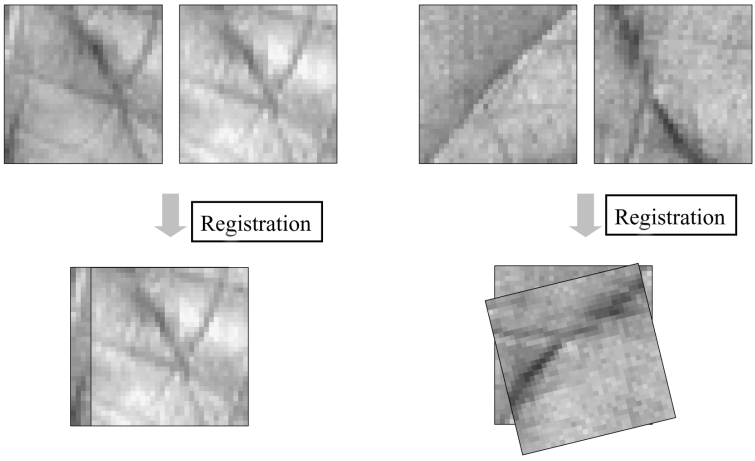
where α , β and γ are the weight factor and $\alpha + \beta + \gamma = 1$. α , β and γ are the experiential values and here we used $\alpha = \beta$ =0.4 and γ =0.2. f_1 , f_2 and f_3 are the exponential sum function which is given by

$$f_1 = \exp(-\sqrt{\sum_{i=1}^{16} (d_i - d_{\max} / u_d)^2}), \quad f_2 = \exp(-\sqrt{\sum_{i=1}^{16} (\theta_i - \theta_{\max} / u_\theta)^2})$$

and

$$f_3 = \exp(-\sqrt{\sum_{i=1}^{16} (\sigma_i - \sigma_{\max} / u_\sigma)^2})$$

where $d_{\max}$, $\theta_{\max}$, $\sigma_{\max}$, u_d , u_θ and u_σ will be found out in training stage and the definitions of $d_{\max}$ and u_d are shown in Fig.4. P(d) is the genuine registration distance distribution and $d_{\max}$ is the distance which have maximum probability, i.e. accounts for the greatest percentage. u_d is the confidence interval and we define $\int_{d_{\max}-u_d}^{d_{\max}+u_d} P(d)dd > 0.9$. Also, $\theta_{\max}$, $\sigma_{\max}$, u_θ and u_σ have the same meanings as $d_{\max}$ and u_d .



(a) Genuine registration (b) Imposter registration

Fig. 3. Example of block registration. a) Genuine registration with $d = 5, \theta = 0$ and $\sigma = 1$, b)Imposter registration with $d = 3.605, \theta = 63.71°$and $\sigma = 1$

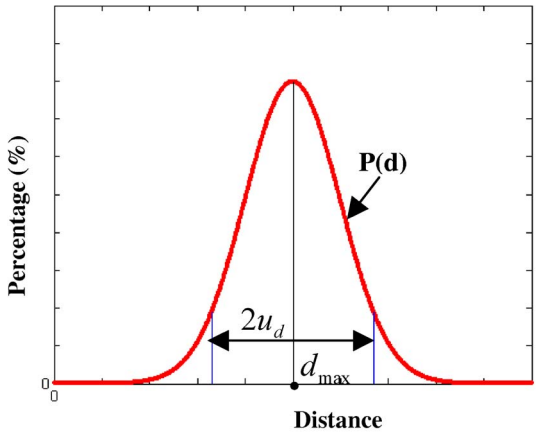


Fig. 4. Definitions of $d_{\max}$ and u_d

Using the equation (6), a similarity score which is between 0 and 1 can be calculated. Two palmprints will be verified as from the same palm if the score is higher than a threshold.

5 Experimental Results

5.1 Palmprint Image Database

The PolyU Palmprint Database is so far the first and the largest open palmprint database, which contains 600 grayscale images corresponding to 100 different palms in

BMP image format. The details of the palmprint image acquisition can be seen in [8]. Six samples from each of these palms were collected from the same person in two sessions, where 3 samples were captured in the first session and the other 3 in the second session. The average interval between the first and the second collection was two months.

5.2 Training Stage

In our experiment, we divide the PolyU Palmprint Database into 2 subsets. The first subset includes total 300 images of the first 50 palms and is used to train the values of $d_{\max}$, $\theta_{\max}$, $\sigma_{\max}$, u_d, u_θ and u_σ. The second one includes the rest of images and is used to test the recognition performance. In training stage, we compute the values of d, θ and σ when genuine and imposter match. According to the distribution of training samples when genuine match, we can obtain the value of $d_{\max}$, $\theta_{\max}$, $\sigma_{\max}$, u_d, u_θ and u_σ, which will be used in test stage. The distribution of d, θ and σ can be seen in Fig.5. As it is clearly seen, d and θ have a strong discriminability in distribution while the discriminability of σ is weaker.

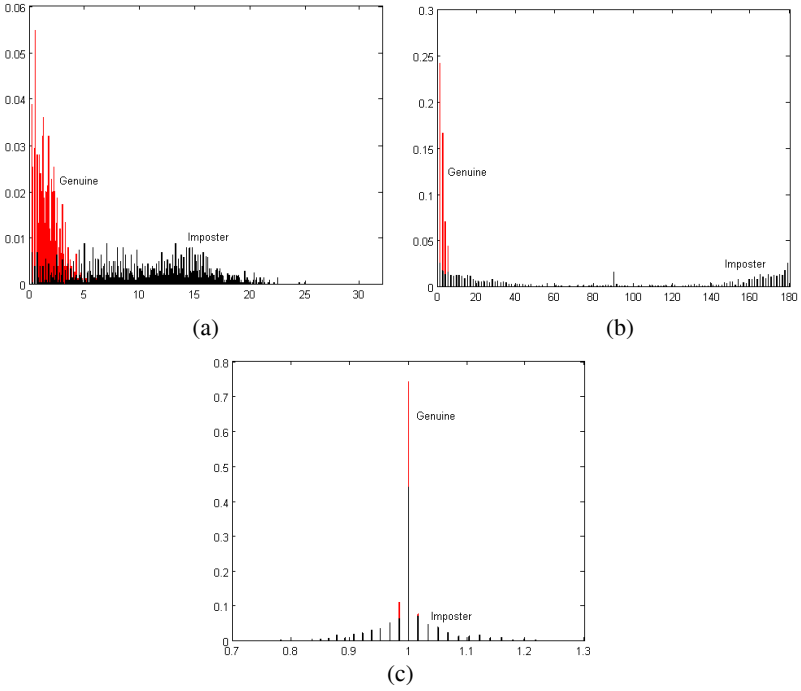


Fig. 5. Distribution of training samples. a) Distribution of d, b) Distribution of θ, c) Distribution of σ

5.3 Test Stage

In this stage, each sample in test subset is matched against the remaining samples of the same palm to compute the False Rejection Rate (FRR). The total number of genuine tests (in case no enrollment rejections occur) is: ((6*5)/2)*50 = 750; The first sample of each palm in the test subset is matched against the first sample of the remaining palms in this subset to compute the False Acceptance Rate(FAR). The total number of imposter tests is: (50*49)/2 = 1225. The distribution of genuine match scores and imposter match scores and the FAR vs. FRR curve can be seen in Fig.6. From the Fig.6(b), we can see that our approach can achieve equal error rate (EER) of 4.9%.

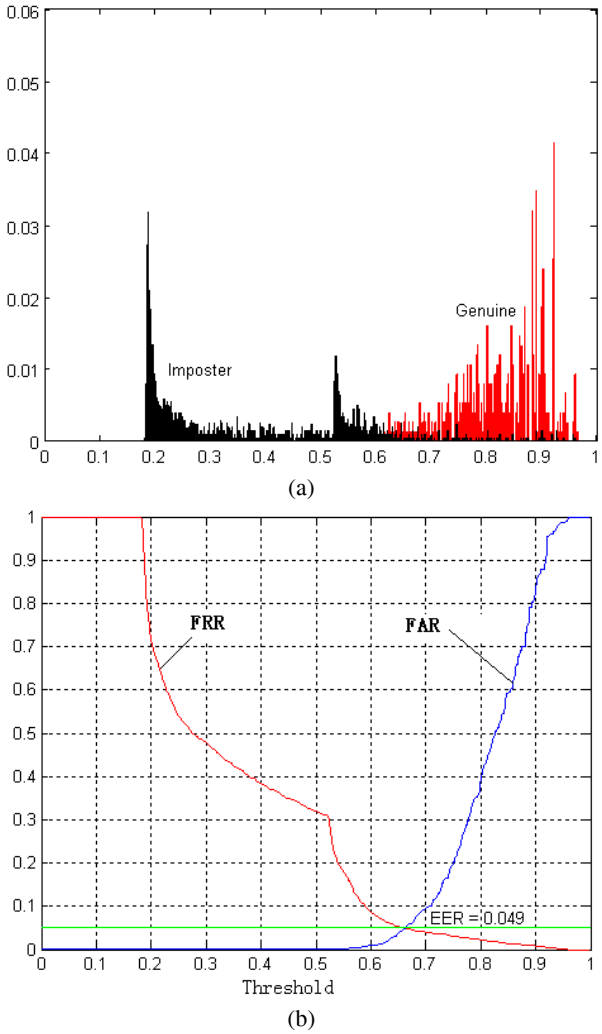


Fig. 6. Test results. (a)Genuine and imposter match score distributions,(b)FAR vs. FRR

6 Conclusions and Future Work

A novel method to feature extraction and matching is proposed in this paper. FMT features are extracted in the divided blocks of palmprint images and then are used to get corresponding block's registering information such as translation, rotation and scaling by registration techniques. Registration-based method is robust to some extent in rotation and translation of palmprint images. The experimental results show that the registration information has a strong discriminability and still have potential to be improved. In the future work, we will investigate the registration method based on the fusion with other features and the design of compact classifier will be investigated as well.

Reference

1. Jain, A.K., Bolle, B., Pankanti, S.: BIOMETRICS Personal Identification in Network Soc., p.411, Kluwer Academic, 1999.
2. Shu, W., Zhang, D.: Automated Personal Identification by Palmprint, Optical Eng.,Vol.37.1998(2659-2362).
3. Zhang, D., Shu, W.: Two novel characteristics in palmprint verification: Datum point invariance and line feature matching. Pattern Recognition,Vol.32,1999(691-702).
4. Duta, N., Jain, A.K., Mardia, K.V.: Matching of Palmprints, Pattern Recognition Letters, Vol.23.2001(477-485).
5. Wu, X., Wang, K., Zhang, D.: An approach to line feature representation and matching for palmprint recognition, Journal of Software, Vol.15.2004(869-880).
6. Wu, X., Wang, K., Zhang, D.: HMMs Based Palmprint Identification, 1st International Conference on Biometric Authentication, Springer Lecture Notes in Computer Science, Vol. 3072. Springer-Verlag, Berlin Heidelberg New York (2004)775–781.
7. Kong, W., Zhang, D.: Palmprint feature extraction using 2-D Gabor filters, Pattern Recogntion, Vol.36.2003(2339-2347).
8. Zhang, D., Kong, W., You, J., Wong, M.: Online Palmprint identification, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.25.2003(1041-1050).
9. Li, W., Zhang, D., Xu, Z.: Palmprint identification by Fourier Transform. Int.J.Pattern Recognit. Artificial Intell., Vol.16.2002(417-432).
10. Lu, G., Zhang, D., Wang, W.: Palmprint recognition using eigenpalms features,Pattern Recognition Letters, Vol.24.2003(1463-1467).
11. Wu, X., Zhang. D., Wang, K.: Fisherpalms based palmprint recognition, Pattern Recognition Letters, Vol.24.2003(2829-2938).
12. Connie, T., Teoh, A., Goh, M., Ngo, D.: Palmprint Recognition with PCA and ICA, Conference of Image and Vision Computing New Zealand 2003 (IVCNZ'03), pp 227-232, November 26th 2003, Massey University, Palmerstone North, New Zealand.
13. You, J., Li, W., Zhang, D.: Hierarchical palmprint identification via multiple feature extraction, Pattern Recognition, Vol.35.2002(847-850).
14. PolyU Palmprint Palmprint Database, http://www.comp.polyu.edu.hk/~biometrics/
15. Reddy, B.S., Chatterji, B.N.: An FFT-based Technique for Translation, Rotation and Scale-Invariant Image Registration, IEEE Transactions on Image Processing, Vol5.1996(1266-1271).

Parametric Versus Non-parametric Models of Driving Behavior Signals for Driver Identification

Toshihiro Wakita[1,2], Koji Ozawa[2], Chiyomi Miyajima[2], and Kazuya Takeda[2]

[1] Toyota Central R&D Labs., Yokomichi, Nagakute, Aichi, 480-1192, Japan
wakita@mosk.tytlabs.co.jp
http://www.tytlabs.co.jp/eindex.html

[2] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
k-ozawa@sp.m.is.nagoya-u.ac.jp, {miyajima,takeda}@is.nagoya-u.ac.jp
http://www.is.nagoya-u.ac.jp/index.html.en

Abstract. In this paper, we propose a driver identification method that is based on the driving behavior signals that are observed while the driver is following another vehicle. Driving behavior signals, such as the use of of the accelerator pedal, brake pedal, vehicle velocity, and distance from the vehicle in front, are measured using a driving simulator. We compared the identification rate obtained using different identification models and different features. As a result, we found the non-parametric models to be better than the parametric models. Also, the driver's operation signals were found to be better than road environment signals and car behavior signals.

1 Introduction

With increased emphasis on the practicality and safety of vehicles, the recognition of drivers and their driving behavior has gained importance. The ability to recognize a driver and his/her driving behavior could form the basis of many applications, such as, driver authentication for security purpose, the ability to detect the driver becoming drowsy, and the customization of the vehicle's functions to suit that driver's personal preferences. A key technology is "human behavior signal processing" which involves the processing and recognition of human behavior signals such as the operation of the accelerator pedal. In this paper, we present a driver identification method that is based on such behavior signals.

"Driving behavior" is a cyclic process, as described below (**Fig.1**).

1. The driver recognizes the road environment, consisting of, for example, the road layout and the distance to the vehicle in front.
2. The driver decides the action that he/she should take, such as, accelerating, braking, and/or steering.
3. The driver operates the accelerator pedal, brake pedal, and/or steering wheel.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 739–747, 2005.
© Springer-Verlag Berlin Heidelberg 2005

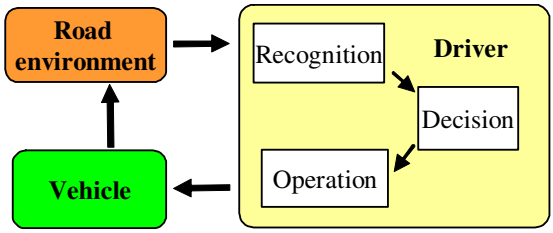


Fig. 1. Basic dynamics of driving behavior, vehicle status, and road environment

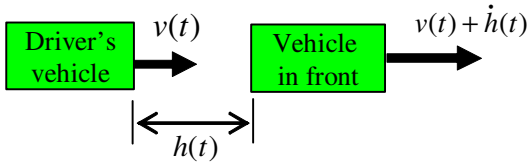


Fig. 2. Car following

4. The vehicle status (ex. velocity, yaw rate) changes according to the driver's operation.
5. The road environment (ex. distance to the vehicle in front) changes according to the vehicle status.

The most elementary and familiar driving behavior is "car following", which involves maintaining a constant distance from the vehicle in front, and adjusting the relative velocity accordingly (**Fig.2**).

In this figure, $v(t)$ is the velocity of the driver's vehicle, and $h(t)$ is the distance to the vehicle in front. The velocity of the vehicle in front is $v(t) + \dot{h}(t)$. $\dot{h}(t)$ is the temporal differential of $h(t)$.

In this research, we aimed to identify a driver by using the driving behavior signals that are observed while the driver is performing the "car following" task.

2 Driving Simulator

We used a driving simulator to collect the driving behavior signals. The driving simulator acquires signals corresponding to the operation of the accelerator pedal, brake pedal, and steering wheel, calculate the corresponding vehicle behavior, and display a representation of the road environment on an LCD monitor (**Fig.3**). The road is a two-lane highway with a layout recognizable as an actual Japanese highway. The vehicle in front acts as if it were negotiating mild traffic congestion.

3 Model Comparison

Two types of model can be used to process the behavior signals. The first is a physical dynamic model. This model explicitly assumes some physical dynamics

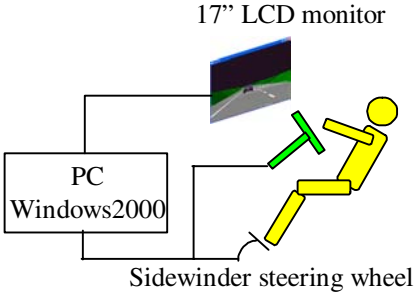


Fig. 3. Driving Simulator

between the observation variables. As the driver's personality directly affects the model parameters, we can identify the driver from those model parameters.

The second is a statistical, non-parametric model. This model assumes the existence of a mathematical relationship between the observation variables, and that the driver can be identified by using a statistical pattern recognition technique.

In the following paragraphs, we compare the two methods.

3.1 Stimulus-Response Model

Model. The most familiar model for car following is the "stimulus-response model" [1][2][3]. A difference in the velocity of the vehicle in front, as well as a change in the distance to that vehicle, act act as stimuli to the driver, who responds by either accelerating or decelerating.

$$\dot{v}(t+T) = C_1\dot{h}(t) + C_2\{h(t) - D\} \tag{1}$$

C_1, C_2 is the response sensitivity to the stimulus, D is the optimum distance to the vehicle in front, and T is the response delay. These values may be the constants or the functions of other variables. While many models have been proposed to represent C_1, C_2, D, T, we chose to use the Helly model [4].

$$\dot{v}(t+T) = \beta_1\dot{h}(t) + \beta_2 h(t) + \beta_3 v(t) + \beta_4 \tag{2}$$

$T, \beta_1, \beta_2, \beta_3, \beta_4$ are constant parameters As this is a linear model, the parameter estimation is stable and the physical meanings of these parameters can be interpreted easily.

Experiment. We performed the experiment described below.

Test subjects. Eight males, all in their twenties, all holding driver's licenses

Task. Three minutes of car following

Sessions. Four attempts at each of two different road layouts (total of eight sessions for each subject)

Measured signals. Velocity of driver's vehicle, velocity of vehicle in front, distance to vehicle in front

Identification Method and Results. For the T parameter, we used a value of 500ms, which we derived from other simple stimulus-response experiments.

The identification process is as follows.

1. Parameter vector $\boldsymbol{x} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ is calculated for the data obtained from each session, using the least-square-error method.
2. For each driver c, the data obtained from the eight sessions was divided into six blocks of learning data and two blocks of estimation data.
3. For each driver c, the average parameter vector $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$ are calculated using the parameter vectors x of the six blocks of learning data.
4. For each block of estimation data, we calculated the Mahalanobis distance D_c between the estimation data and the average for each driver. The estimation data is identified as the driver having the least Mahalanobis distance.

$$D_c = (\boldsymbol{x} - \boldsymbol{\mu}_c)' \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \tag{3}$$

A cross-validation test with the process above gave an identification rate of 43.8%.

3.2 Optimal Velocity Model

Model. Another model for the car following task is the "optimal velocity model" [5]. This model assumes that a driver has his/her own optimal velocity for a given distance to the vehicle in front, and accelerates/decelerates according to the difference between the current velocity and the optimal velocity.

$$\dot{v}(t+T) = \alpha\{V_{\mathrm{opt}}(h(t)) - v(t)\} \tag{4}$$

$$V_{\mathrm{opt}}(h) = V_{\max}[1 - \exp\{-a(h - h_0)\}] \tag{5}$$

$V_{\mathrm{opt}}(h)$ is the optimal velocity function , α is the sensitivity parameter, $V_{\max}$ is the maximum velocity, and a, h_0 is the parameter that represents the driver's optimal velocity property.

Identification Method and Results. For the parameter T, V_{max}, we used 500ms and 32m/s which we derived from another simple experiment.

The identification method is same as that described in Section 3.1. The parameter for identification is a, h_0, α.

The identification rate was found to be 54.7% with a cross-validation test.

3.3 Gaussian Mixture Model

Model. The Gaussian Mixture Model (GMM) is well known and used in many applications[6]. GMM is a statistical model that is a linear combination of Gaussian basis functions. The output probability of a GMM λ to the observation vector $\boldsymbol{o}$ is as follows:

$$b(\boldsymbol{o} \mid \lambda) = \sum_{m=1}^{M} \omega_m \mathcal{N}_m(\boldsymbol{o}) \tag{6}$$

$$\lambda = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \mid m = 1, 2, \ldots, M\} \tag{7}$$

where $\boldsymbol{o}$ is an observation vector, λ is a Gaussian mixture model, $b()$ is an output probability, M is the number of mixture functions, $\boldsymbol{\mu}_m$ is the centroid vector of the mth mixture function, and $\boldsymbol{\Sigma}_m$ is the covariance matrix of the mth mixture function.

ω_m is the mixture weight for the mth mixture function and satisfies the following equation:

$$\sum_{m=1}^{M} \omega_m = 1 \tag{8}$$

$\mathcal{N}_m(\boldsymbol{o})$ is the mth mixture function and is defined by the equation below:

$$\mathcal{N}_m(\boldsymbol{o}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_m|}} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu}_m)' \ \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{o} - \boldsymbol{\mu}_m)\right\} \tag{9}$$

where $\boldsymbol{\Sigma}_m, \boldsymbol{\Sigma}_m^{-1}$ is the covariance matrix and the inverse of the covariance matrix,and $(\boldsymbol{o} - \boldsymbol{\mu}_m)'$ is the transpose of $(\boldsymbol{o} - \boldsymbol{\mu}_m)$. In this work, we use a diagonal matrix for $\boldsymbol{\Sigma}_m$.

The likelihood of the model λ to the observation vector $O = (o_1, o_2, ...)$ is defined by the next equation:

$$P(\boldsymbol{O} \mid \lambda) = \prod_{t=1}^{T} b(\boldsymbol{o}_t) = \prod_{t=1}^{T} \sum_{m=1}^{M} \omega_m \, \mathcal{N}_m(\boldsymbol{o}_t) \tag{10}$$

Identification Method. The experimental data was the same as that described in Section 3.1. The identification process is as follows:

1. For each driver c, the eight items of session data are divided into six blocks of learning data and two blocks of estimation data.
2. For each driver c, we estimated the Gaussian mixture model λ_c. The mixture weight ω_m, centroid vector $\boldsymbol{\mu}_m$, and covariance matrix $\boldsymbol{\Sigma}_m$ are calculated using feature vectors o of six blocks of learning data with the EM algorithm. The elements of the feature vector are some of $v, \Delta v, h, \Delta h$, where Δx represents the temporal change in value x and is calculated using the the following equation:
$$\Delta x(t) = \frac{\sum_{k=-K}^{K} kx(t+k)}{\sum_{k=-K}^{K} k^2} \tag{11}$$
where $x(t)$ is the original feature, K is the time window duration (in this work, $2K = 600ms$). The mixture number is one of 2,4,8, and 16.
3. For each block of estimation data, we calculated the likelihood $P(O|\lambda_c)$ for each driver c. The estimation data is identified for the driver for whom the likelihood is the greatest.

A cross-validation test was done using the above process.

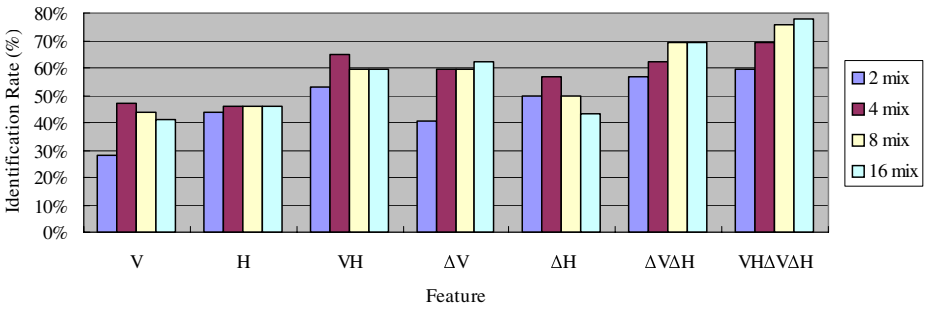


Fig. 4. Result of identification with Gaussian mixture model

Results. The identification results are shown in **Fig.4**. V is the velocity, H is the distance to the vehicle in front, and Δ represents the temporal change. We can see that the dynamic features are effective for the likes of speech recognition [7]. The best identification rate was 78%, which was obtained using $V, \Delta V, H, \Delta H$.

The stimulus-response model described in Section 3.1 uses the variable $v, \dot{v}, h, \dot{h}$ and the identification rate was 43.8%. The identification rate of GMM using a similar feature $V, \Delta V, H, \Delta H$ was 78%. The optimal velocity model described in Section 3.2 uses the variable $v, \dot{v}, h$ and the identification rate is 54.7%. The identification rate of GMM using less feature V, H is 69%. In each case, the non-parametric GMM model was found to be better than the parametric physical model. This result suggests that:

- GMM can represent the underlying dynamics between features with the joint distribution function.
- GMM can represent the non linearity and stochastic aspects with a probabilistic distribution function.

4 Feature Comparison

In the previous section, we showed that the GMM model exhibits good identification performance. In this section, then, we compare the features of GMM.

4.1 Experiment and Identification Method

To check the properties of the features, we performed another experiment.

Test subjects Twelve males, all in twenties, all holding driver's license
Task Three minutes of car following
Sessions Four attempts at each of two different road layouts (total of eight sessions for each subject)
Measured signals Driver's vehicle velocity V, distance to the vehicle in front H, accelerator pedal angle A, brake pedal angle B

The identification method was the same as that described in Section 3.3.

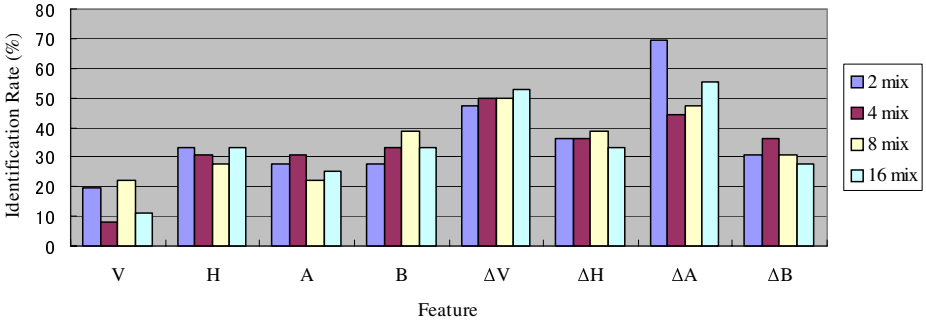


Fig. 5. Identification result for single feature

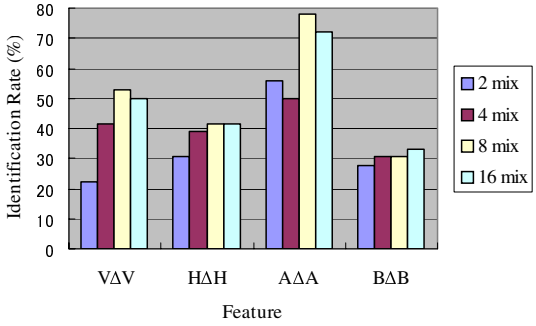


Fig. 6. Identification result for single feature

4.2 Single Feature

The identification results are shown in **Fig. 5** and **Fig. 6**. V is the driver's vehicle velocity, H is the distance to the vehicle in front, A is the accelerator pedal angle, B is the brake pedal angle, and Δ represents the temporal change. This result shows that the accelerator pedal behavior signal offers the best means of identification. We believe that the reason for this is as follows:

- As the accelerator pedal is operated directly by the driver, it is best at preserving the personal property information.
- The accelerator pedal is operated more frequently than the brake pedal.
- As the vehicle velocity and the distance to the vehicle in front are both results of the convolution of the driver's operation, the physical properties of the vehicle, and the properties of the vehicle in front (**Fig. 7**), the personal property information can be unclear.

4.3 Multiple Features

Fig. 8 shows the results for multiple features. This result shows that the feature of the accelerator pedal and the distance to the vehicle in front offer the best

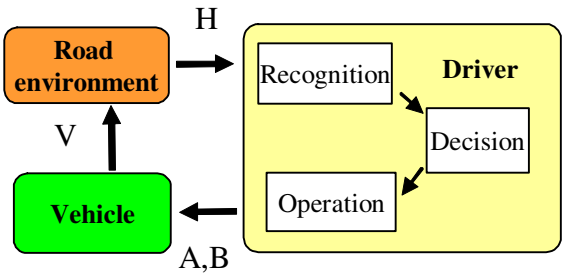


Fig. 7. Basic dynamics and feature variables

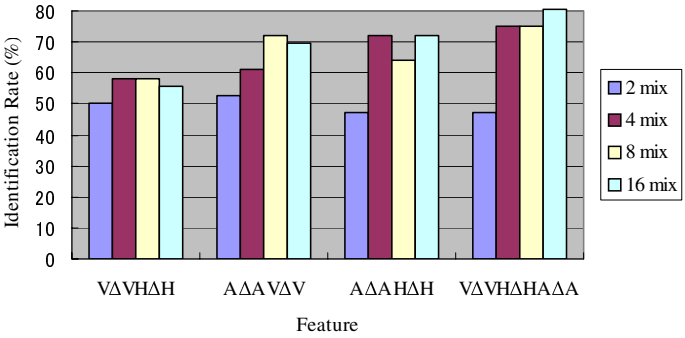


Fig. 8. Identification result for multiple features

combination. This is reasonable, because that these features provide the input and the output for the driver (**Fig. 7**).

5 Conclusion and Future Work

We have proposed a driver identification method based on the driving behavior signals that are observed while car following. The driving behavior signals of the accelerator pedal, brake pedal, vehicle velocity, and distance to the vehicle in front were measured using a driving simulator. We compared the identification rate using different identification models and different features. We obtained two results.

- Non-parametric models are superior to parametric models.
- A driver's operation signals are better than the road environment signals and vehicle behavior signals.

The physical model and statistical model is not competitive model. As the next step of this research, we aim to analyze the underlying properties of the behavior signals, merge these two models, and develop a more precise identification method.

References

1. Oguchi,T. : Analysis of Bottleneck Phenomena at Basic Freeway Segments - Car-following Model and Future Exploration - (in Japanese), Journal of the Japan Society of Civil Engineers, **660, IV-49** (2000) 39–51
2. Brackstone,M., McDonald,M.: Car-following: a historical review, Transportation Research **Part F 2** (1999) 181–196
3. Ranjitkar,P., Nakatsuji,T., Asano,M.: Performance Evaluation of Microscopic Traffic Flow Models Using Test Track Data, 2004 TRB Annual Meeting (2004)
4. Helly,W.: Simulation of Bottlenecks in Single Lane Traffic Flow, Proc. of the Symposium on Theory of Traffic Flow, Research Laboratories, General Motors (1959) 207–238
5. Bando,M., Hasebe,K., Nakayama,A., Shibata,A., Sugimaya,Y.: Dynamical Model of Traffic Congestion and Numerical Simulation Physical Review **E51** (1995) 1035–1042
6. Reynolds,D.A., Rose,R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Trans. Speech and Audio Processing, **3** (1995) 72–83
7. Furui,S. : Speaker-independent isolated word recognition using dynamic features of speech spectrum, IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-34, no.1** (1986) 52–59
8. Igarashi,K., Miyajima,C., Itou,K., Takeda,K., Itakura,F., Abut,H.: Biometric identification using driving behavioral signals, Proc. 2004 IEEE International Conference on Multimedia and Expo (2004)

Performance Evaluation and Prediction for 3D Ear Recognition

Hui Chen, Bir Bhanu, and Rong Wang

Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{hchen,bhanu,rwang}@vislab.ucr.edu

Abstract. Existing ear recognition approaches do not give theoretical or experimental performance prediction. Therefore, the discriminating power of ear biometric for human identification cannot be evaluated. This paper addresses two interrelated problems: (a) proposes an integrated local descriptor for representation to recognize human ears in 3D. Comparing local surface descriptors between a test and a model image, an initial correspondence of local surface patches is established and then filtered using simple geometric constraints. The performance of the proposed ear recognition system is evaluated on a real range image database of 52 subjects. (b) A binomial model is also presented to predict the ear recognition performance. Match and non-matched distances obtained from the database of 52 subjects are used to estimate the distributions. By modeling cumulative match characteristic (CMC) curve as a binomial distribution, the ear recognition performance can be predicted on a larger gallery.

1 Introduction

Ear is a viable new class of biometrics since the ear has desirable properties such as universality, uniqueness and permanence [1]. For example, ear is rich in features; it is a stable structure which does not change with the age; it doesn't change its shape with facial expressions, cosmetics and hair styles. Although it has above advantages over other biometrics, it has received little attention compared to other popular biometrics such as face, fingerprint and gait.

In recent years, some approaches have been developed for the ear recognition from 2D images. Burge and Burger [2] proposed an adjacency graph, which is built from the Voronoi diagram of the ear's edge segments, to describe the ear. Ear recognition is done by subgraph matching. Hurley et al. [3] applied force field transform to ear images and wells and channels are shown to be invariant to affine transformations. Chang et al. [4] used principal component analysis (PCA) to ear images. All of these works for ear recognition have used 2D intensity images and therefore the performance of their systems is greatly affected by imaging conditions such as lighting and shadow. However currently available range sensors can directly provide us 3D geometric information which is insensitive to above imaging problems. Therefore, it is desirable to design a human ear recognition system from 3D side face images obtained at a distance. In fact, different methods to design biometrics system based on 3D data have been addressed [5–10].

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 748–757, 2005.
© Springer-Verlag Berlin Heidelberg 2005

However, no existing ear recognition approaches gives theoretical or experimental performance prediction. Evaluation and prediction of the performance of biometrics system to identify individuals is always considered in real world applications. Researchers have built mathematical models to evaluate and predict the performance on biometrics such as face, fingerprint, iris and gait. Bhanu et al. [11] develop a binomial model to predict fingerprint recognition performance. Tan et al. [12] present a two-point model and a three-point model to estimate the error rate for the point based fingerprint recognition. Johnson et al. [13] build a CMC model that is based on the feature space to predict the gait identification performance. Wayman [14] derives equations for the general biometric identification system. Daugman [15] analyzes the statistical variability of iris recognition using a binomial model. Johnson et al. [16] model a CMC curve to estimate recognition performance for larger galleries. Grother et al. [17] introduce the joint density function of the match and non-match scores to predict open- and closed-set identification performance.

In this paper, we first introduce an integrated local surface descriptor for 3D ear representation. A local surface descriptor is defined by a centroid, its surface type and 2D histogram. The 2D histogram consists of shape indexes, calculated from principal curvatures, and angles between the normal of reference point and that of its neighbors. The local surface descriptors are calculated only for the feature points which are defined as the local minimum and maximum of shape indexes. By comparison of local surface descriptors between a test and a model image, correspondences of local surface patches are established and then filtered using simple geometric constraints. The initial transformation is estimated based on the corresponding surface patches and applied to randomly selected locations of model ears. Iterative closest point (ICP) algorithm [18] iteratively refines the transformation to bring model ears and test ear into best alignment. The root mean square (RMS) registration error is used as the matching error criterion.

Next, a binomial model is presented to predict the proposed ear recognition performance. We calculate the RMS registration errors between 3D ears in the probe set with 3D ears in the gallery. RMS errors are used as matching distances to estimate the distribution of match and non-match distances. Then the *cumulative match characteristic* (CMC) curve is modeled by a binomial distribution and the probability that the match score is within rank r can be calculated. Using this model we can predict ear recognition performance for a large gallery.

2 Technical Approach

2.1 Feature Extraction

In our approach, feature points are defined as local minimum and maximum of shape indexes, which can be calculated from principal curvatures. In order to estimate curvatures, we fit a quadratic surface $f(x, y) = ax^2 + by^2 + cxy + dx + ey + f$ to a local window and use the least square method to estimate the parameters of the quadratic surface, and then use differential geometry to calculate the surface normal, Gaussian and mean curvatures and principal curvatures [19].

Shape index (S_i), a quantitative measure of the shape of a surface at a point p, is defined by (1) where k_1 and k_2 are maximum and minimum principal curvatures respectively. With this definition, all shapes are mapped into the interval $[0, 1]$ [20].

$$S_i(p) = \frac{1}{2} - \frac{1}{\pi} tan^{-1} \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \tag{1}$$

Within a $w \times w$ window, the center point is marked as a feature point if its shape index is higher or lower than those of its neighbors.

2.2 Local Surface Patches

We define a "local surface patch" as the region consisting of a feature point P and its neighbors N. The neighbors should satisfy these two conditions,

$$\begin{aligned} N = \{pixels \quad & N, ||N - P|| \leq \epsilon_1\} \\ & and\ acos(n_p \bullet n_n < A), \end{aligned} \tag{2}$$

where n_p and n_n are the surface normal vectors at point P and N. The two parameters ϵ_1 and A are important since they determine how the local surface patch is resistant to clutter and occlusion. Johnson [21] discussed the choices for the two parameters. For every local surface patch, we compute the shape indexes and normal angles between point P and its neighbors. Then we can form a 2D histogram. One axis of this histogram is the shape index which is in the range [0,1]; the other is the dot product of surface normal vectors at P and N which is in the range [-1,1]. In order to reduce the effect of the noise, we use bilinear interpolation when we calculate the 2D histogram.

We also compute the centroid of local surface patches. We classify surface shape of the local surface patch into three types: concave, saddle and convex based on shape index value of the feature point. The shape index range and its corresponding surface type are listed in Table 1 [22]. Note that a feature point and the centroid of a patch may not coincide.

Table 1. Surface type Tp based on the shape index

Type tag (Tp)	S_i range	Surface type
0	[0,5/16)	Concave
1	[5/16,11/16)	Saddle
2	[11/16,1]	Convex

In summary, every local surface patch is described by a 2D histogram, surface type and the centroid. The local surface patch encodes the geometric information of a local surface.

2.3 Off-Line Model Building

Considering the uncertainty of location of a feature point, we repeat the above process to calculate descriptor of local surface patches for neighbors of feature point P and save these descriptions into the model database. For each model object, we repeat the same process to build the model database.

2.4 Surface Matching

Comparing Local Surface Patches. Given a test range image, we repeat the above steps and get local surface patches. Considering the inaccuracy of feature points' location, we also extract local surface patches from neighbors of feature points. Then we compare them with all of the local surface patches saved in the model database. This comparison is based on the surface type and histogram dissimilarity. Since histogram can be thought of as an approximation of probability distributed function, we use statistical method to assess the dissimilarity between two probability distributions. The $\chi^2 - divergence$ is among the most prominent divergence used in statistics to assess the dissimilarity between two probability density functions. We use it to measure the dissimilarity between two observed histograms Q and V, which is defined by (3) [23].

$$\chi^2(Q,V) = \sum_i \frac{(q_i - v_i)^2}{q_i + v_i} \tag{3}$$

Figure 1 shows an experimental validation that the local surface patch has the discriminative power to distinguish shapes. We do experiments under three cases. 1) a local surface patch (Lsp1) generated for an ear is compared to another local surface patch (Lsp2) corresponding to the same physical area of the same ear imaged at different viewpoints; a low dissimilarity exists. 2) The Lsp1 is compared to Lsp3 which lies in different area of the same ear; the dissimilarity is high. 3) The Lsp1 is compared to Lsp4 which lies in the similar area as the Lsp 1 but not the same ear, there exists a higher dissimilarity than the first case. The experimental results suggest that the local surface patch can be used for differentiation between ears. Table 2 lists the comparison results.

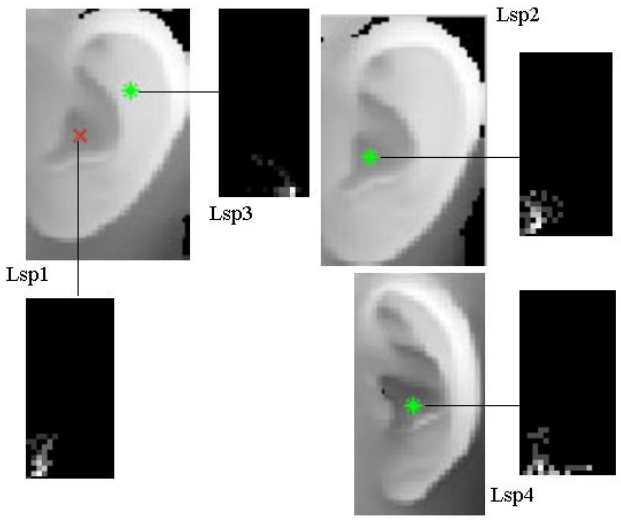


Fig. 1. Experimental validation of discriminatory power of Local Surface Patches

Table 2. Comparison results of local surface patches

Surface Type	Lsp1	Lsp2	Lsp3	Lsp4
	Tp=0	Tp=0	Tp=2	Tp=1
$\chi^2 - divergence$	$\chi^2(Lsp1, Lsp2)$	$\chi^2(Lsp1, Lsp3)$	$\chi^2(Lsp1, Lsp4)$	
	0.479	1.99	0.984	

Grouping Corresponding Pairs of Local Surface Patch. For every local surface patch from the test ear, we choose the local surface patch from the database with minimum dissimilarity and the same surface type as the possible corresponding patch. We filter the possible corresponding pairs based on the geometric constraints given below.

$$d_{C_1,C_2} = |d_{S_1,S_2} - d_{M_1,M_2}| < \epsilon_2, \tag{4}$$

where d_{S_1,S_2} and d_{M_1,M_2} are Euclidean distance between centroids of two surface patches. For two correspondences $C_1 = \{S_1, M_1\}$ and $C_2 = \{S_2, M_2\}$ where S means test surface patch and M means model surface patch, they should satisfy (4) if they are consistent corresponding pairs. Thus, we use geometric constraints to partition the potential corresponding pairs into different groups. The largest group would more likely to be the true corresponding pair.

Given a list of corresponding pairs $L = \{C_1, C_2, \ldots, C_n\}$, the grouping procedure for every pair in the list is as follows: Initialize each pair of a group. For every group, add other pairs to it if they satisfy (4). Repeat the same procedure for every group. Select the group which has the largest size.

Aligning Model Ears with Test Ears. We estimate the initial rigid transformation based on the corresponding local surface patches using quaternion representation [24]. Starting with the initial transformation, Iterative closest point (ICP) algorithm [18] is run to refine the transformation by minimizing the distance between the control points of the model ear and their closest points of the test ear. Since the ear is assumed to be in the center of the image, the control points are selected around the center of the image. Every time the control points are randomly selected from model ears and ICP is applied to those points. We repeat the same procedure several times and choose the minimum RMS error as the final result.

2.5 Prediction Model

The mathematical prediction model is based on the distribution of match and non-match scores [11]. We use $ms(x)$ and $ns(x)$ to denote the distributions of match and non-match scores. If the similarity score is higher, the match is closer. The error occurs when any given match score is less than any of the non-match scores. The probability that the non-match score is greater than or equal to the match score x is $NS(x)$ computed by (5).

$$NS(x) = \int_x^\infty ns(t)dt \tag{5}$$

The probability that the match score x has rank r exactly, is given by the binomial probability distribution:

$$\mathrm{C}_{r-1}^{N-1}(1-NS(x))^{N-r}NS(x)^{r-1} \tag{6}$$

By integrating over all the match scores, we get

$$\int_{-\infty}^{\infty}\mathrm{C}_{r-1}^{N-1}(1-NS(x))^{N-r}NS(x)^{r-1}ms(x)dx \tag{7}$$

In theory the match scores can be any value within $(-\infty,\infty)$. Therefore the probability that the match score is within rank r, which is definition of a CMC curve, is

$$P(N,r)=\sum_{i=1}^{r}\int_{-\infty}^{\infty}\mathrm{C}_{i-1}^{N-1}(1-NS(x))^{N-i}NS(x)^{i-1}ms(x)dx \tag{8}$$

In above equations N is the size of large population whose performance needs to be estimated. Here we assume that the match score and non-match score are independent and the match and non-match score distributions are the same for all the persons. The small size gallery is used to estimate distributions of $ms(x)$ and $ns(x)$.

For the ear recognition case, every 3D ear in the probe set is matched to every 3D ear in the gallery and the RMS registration error is calculated using the procedure described in Section 2.4. The RMS registration error is used as matching error criterion. In our case, the matching error is smaller, the match is closer. In order to use the above prediction model, we modify the equations accordingly.

3 Experimental Results

3.1 Data Acquisition

We use real range data acquired using Minolta Vivid 300. The range image contains 200×200 grid points and each grid point has a 3D coordinate (x, y, z). There are 52 subjects in our database and every subject has two left side face range images taken at different viewpoints. The ears are manually extracted from side face images. The data is split into model and test sets. Each set has 52 ears. The extracted model ears and corresponding test ears are shown in Figure 2 and Figure 3 respectively.

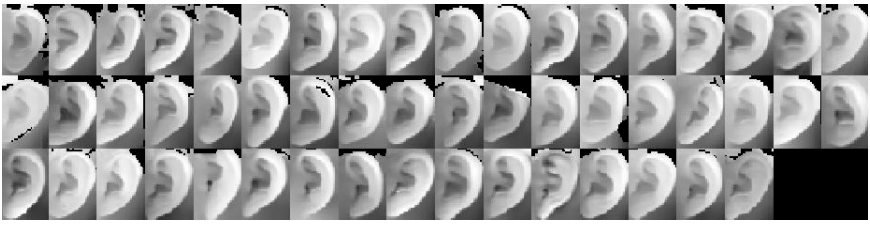


Fig. 2. 3D model ears shown as gray scale images

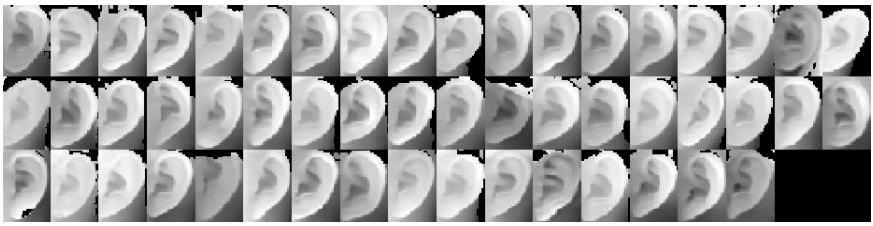
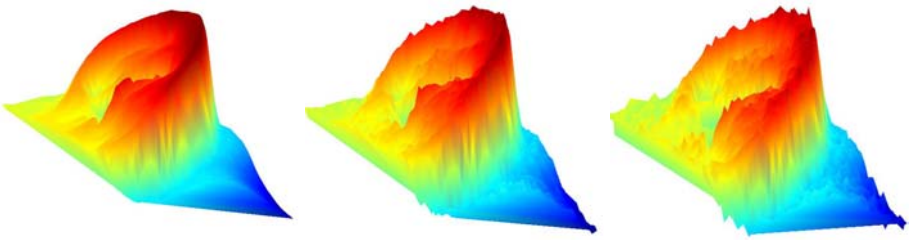


Fig. 3. 3D test ears shown as gray scale images



(a) No added noise (b) $\sigma = 0.6mm$ (c) $\sigma = 1.0mm$

Fig. 4. Test scans corrupted with Gaussian noise

3.2 Performance Evaluation

To test the proposed system's performance, we add Gaussian noise to the test scans along the viewing direction (Z-axis). The standard deviation of Gaussian noise we add depends on the mesh resolution of test scans. However the mesh resolution is not well defined. We use the Johnson's definition [21] "Mesh resolution is defined as the median of all edge lengths in a mesh". Given a test range image, we triangulate it and get a triangular mesh. Then we calculate the median of all edge lengths in the mesh. The average median calculated from test scans is about $1.25mm$. We add Gaussian noise with $\sigma = 0.6mm$ and $\sigma = 1.0mm$ to test scans. Therefore, we have three probe sets: one probe set has no added Gaussian noise; the second probe set has Gaussian noise $N(0, \sigma = 0.6mm)$; the third probe set has Gaussian noise $N(0, \sigma = 1.0mm)$. Examples of one test scan corrupted with Gaussian noise are shown in Figure 4.

The CMC curve is used to evaluate the system's recognition performance. The rank-1, rank-2 and rank-3 recognition rates for three probe sets are listed in Table 3. The verification performance results are given by the receiver operating characteristic (ROC)

Table 3. Recognition results for three probe sets

Probe set	Rank-1	Rank-2	Rank-3
No added noise	90.4%	96.2%	96.2%
$N(0, \sigma = 0.6mm)$	76.9%	86.5%	86.5%
$N(0, \sigma = 1.0mm)$	44.2%	61.5%	67.3%

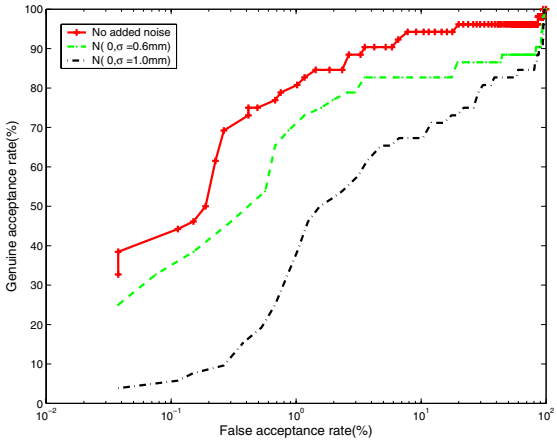


Fig. 5. ROC curve for three probe sets

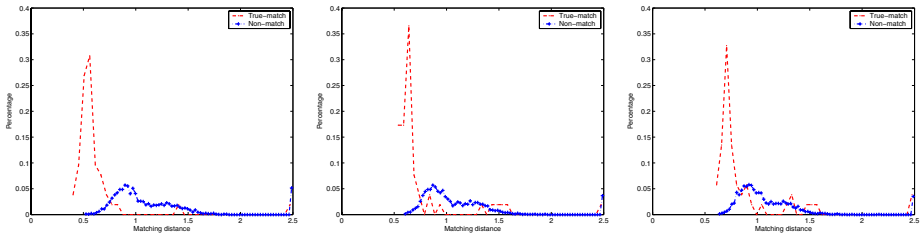
curve which is defined as the plot of genuine acceptance rate against false acceptance rate. Figure 5 shows the ROC curves for three probe sets. From Table 3 and Figure 5, we can see that the performance of the proposed system degrades as the scene noise increases. It's reasonable since Gaussian noise corrupts the surface normals and shape index resulting in the corruption of the local surface patch representation.

3.3 Prediction Results

Every 3D scan in three probe sets is matched to every 3D ear in the gallery and the RMS registration error is calculated using the procedure described in Section 2.4. The RMS registration error is used as the matching distance. Therefore, we obtain 52 true-match distances and 2652 non-match distances for every probe set. The matching distance distribution for true-match and non-match for three probe sets are shown in Figure 6. Based on the distributions, we can predict CMC curve $P(N, r)$ where $r = 1, 2, 3$ and N is 52. We also calculate the CMC curve based on the experimental results for three probe sets. The results of the directly calculated CMC curve and the predicted CMC curve are shown in Table 4. Table 4 shows that the predicted CMC values are close to the calculated CMC values, which demonstrates the effectiveness of our prediction model. We'd like to predict CMC values for larger galleries from the original range image database of 52 subjects. Table 5 shows the predicted CMC values for three probe

Table 4. Predicted and calculated CMC values for three probe sets on 52 subjects

Probe set	Rank-1		Rank-2		Rank-3	
	Predicted	Calculated	Predicted	Calculated	Predicted	Calculated
No added noise	92.5%	90.4%	94.6%	96.2%	95.7%	96.2%
$N(0, \sigma = 0.6mm)$	80.4%	76.9%	83.9%	86.5%	85.8%	86.5%
$N(0, \sigma = 1.0mm)$	51.5%	44.2%	60.2%	61.5%	66.1%	67.3%



(a) (b) (c)

Fig. 6. Matching distance distribution for match and non-match pairs for three probe sets: (a) without added noise, (b) with Gaussian noise $N(0, \sigma = 0.6mm)$, (c) with Gaussian noise $N(0, \sigma = 1.0mm)$

Table 5. Predicted CMC values for three probe sets for larger galleries (Recognition rate shown as percentage)

Probe set	N=100			N=200			N=300			N=400		
	R-1	R-2	R-3	R-1	R-2	R-3	R-1	R-2	R-3	R-1	R-2	R-3
No added noise	91.2	92.8	94.4	90.5	90.9	91.8	90.4	90.5	90.8	90.4	90.4	90.5
$N(0, \sigma = 0.6mm)$	78.3	80.9	83.6	77.1	77.9	79.3	76.9	77.1	77.6	76.9	76.9	77.1
$N(0, \sigma = 1.0mm)$	46.8	52.1	57.9	44.6	45.9	48.6	44.3	44.6	45.4	44.2	44.3	44.5

sets for different gallery size (N=100, 200, 300, 400). Confidence in prediction is of interest and we plan to work on it.

4 Conclusions

In this paper, we first propose an integrated local descriptor for representation to recognize human ears in 3D. We evaluate the proposed ear recognition performance by means of CMC and ROC curves on three different probe sets using a real range image database of 52 subjects. One probe set has no added Gaussian noise; the second probe set has Gaussian noise $N(0, \sigma = 0.6mm)$; the third probe set has Gaussian noise $N(0, \sigma = 1.0mm)$. We obtain rank-one recognition rate of 90.4% for test scans without added noise and the system's performance degrades as the scene noise increases. We also predict the ear recognition performance on larger galleries by modeling cumulative match characteristic curve as a binomial distribution. The predicted rank-one recognition rate is 90.4% on test scans without added noise for a database of 400 subjects. Table 5 demonstrates that we can predict the recognition performace for lager galleries.

References

1. Iannarelli, A.: Ear Identification. Forensic Identification Series. Paramont Publishing Company (1989)
2. Burge, M., Burger, W.: Ear biometrics in computer vision. Proc. Int. Conf. on Pattern Recognition **2** (2000) 822–826

3. Hurley, D., Nixon, M., Carter, J.: Automatic ear recognition by force field transformations. IEE Colloquium on Visual Biometrics (2000) 7/1 –7/5
4. Chang, K.C., Bowyer, K.W., Sarkar, S., Victor, B.: Comparison and combination of ear and face images in appearance-based biometrics. IEEE Trans. Pattern Analysis and Machine Intelligence **25** (2003) 1160–1165
5. Bhanu, B., Chen, H.: Human ear recognition in 3D. Workshop on Multimodal User Authentication (2003) 91–98
6. Bronstein, A., Bronstein, M., Kimmel, R.: Expression-invariant 3D face recognition. Audio and Video based Biometric Person Authentication (2003) 62–70
7. Chang, K.C., Bowyer, K.W., Flynn, P.J.: Multi-modal 2D and 3D biometrics for face recognition. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (2003) 187–194
8. Chua, C.S., Han, F., Ho, Y.: 3D human face recognition using point signatures. Int. Conf. on Automatic Face and Gesture Recognition (2000) 233–238
9. Lee, J.C., Milios, E.: Matching range images of human faces. Proc. Int. Conf. on Computer Vision (1990) 722–726
10. Lu, X., Colbry, D., Jain, A.K.: Three-dimensional model based face recognition. Proc. Int. Conf. on Pattern Recognition **1** (2004) 362–366
11. Bhanu, B., Wang, R., Tan, X.: Predicting fingerprint recognition performance from a small gallery. ICPR Workshop on Biometrics: Challenges arising from Theory to Practice (2004) 47–50
12. Tan, X., Bhanu, B.: On the fundamental performance for fingerprint matching. Proc. IEEE Conf. Computer Vision and Pattern Recognition **2** (2003) 499–504
13. Johnson, A.Y., Sun, J., Boick, A.F.: Predicting large population data cumulative match characteristic performance from small population data. Audio and Video based Biometric Person Authentication (2003) 821–829
14. Wayman, J.L.: Error-rate equations for the general biometric system. IEEE Robotics & Automation Magazine **6** (1999) 35–48
15. Daugman, J.: The importance of being random: statistical principles of iris recognition. Pattern Recognition **36** (2003) 279–291
16. Johnson, A.Y., Sun, J., Boick, A.F.: Using similarity scores from a small gallery to estimate recognition performance for large galleries. IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures (2003) 100–103
17. Grother, P., Phillips, P.J.: Models of large population recognition performance. Proc. IEEE Conf. Computer Vision and Pattern Recognition **2** (2004) 68–75
18. Besl, P., Mckay, N.D.: A method of registration of 3-D shapes. IEEE Trans. Pattern Analysis and Machine Intelligence **14** (1992) 239–256
19. Flynn, P., Jain, A.: On reliable curvature estimation. Proc. IEEE Conf. Computer Vision and Pattern Recognition (1989) 110–116
20. Dorai, C., Jain, A.: COSMOS-A representation scheme for free-form surfaces. Proc. Int. Conf. on Computer Vision (1995) 1024–1029
21. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Analysis and Machine Intelligence **21** (1999) 433–449
22. Koenderink, J.J., Doorn, A.V.: Surface shape and curvature scales. Image Vision Computing **10** (1992) 557–565
23. Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. International Journal of Computer Vision **36** (2000) 31–50
24. Horn, B.: Close-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America **4** (1987) 629–642

Optimal User Weighting Fusion in DWT Domain On-Line Signature Verification

Isao Nakanishi[1], Hiroyuki Sakamoto[2], Yoshio Itoh[2], and Yutaka Fukui[2]

[1] Faculty of Regional Sciences, Tottori University, Japan
isao@rstu.jp
[2] Faculty of Engineering, Tottori University, Japan
{itoh,fukui}@ele.tottori-u.ac.jp

Abstract. DWT domain on-line signature verification method has bee proposed. Time-varying pen-position signal is decomposed into sub-band signals by using the DWT. Individual features are extracted as high frequency signals in sub-band. By using the extracted feature, verification is achieved at each sub-band and then total decision is done by combining such verification results. In this paper, we introduce a user weighting fusion into the total decision for improving verification performance. Through many verification experiments, it is confirmed that there is an optimal weight combination for each user and verifiaction rate can be improved when the optimal weight combination is applied. Such the optimal weight combination also becomes an individual feature which can not be known by others.

1 Introduction

Recently, multiple biometric systems have been attracted attentions to improve the performance of single biometric systems. Five scenarios of the multiple biometric system are considered in [1], that is, multi-sensor system, multi-modal system, multi-unit system, multi-impression system, and multi-matcher system. Among of them, the multi-matcher system which uses multiple representation and matching algorithm for the same input biometric signal is the most cost-effective way to improve the performance of the biometric system [1]. In addition, the multi-matcher system requires capturing biometrics only once.

We have proposed the on-line signature verification system in the Discrete Wavelet Transform (DWT) domain [2, 3]. This system utilized only pen-position parameter, that is, x and y coordinates since it was detectable even in portable devices such as the Personal Digital Assistants (PDA). Each time-varying signal of x and y coordinates was decomposed into sub-band signals by using the DWT. Verification was achieved by using the adaptive signal processing in each sub-band. Total decision for verification was done by averaging the verification results of several sub-bands in x and y coordinates. Verification rate was about 95%, which was improved by about 10% comparing with a time-domain verification system.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 758–766, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Our proposed system is regarded as the multi-matcher system. In general, the multi-matcher system combines at most a few verification results [1]. On the other hand, the verification of our proposed system is achieved at several sub-bands in both x and y coordinates; therefore, there are much more verification results than general multi-matcher systems. This enables to adopt more unrestrained weighting of the verification results. If an optimal weighting for each user (signature) is applied in the total decision, the verification rate is expected to be improved. In this paper, we introduce a user weighting fusion into the total decision. Through many verification experiments, it is confirmed that there is an optimal weight combination for each signature and the verifiaction rate is improved when the optimal weight combination is applied. Moreover, the optimal weight combination also becomes an individual feature which can not be known by others.

2 On-Line Signature Verification in DWT Domain

2.1 On-Line Signature

The on-line signature is digitized with the electronic pen-tablet. Especially, we utilize only pen-position parameter since it is provided even in such as the PDA for handwriting or pointing. Actually, the pen-position parameter consists of discrete time-varying signals of x and y coordinates, which are $x^*(n^{'})$ and $y^*(n^{'})$, respectively. $n^{'}(= 0, 1, \cdots, N_{max}-1)$ is a sampled time index. N_{max} is the total number of sampled data. As the one-line signature is a dynamic biometrics, each writing time is different from the others. This results in the different number of sampled data even in genuine signatures. Moreover, different writing place and different size of signature cause variations in pen-position parameter. To reduce such variations, pen-position data are normalized in general. The normalized pen-position parameter is defined as

$$x(n) = \frac{x^*(n) - x_{min}}{x_{max} - x_{min}} \cdot \alpha_x \tag{1}$$

$$y(n) = \frac{y^*(n) - y_{min}}{y_{max} - y_{min}} \cdot \alpha_y \tag{2}$$

where $n(= 0 \sim 1)$ is a normalized sampled time index given by $n = n^{'}/(N_{max} - 1)$. x_{max} and y_{max} are maximum and minimum values of $x^*(n)$ and $y^*(n)$, respectively. α_x and α_y are scaling factors for avoiding underflow calculation in sub-band decomposition described later.

However, such normalization makes the difference between a genuine signature and its forgery unclear. In addition, the on-line signature is relatively easy to forge if the written signature is known. Easiness of imitating pen-position data decreases the difference between the genuine signature and the forgery further. Figure 1 shows examples of the time-varying signal of x coordinate in a genuine signature and its forgery. The forgery data was obtained by tracing the

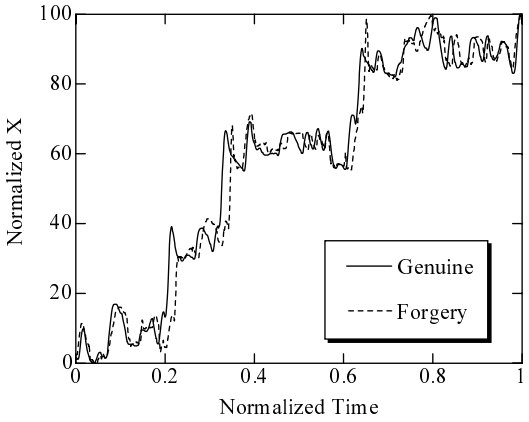


Fig. 1. Examples of the time-varying signal of x coordinate

genuine signature. It is clear that to distinguish between the genuine signature and the forgery is difficult by using the time-varying signal of the pen-position parameter.

2.2 Feature Extraction by Sub-band Decomposition

In order to enhance the difference between a genuine signature and its forgery, we have proposed to verify the on-line signature in DWT domain [2, 3]. In the following, $x(n)$ and $y(n)$ are represented as $v(n)$ for convenience. The DWT of the normalized pen-position $v(n)$ is defined as [4]

$$u_k(m) = \sum_n v(n)\overline{\Psi_{k,m}(n)} \tag{3}$$

where $\Psi_{k,m}(n)$ is the wavelet function and $\bar{\cdot}$ denotes the conjugate. k is a frequency (level) index.

Moreover, it is well known that the DWT corresponds to the octave-band filter bank. Figure 2 shows a parallel structure of the sub-band decomposition where M_d is a decomposition level and is set to guarantee the following relation

$$2^{M_d+1} \leq N_{tmp} < 2^{M_d+2} \tag{4}$$

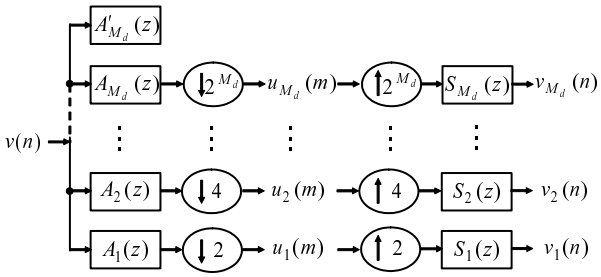


Fig. 2. Parallel structure of sub-band decomposition by DWT

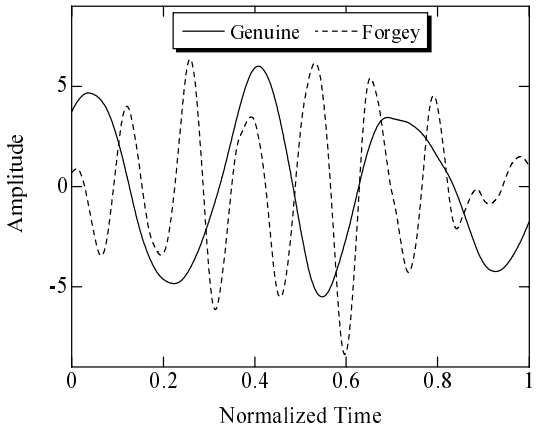


Fig. 3. Examples of *Detail*

N_{tmp} is the number of sampled data of pen-position template described later. Also, M_d has the upper limit: M_d^{max}. The synthesized signal $v_k(n)$ $(k = 1, 2, \cdots, M_d)$ is called *Detail*. The *Detail* is the signal in high frequency band and so it contains differences between signals. Therefore, we consider the *Detail* as an enhanced individual feature in pen-position.

Figure 3 shows examples of the *Detail* [2, 3]. We can confirm that the difference between a genuine signature and its forgery become remarkable by the sub-band decomposition even if the genuine signature is traced by the forger.

2.3 Verification System

Figure 4 shows a system overview. Pen-position, actually x and y coordinates are separately processed in verification block. Figure 5 describes the verification block. Firstly, the time-varying signal of x or y coordinate is decomposed into *Detail*s and then each *Detail* is verified with a corresponding template using the adaptive signal processing at each sub-band level.

Before verification, templates must be enrolled to be compared with input signatures. As the template, T genuine signatures which have equal number of

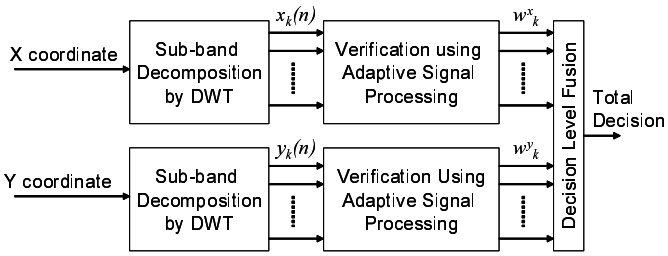


Fig. 4. System overview

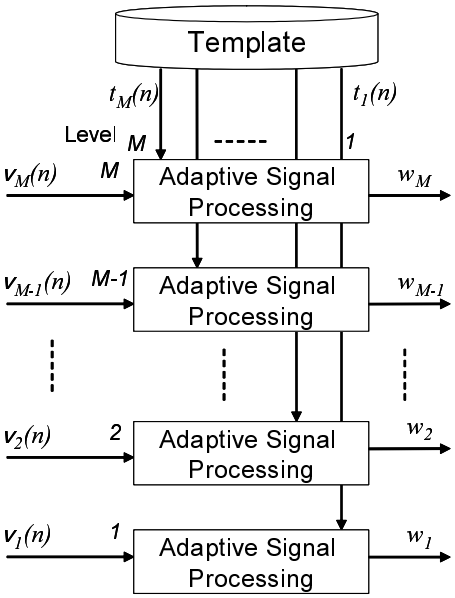


Fig. 5. Verification block

strokes are prepared and then their pen-position parameter is decomposed into sub-band signals by the DWT each other. Decomposition level is decided after examinations of those genuine signatures. Extracted *T Details* are averaged at the same level each other.

By the way, if the number of strokes in an input signature is different from that in a template, it is natural to consider the input signature as a forgery. However, not all genuine signatures have the same number of strokes. We adopt the dynamic programming (DP) matching method to identify the number of strokes in an input signature with that in a template. The procedure of the stroke matching is omitted for lack of space. It is described in detail in [2, 3].

2.4 Verification Using Adaptive Signal Processing

After enrollment of the template, verification is achieved by using the adaptive signal processing. The purpose of the adaptive signal processing is to reduce the error between the input signal and the desired signal sample by sample [5]. When an input signal is of a genuine signature, the error between the input and its template becomes small; therefore, adaptive weights are expected to converge close on 1. Inversely, if the input signature is a forgery, adaptive weights converge far from 1. In this way, the verification can be achieved by examining whether converged value is nearly 1 or not [2, 3].

As the adaptive algorithm, we use a new kind of steepest descent algorithm [5] defined as follows.

$$w_k(n+1) = w_k(n) + \mu E\left[e_k(n) v_k(n)\right] \tag{5}$$

$$e_k(n) = t_k(n) - w_k(n)v_k(n) \tag{6}$$

$$E\left[e_k(n)v_k(n)\right] = \frac{1}{N_{tmp}} \sum_{l=0}^{N_{tmp}-1} e_k(r-l)\ v_k(r-l) \tag{7}$$

$$\mu = \mu_0 / \left\{E\left[|v_k(n)|\right]\right\}^2 \tag{8}$$

$$E\left[|v_k(n)|\right] = \frac{1}{N_{in}} \sum_{l=0}^{N_{in}-1} v_k(n-l) \tag{9}$$

where N_{in} is the number of sampled data in an input *Detail*. N_{tmp} is the number of sampled data in a template. μ is a step size parameter which controls the convergence in the adaptive algorithm. The step size parameter is normalized by input power as shown in Eqs.(8) and (9), so that convergence is always guaranteed. μ_0 is a positive constant.

The verification is done in all sub-bands in parallel. After enough iterations for convergence, $w_k(n)$ is averaged in past N_{tmp} samples and then we obtain the converged value w_k.

Total verification score (TS) is obtained by combining converged values at several sub-band levels in x and y coordinates.

$$\mathrm{TS} = c_x \left(\sum_{p=0}^{L-1} f_p \cdot w^x_{M-p}\right) + c_y \left(\sum_{p=0}^{L-1} f_p \cdot w^y_{M-p}\right) \tag{10}$$
$$c_x + c_y = 1, \quad c_x > 0, c_y > 0, \quad \Sigma f_p = 1, \quad f_p > 0$$

where w^x_{M-p} and w^y_{M-p} respectively denote the converged values of x and y coordinates at level $M-p$. L is the number of used sub-band levels in decision fusion. c_x and c_y are the weights for x and y coordinates, respectively and f_p is the weight for sub-band.

In our conventional results, we set $c_x = c_y = 1/2$ and $f_p = 1/L$, that is, the total verification score was obtained by averaging all converged values. In that case, verification rate was about 95% [2, 3].

3 User Weighting Fusion

In our proposed system, total verification score is obtained by fusing $2 \times L$ converged values. In other words, it is possible to set the weights more unrestrained than the time-domain verification system which has only c_x and c_y.

There have been proposed many fusion methods such as the sum rule, the minimum score, the maximum score and so on [6]. In this paper, we introduce user weighting fusion into the total decision for verification. The total verification score is re-defined as

$$\mathrm{TS}^i = c^i_x \left(\sum_{p=0}^{L-1} f^i_p \cdot w^x_{M-p}\right) + c^i_y \left(\sum_{p=0}^{L-1} f^i_p \cdot w^y_{M-p}\right) \tag{11}$$
$$c^i_x + c^i_y = 1, \quad c^i_x > 0, c^i_y > 0, \quad \Sigma f^i_p = 1, \quad f^i_p > 0$$

where i $(i = 1, 2, \cdots, I)$ presents enrolled user (signature) identification number. In general verification systems, such a user identifier is used for one-to-one matching between an input and its template [7]. The user weighting fusion enables to set optimal weights for each user.

Next, in order to find such optimal weights, we carried out verification experiments in various weight combinations. In this experiment, we assumed the following severe situation. Before signing, the subjects were called upon to practice using the pen tablet for becoming skilled. This suppresses the variation of signature due to inexperienced pen-tablet. When the subjects signed genuine signatures, they were not able to refer to their already written signatures. This tends to increase the intra-class variation in signatures of one individual. On the other hand, assuming that the signature shape was easily imitated, forgers were permitted to trace the genuine signature by putting the paper to which the signature was written over the pen tablet.

On the above situation, we prepared an original database. Four subjects were requested to sign their own signatures and then we obtained 118 genuine signatures. The four subjects were labeled "a", "b", "c" and "d" in the following. Five genuine signatures for each subject were used to make a template and the remaining 98 genuine signatures were used for verification. Five subjects were required to counterfeit the genuine signature 10 times each, so that 200 forgeries were prepared in total.

Other conditions of simulation are summarized as follows.

- Scaling parameter: $\alpha_x = \alpha_y = 100$
- Wavelet function: Daubechies8
- Number of signatures for making a template: $T = 5$
- Upper limit decomposition level: $M^{max} = 8$
- Number of processed level: $L = 4$
- Step size constant: $\mu_0 = 0.0001$
- Number of iterations: 10^5

The weight for pen-position was changed from 0.0 to 1.0 every 0.1. Also, three combinations of weight for sub-band, (0.1, 0.2, 0.3, 0.4), (0.25, 0.25, 0.25, 0.25), (0.4, 0.3, 0.2, 0.1) were examined. Totally 33 weight combinations were evaluated. Verification performance was estimated by the Equal Error Rate (EER) where the False Rejection Rate (FRR) is equal to the False Acceptance Rate (FAR).

Results are shown in Table 1. When the case of $c_x = c_y = 0.5$ and $f_3 = f_2 = f_1 = f_0 = 0.25$ corresponds to the conventional setting. In that case, the total EER was 5% [2, 3].

Next, we defined an optimal combination as the weights which achieved the smallest EER and made it easier to set threshold value in total decision using the FAR and FRR curves. The optimal weight combinations are summarized in Table 2. Total EER was 4%. As a result, user optimal weighting improved the total EER by 1%.

It is interesting that each user (signature) has different optimal weight combination and the EER can be greatly decreased when the optimal weight is applied. Especially, the weight combination for user "b" is contrary to that for user "d".

Table 1. Weight combination vs. EER

Weights for pen-position		Weights for sub-band				EER(%)			
c_x	c_y	f_3	f_2	f_1	f_0	a	b	c	d
0.0	1.0	0.1	0.2	0.3	0.4	12.0	0.0	6.8	5.0
0.1	0.9	0.1	0.2	0.3	0.4	9.0	0.0	4.2	3.5
0.2	0.8	0.1	0.2	0.3	0.4	6.5	0.0	6.8	5.5
0.3	0.7	0.1	0.2	0.3	0.4	4.0	0.0	6.5	6.0
0.4	0.6	0.1	0.2	0.3	0.4	4.0	0.0	8.2	4.0
0.5	0.5	0.1	0.2	0.3	0.4	2.0	0.0	8.2	6.0
0.6	0.4	0.1	0.2	0.3	0.4	2.5	0.0	8.2	6.0
0.7	0.3	0.1	0.2	0.3	0.4	1.8	0.0	8.2	6.0
0.8	0.2	0.1	0.2	0.3	0.4	2.0	0.0	8.2	11.5
0.9	0.1	0.1	0.2	0.3	0.4	2.0	0.0	9.5	4.0
1.0	0.0	0.1	0.2	0.3	0.4	2.5	0.0	12.5	14.3
0.0	1.0	0.25	0.25	0.25	0.25	10.5	0.0	5.5	2.0
0.1	0.9	0.25	0.25	0.25	0.25	9.5	0.0	4.2	2.0
0.2	0.8	0.25	0.25	0.25	0.25	7.0	0.0	5.0	2.0
0.3	0.7	0.25	0.25	0.25	0.25	5.2	0.0	7.5	1.8
0.4	0.6	0.25	0.25	0.25	0.25	3.0	0.0	8.2	2.5
0.5	0.5	0.25	0.25	0.25	0.25	2.0	0.0	8.2	3.5
0.6	0.4	0.25	0.25	0.25	0.25	1.3	0.0	8.2	4.8
0.7	0.3	0.25	0.25	0.25	0.25	2.0	0.0	8.2	4.0
0.8	0.2	0.25	0.25	0.25	0.25	1.6	0.0	8.2	6.0
0.9	0.1	0.25	0.25	0.25	0.25	2.0	0.0	8.2	8.5
1.0	0.0	0.25	0.25	0.25	0.25	3.0	0.0	8.2	12.0
0.0	1.0	0.4	0.3	0.2	0.1	8.0	0.0	4.2	0.0
0.1	0.9	0.4	0.3	0.2	0.1	8.0	0.0	4.2	0.0
0.2	0.8	0.4	0.3	0.2	0.1	8.0	0.0	6.0	0.0
0.3	0.7	0.4	0.3	0.2	0.1	5.5	0.0	6.0	0.0
0.4	0.6	0.4	0.3	0.2	0.1	4.0	0.0	8.2	0.0
0.5	0.5	0.4	0.3	0.2	0.1	4.0	0.0	8.2	2.8
0.6	0.4	0.4	0.3	0.2	0.1	2.8	0.0	9.5	2.8
0.7	0.3	0.4	0.3	0.2	0.1	3.0	0.0	9.5	4.0
0.8	0.2	0.4	0.3	0.2	0.1	1.5	1.5	10.0	4.2
0.9	0.1	0.4	0.3	0.2	0.1	3.0	3.0	10.5	4.2
1.0	0.0	0.4	0.3	0.2	0.1	4.0	4.0	12.0	4.2

In the case of user "b", verification results at lower levels have more effect on verification performance than those at higher levels. Inversely, the verification results at higher levels play an important role in the total decision in user "d". These matters depend on the figure of signature and the user's habit in writing process. In other words, the optimal weight combination is also an individual feature which can not be known by others.

Table 2. Optimal user weighting

User	Weights for pen-position		Weights for sub-band				EER
	c_x	c_y	f_3	f_2	f_1	f_0	(%)
a	0.6	0.4	0.25	0.25	0.25	0.25	1.3
b	0.7	0.3	0.1	0.2	0.3	0.4	0.0
c	0.1	0.9	0.4	0.3	0.2	0.1	4.2
d	0.1	0.9	0.4	0.3	0.2	0.1	0.0

4 Conclusion

We introduced user weighting fusion into the total decision in the DWT domain on-line signature verification. Verification experiments showed that there was an optimal weight combination for each user and then verifiaction rate could be improved when the optimal weights were applied. In addition, the optimal weight combination is expected to be a new individual feature which can not be known by others. As amount of data of optimal weight combinations is quite small, they can be enrolled in the database as well as the template. It is easy to implement the proposed optimal fusion method in the on-line signature verification system.

In this evaluation, we used not only genuine signatures but also their forgeries. However, it may not be realistic for a real system. It must be studied to develop some statistical method for determining optimal weights by using only genuine signatures. Moreover, we will study to implement our on-line signature verification system in a portable device such as the PDA in the near future.

References

1. S. Prabhakar, A. K. Jain, "Decision-level Fusion in Fingerprint Verification," *Pattern Recognition*, vol.35, pp.861-874, 2002.
2. I. Nakanishi, N. Nishiguchi, Y. Itoh, and Y. Fukui, "On-line Signature Verification Method Based on Discrete Wavelet Transform and Adaptive Signal Processing," *Proc. of Workshop on Multimodal User Authentication*, Santa Barbara, USA, pp.207-214, Dec. 2003.
3. I. Nakanishi, N. Nishiguchi, Y. Itoh, and Y. Fukui, "On-line signature verification based on subband decomposition by DWT and adaptive signal processing," Electronics and Communications in Japan (Part III: Fundamental Electronics Science), vol.88, no.6, 2005.
4. G. Strang, T. Nguyen, Wavelet and Filter Banks, Wellesley-Cambridge Press, Massachusetts, 1997.
5. S. Haykin, Introduction to Adaptive Filters, Macmillan Publishing Company, New York, 1984.
6. M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," *Proc. of Workshop on Multimodal User Authentication*, Santa Barbara, USA, pp.99-106, Dec. 2003.
7. A. K. Jain, F.D. Griess, and S.D. Connell, "On-Line Signature Verification," *Pattern Recognition*, vol.35, pp.2963-2972, 2002.

Gait Recognition Using Spectral Features of Foot Motion

Agus Santoso Lie[1], Ryo Shimomoto[1], Shohei Sakaguchi[1], Toshiyuki Ishimura[1], Shuichi Enokida[1], Tomohito Wada[2], and Toshiaki Ejima[1]

[1] Kyushu Institute of Technology
[2] National Institute of Fitness and Sports in Kanoya

Abstract. Gait as a motion-based biometric has the merit of being non-contact and unobtrusive. In this paper, we proposed a gait recognition approach using spectral features of horizontal and vertical movement of ankles in a normal walk. Gait recognition experiments using the spectral features in term of the magnitude, phase and phase-weighted magnitude show that both magnitude and phase spectra are effective gait signatures, but magnitude spectra are slightly superior. We also proposed the use of geometrical mean based spectral features for gait recognition. Experimental results with 9 subjects show encouraging results in the same-day test, while the effect of time covariate is confirmed in the cross-month test.

1 Introduction

Advances in sensor technologies enable us to create systems with a capability to sense and collect information related to human activities within an operational space, and use the information for surveillance or human machine interaction purposes. Such systems need to be able to translate the sensor input into abstract descriptions about the human subject or the activity involved. For example, a system with gait recognition capability may tell who is doing a certain activity based on the observation of the person's gait.

Gait recognition as a motion-based biometric has the merit of being non-contact and unobtrusive. Therefore, privacy issues that often arise in biometrics using facial image can be avoided. In addition, the process can be accomplished from a distance and the subjects do not have to initiate or even be aware or distracted by the procedure.

Interests in gait recognition are supported by the experiments using Moving Light Displays (MLDs), showing that while human cannot recognize gait from a single static image in MLDs, a sequence of MLDs frames provides enough information for differentiating gaits [1]. Further experiments also showed that from MLDs information human could identify their friends and discriminate genders [2].

This paper focuses on using the foot motion pattern of normal walks for personal identification. From the planar projections of the three-dimensional motions of fixed points at left and right feet, such as ankle points, we extract spectral features which represent the horizontal and vertical dynamics of a gait and investigate the discriminatory characteristics of these cues for personal identification. We also investigate how each vertical and horizontal movement of the ankles contributes in identifying the individuals. Utilizing ankles points is advantageous, because the features are relatively easy to track in a video input, and are not sensitive to changes in clothing styles.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 767–776, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Related Work

Gait holds individual specific characteristics in both the structural and transitional sense. Each time capture of a gait represents the human body structure, which is static in nature. The human gait also has the transitional characteristics that define the dynamics of the changes in the gait structural view. Therefore, a gait can be represented as a transition of poses in a state-space model [5], [7] or as a feature that reflects the spatiotemporal distribution of the gait as a continuum [8], [9], [10].

There are two prominent methods for gait recognition: model-based and appearance-based. In model-based approaches [3], [6], [12], the observation at each frame is fitted into an explicit structural model of human body, and recognition is achieved from the analysis of the trajectories of high-level body parts. Niyogi *et al.* [3] used a spatiotemporal analysis for gait detection and constructed a stick model of human body for gait recognition. In [6], [12], the human thigh and lower leg are modeled as a pendulum joint at the knee and the Fourier descriptions of the periodic change of thigh and lower leg orientations are used to form gait signature. In the definition of gait signature, they considered both the magnitude and the phase spectra, as the phase spectra of frequency components of low magnitude are insignificant.

Appearance-based approaches mostly use silhouette features, and are more sensitive to changes in clothing styles and noise in human segmentation process. Wang *et al.* [10] used a distance signal to encode a silhouette into 1D form and then reduced the dimensionality using Principal Components Analysis Method (PCA). Mowbray and Nixon [11] described the boundary of a silhouette as Fourier descriptors. He [7] used Hu moments to represent the silhouette at each frame and utilized HMM for the recognition stage. Lee and Grimson [9] segmented a silhouette into seven blobs and then used the moment features of each blob and the silhouette centroid height as the feature vector. To aggregate these appearance-based features across time in a gait sequence, they proposed three methods: Gaussian representation, histogram representation and fundamental spectral decomposition. Their approach is robust for not systematically biased motion segmentation noise, but cross-day evaluations showed that substantial changes in clothing style result in poor recognition. The effects of time and other covariates on gait recognition are investigated extensively in [14], using a baseline algorithm on a large database.

MLDs has been used intensively for investigating the potential of spectral analysis of gait for classification [13] and recognition [4]. Li and Holstein [13] showed that frequency domain method can be used effectively for classifying periodic motions such as walking, running, jumping and skipping. They utilized the power spectra of the vertical components of 16 unidentified feature points of human body. The power spectrum of each point is then reduced to into 7 dimensional features, expressing the average height from the floor, the total activeness of the respective body part and the power distribution in the neighborhood of the 1^{st} to the 3^{rd} gait-cycle (Gc). Lakany and Hayes [4] used the trajectories of the head and 12 joints of a walking subject to recognize the walker. The magnitude spectra of the trajectories are used as feature vectors for a neural network built for discriminating walkers. Their experiments using a data set of 4 individuals showed a promising result.

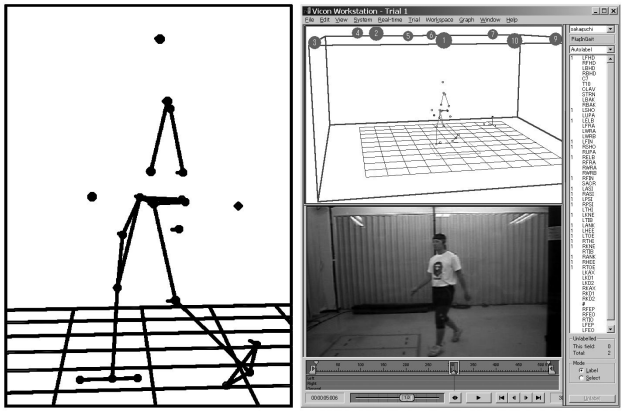


Fig. 1. The motion capture environment

3 Data Acquisition

We acquired the 3D data used for our experiments using Vicon optical motion capture system, which provides high accuracy positional measurements of markers attached to the body. The capture environment is shown in Fig. 1.

Nine subjects (all males) were asked to walk in their normal speed. The subjects are about 22-30 years old, 160-182 cm in height, and weight 54-130 kilograms. The data were taken in four separate sessions. The first two sessions were taken on the same day with a 30 minutes interval in between. The third and forth sessions were taken on two separate days, three months after the first two sessions. Ten gait sequences per person were sampled in a sampling rate of 60 Hz at each session. Some of the gaps in point trajectories due to capture failures are filled by interpolation. We conduct our experiments using 357 gait sequences that have trajectories of left and right ankles in more than 64 consecutive frames. The average length of the sequences is 150 frames.

4 Spectral Features as Gait Representation

4.1 Spectral Analysis

Spectral analysis represents a time domain signal as the sum of its sinusoidal components. The Fourier decomposition of a continuous time signal $x(t)$ is given by

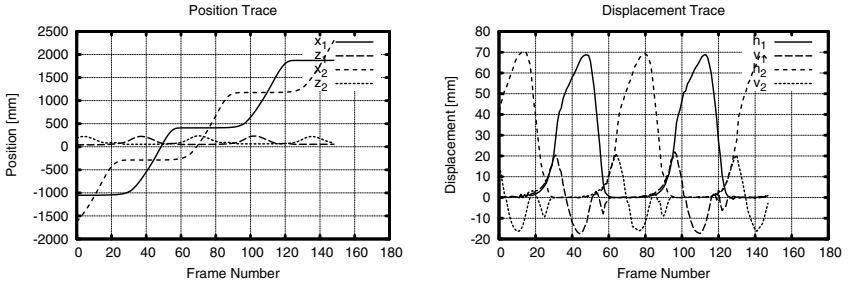
$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega, \tag{1}$$

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \tag{2}$$

where ω denotes the angular frequency of a sinusoidal component.

4.2 Motion Signals and Spectral Features as Gait Representation

Among the temporal MLDs observations of N points $P_i(t) = (x_i(t), y_i(t), z_i(t))^T$, we will only consider the movement of the left and right ankles. Each displacement vector $\boldsymbol{r}_i(t) = P_i(t) - P_i(t-1)$ approximates the velocity of a point at time t. The displacement vectors should be expressed in a consistent coordinate system defined by direction of the motion. In our approach, we only consider the projection of the displacement vector onto a two dimensional plane that corresponds to the horizontal and vertical movement of each point in reference to direction of the walk and the ground level plane. We denote the left and right ankles displacement vectors as $\boldsymbol{s}_i(t) = (h_i(t), v_i(t))^T, (i = 1, 2)$, where $h_i(t)$ and $v_i(t)$ are the displacements in horizontal and vertical directions. We will call them motion signals. Figure 2 shows the position and displacement traces of left and right ankles.



(a) (b)

Fig. 2. Traces of (a) positions and (b) positional displacements of left and right ankles as a function of time, sampled in 60 Hz. (x_1, z_1), (x_2, z_2) are the planar positions and (h_1, v_1), (h_2, v_2) are the positional displacements

We then use Discrete Fourier Transform to extract spatiotemporal features of the movement in each direction of the two points. The spectral features from $X(\omega)$ are defined as follows:

$$S_1(X(\omega)) = ||\bar{X}(\omega)||, \tag{3}$$

$$S_2(X(\omega)) = e^{j \arg(\bar{X}(\omega))}, \tag{4}$$

$$S_3(X(\omega)) = S_1(X(\omega)) \cdot S_2(X(\omega)), \tag{5}$$

where $\bar{X}(\omega)$ is the normalized Fourier coefficient:

$$\bar{X}(\omega) = \frac{1}{\int ||X(\omega)|| \mathrm{d}\omega} X(\omega). \tag{6}$$

The normalization eliminates the need for depth compensation when a motion is acquired using a video camera and the distance to the subject varies. S_1 is the magnitude spectrum, S_2 is the phase spectrum and S_3 is known as phase-weighted magnitude spectrum [6]. The magnitude distributions of motion signals are only dominant in the

lower frequency components and thus the elements of higher frequency components in phase spectra are trivial. In $S3$, only the phases of frequency components with relatively large magnitude will remain.

Phase spectra are not time shift invariant. Therefore, a uniform start point of each motion sequence is required to create a valid phase spectra. However, instead of aligning the start point of the motion signal (or the partial data block as explained later),we shift the phases in S_2 so that the phases of the fundamental frequency components are equal. In other words, the phase spectra are defined relative to phase of the fundamental frequency component.

In addition, for two signals that relate to the same motion, we can define a spectrum that is a function of the dynamics of both signals. Consider $X(\omega)$ and $Y(\omega)$ to be the Fourier coefficients of synchronized signals that belong to a motion. We propose the use of the geometrical mean of $X(\omega)$ and $Y(\omega)$ to extract a more compact spectral feature for gait recognition. Using geometrical mean based spectral features offers another simplification in that we do not need to have strict correspondence between signals and their origins. This means, for example, that we do not have to differentiate between signals from the left and right ankles. Further, we can expect the geometrical mean to be more stable against noise at the originating elements.

We defined the following spectral features to be extracted from two different motion signals:

$$T_1(X(\omega), Y(\omega)) = \sqrt{||\bar{X}(\omega)||\,||\bar{Y}(\omega)||}, \tag{7}$$

$$T_2(X(\omega), Y(\omega)) = (e^{j\arg(\bar{X}(\omega)\bar{Y}(\omega))})^{1/2}, \tag{8}$$

$$T_3(X(\omega), Y(\omega)) = T_1(X(\omega), Y(\omega)) \cdot T_2(X(\omega), Y(\omega)), \tag{9}$$

each of which corresponds to S_1, S_2 and S_3 of the geometrical mean of $X(\omega)$ and $Y(\omega)$.

In the spectral feature extraction, first we divide a motion signal into overlapping blocks of a certain length (in our experiments, we set the length to 64 with 8 sampling intervals in between). After shifting the average of signal at each block so that it averages to zero, we apply Hamming window to each data block in order to reduce DFT leakage. The final spectral feature is the average of the spectral features from the data blocks. Given that our data were taken at a sampling rate of 60 Hz, the fundamental frequency and also the frequency resolution of the spectra produced is 60/64 $\approx$ 0.94 Hz, which is consistent to the gait cycle of approximately 1 Hz as reported in [13].

4.3 Foot Motion Spectral Features

We can consider each vertical and horizontal component of any ankles as an independent entity, and use the spectral feature as a gait signature. Let the Fourier coefficients of the motion signals be $F = \{ V_1(\omega), V_2(\omega), H_1(\omega), H_2(\omega) \}$. For each non-empty element of the power set of F, we can apply any of (3), (4) or (5) to its elements and use the concatenation of the results as the spectral feature. By concatenating only spectral features of the same type, we can evaluate the characteristics of magnitude or phase spectra as well as the contribution of each motion signals in discriminating individuals.

Spectra from the average of Fourier coefficients can be used as more compact spectral features of a gait. Given the Fourier coefficients of motion signal pairs $G = \{ \{ X(\omega), Y(\omega) \} \mid X(\omega), Y(\omega) \in F, X(\omega) \neq Y(\omega) \}$, we can apply the spectra of (7), (8) or (9) to each element of non-empty elements of the super set of G, and concatenate the results as spectral features.

5 Experimental Results

The data set of our experiments comprises of 9 individuals. We group the data from the four separate sessions into 2 data sets, so that data taken at the same month are in the same data set. Each data set has about 20 gait samples per subject. We evaluate the spectral features based on the correct classification rates (CCRs) at *same-day* and *cross-month* test. In same-day test, we use leave-one-out validation method on each data set separately and average the CCRs. For the cross-month test, we alternately use one data set as training data set and the other as the test set. The purpose of this test is to evaluate the effect of time covariates in gait recognition, especially in its effects on each spectral features. The results shown are the average CCRs from alternating data and training data sets.

Individuals are recognized using a k-NN classifier (k=5 in our experiments), and Euclidean distance is used for comparing similarity of features. We disregard the DC component of Fourier decomposition, which corresponds to the average speed of movements, and truncated the spectra into 8 or 16 lower harmonics. The 16 lower harmonics represent the spectra of frequency components up to 15 Hz. We do not normalize the spectra with regard to the period of the gait cycle, but leave the variation of the period as a characteristic of an individual.

5.1 Same-Day Gait Recognition

Figure 3 shows the CCRs of the basic spectral features and Fig. 4 shows the CCRs of the more compact spectral features based on geometrical mean at same-day test.

The chance recognition rate for 9 individuals is about 11%, and the experiments show recognition rates up to 95%. Some characteristics that we observed here are:

- In general, the magnitude spectra show better discriminatory capability than phase or phase-weighted magnitude spectra, while phase-weighted magnitude spectra do not differ much from the phase spectra. The magnitudes of high order harmonics are usually exponentially smaller relative to the magnitude of the peak harmonic. Since we are using a k-NN classifier with Euclidean distance, a significant difference in the magnitudes of a higher order frequency component of two spectral features may result in only a slight difference of distance between the features. Our preliminary experiments show that using the spectra in logarithmic scale improves the recognition results.
- Although the motions of left and right feet independently are more likely to be identical for general population, using the motion signals from both feet is better, implying that the difference of the motion of left and right feet is a good cue for gait recognition.

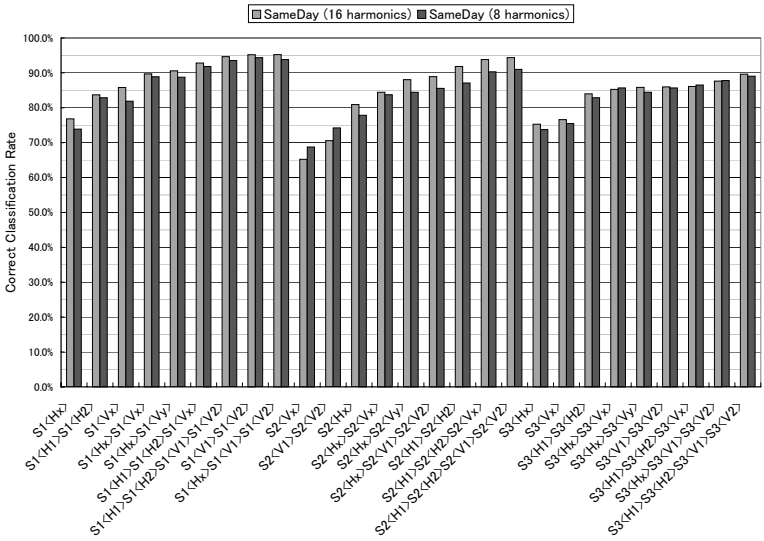


Fig. 3. CCRs for same-day test, using concatenation of basic spectral features. The CCR of `S1<Hx>S1<Vx>` is the average of the CCRs of `S1<H1>S1<V1>` and `S1<H2>S1<V2>`. Similarly `S1<Hx>S1<Vy>` represents `S1<H1>S1<V2>` and `S1<H2>S1<V1>`, which are the concatenation of magnitude spectra of different directional movement of different foot

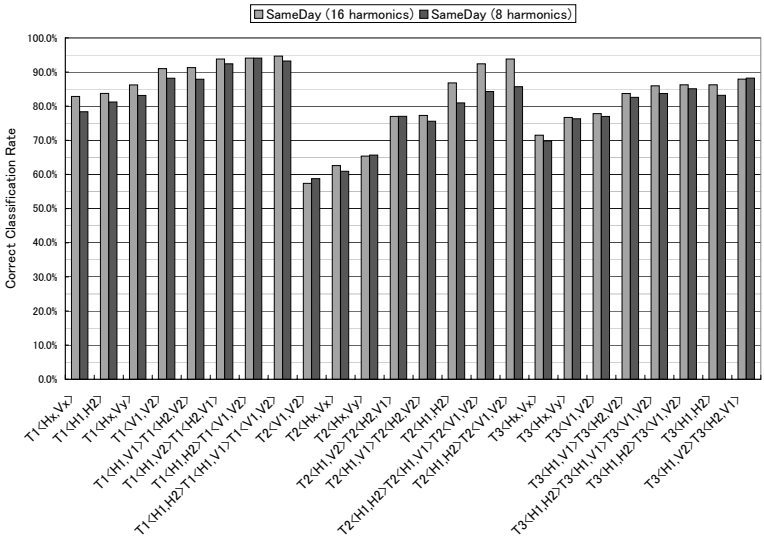


Fig. 4. CCRs for same-day test, using concatenation of geometrical mean based spectral features

– The magnitude spectra of vertical movement have better discriminatory capability than those of horizontal movement. On the other hand, the phase related spectra of horizontal movement perform better than the corresponding spectra of vertical movement.

– The compact representation of spectral features, as shown in Fig. 4, performs reasonably good, especially for the magnitude spectra. For example `T1<H1,H2>` `T1<V1,V2>` gives a CCR almost similar to `S1<H1>S<H2>S1<V1>S1<V2>`.
– The CCRs when using 16 lower harmonics are generally above the CCRs of spectral features truncated to 8 lower harmonics. As more higher harmonics are incorporated into the spectral features, the identification rates tend to increase, but not significantly.

5.2 Cross-Month Gait Recognition

Figure 5 and Fig. 6 show the CCRs at cross-month test. The discriminatory capability characteristics of the types of spectral features as well as the horizontal and vertical movement of the feet resemble that of Fig. 3 and Fig. 4.

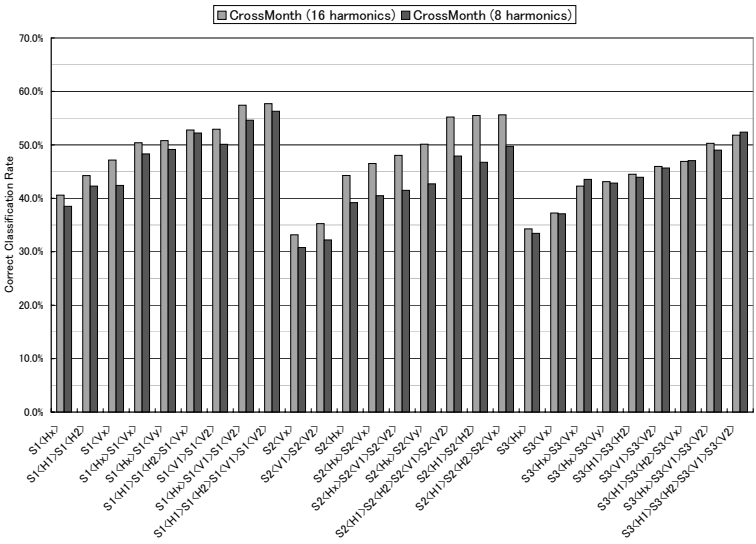


Fig. 5. CCRs for cross-month test, using concatenation of bacic spectral features

The best identification rates obtained for the cross-month test are about 60%, relatively low compared to the test within the same data set. This drop of performance reflects a change of walking pattern over time, but further analysis is needed to understand the factors underlying this change, as well as how to extract the features of gait specific to individuals that are persistent over time. Also we noticed that expressing the magnitude in logarithmic scale improves the recognition rates.

6 Conclusions

The spectral features derived from horizontal and vertical motions of ankles during a normal walk are effective gait signatures for personal identification. In experiments on a database of 9 subjects, we achieved recognition rate up to 95% in same-day test, while

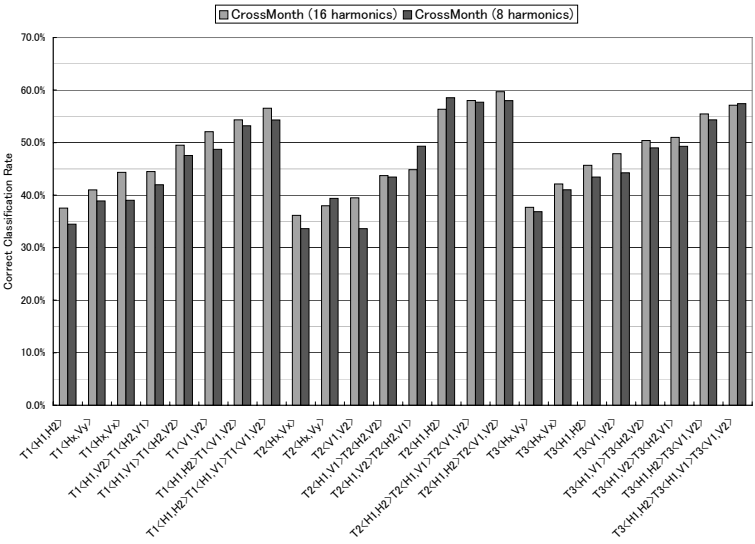


Fig. 6. CCRs for cross-month test, using concatenation of geometrical mean based spectral features

the best identification rates for cross-month test is about 60%. Furthermore, both the magnitude and phase spectra are shown to be good candidates for gait signatures.

In the proposed method, each motion signal is analyzed *independently* and direct information about interrelationships between motion signals are not represented in the gait signatures. Therefore, the approach do not consider the configuration of left and right feet during gait, but only the dynamics of both feet seen separately. Information about the structure(shape) of gait can further be added to gait signatures to obtain a better result.

Our model-based approach that utilizes foot motion is not sensitive to changes in clothing styles. However, the effects of other covariates, such as footwear and walking surface, as well as the performance for a larger data set need further investigation.

We have also evaluated the use of spectra from the geometrical mean of spectra of two motion signals for gait recognition. This approach offers reduced dimensionality of features, correspondence-free gait signature, and gives better or equal recognition rate compared to concatenation of spectra.

Spectral features from motion signals can be extended into taking more feature points, or using other motion signals accessible in data acquisition. In the implementation, gaits can be measured using video camera, laser range finder, or other alternative sensing technologies. Works on sensing the trajectory of foot motion from a video input are also necessary for a practical gait recognition system.

References

1. G. Johansson, Visual Motion Perception, *Scientific American*, pp. 75-80, 85-88, June 1975.
2. J. Cutting and L. Kozlowski, Recognizing Friends by Their Walk: Gait perception Without Familiarity Cues, *Bulletin Psychonomic Society*, vol. 9, no. 5, pp. 353-356, 1977.

3. S.A. Niyogi and E.H. Adelson, Analyzing and recognizing walking figures in XYT, *Proc. Conf. Computer vision and Pattern Recognition 1994*, pp. 467-474, 1994.
4. H.M. Lakany and G.M. Hayes, An Algorithm for Recognising Walkers, *Proc. of the 1st Int'l Conference on Audio- and Video-based Person Authentication*, March 1997.
5. H. Murase and R. Sakai, Moving Object Recognition in Eigenspace Representation: Gait Analysis and Lip Reading, *Pattern Recognition Letters*, vol. 17 issue 2, pp. 155-162, 1997.
6. D. Cunado, M.S. Nixon and J.N. Carter, Using Gait as a Biometric, via Phase-Weighted Magnitude Spectra, *Proceedings of 1st Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pp. 95-102, March 1997.
7. Q. He and C. Debruner, Individual Recognition from Periodic Activity Using Hidden Markov Models, *Proc. IEEE Workshop on Human Motion*, 2000.
8. C. BenAbdelkader, R. Cutler and L. Davis, Motion-based Recognition of People in Eigen-Gait Space, *Proc. of the 5th IEEE Int'l Conference on Automatic Face and Gesture Recognition*, pp. 267-272, May 2002.
9. L. Lee and W.E.L. Grimson, Gait Appearance for Recognition, *ECCV Workshop on Biometric Authentication*, June 2002.
10. L. Wang, T. Tan, H. Ning and W. Hu, Silhouette Analysis-Based Gait Recognition for Human Identification, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505-1518, December 2003.
11. S.D. Mowbray and M.S. Nixon, Automatic Gait Recognition via Fourier Descriptors of Deformable Objects, *Proceedings of Audio Visual Biometric Person Authentication*, pp. 566-573, June 2003.
12. C. Yam, M.S. Nixon and J.N. Carter, Automated Person Recognition by Walking and Running via Model-Based Approaches, *Pattern Recognition*, vol. 37 issue 5, pp. 1057-1072, May 2004.
13. B. Li and H. Holstein, Perception of Human Periodic Motion in Moving Light Displays - a Motion-Based Frequency Domain Approach, *Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behaviour (AISBJ)*, vol.1, no. 5, pp. 403-416, 2004.
14. S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother and K.W. Bowyer, The HumanID Gait Challenge Problem: Data Sets, Performance and Analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162-177, February 2005.

VALID: A New Practical Audio-Visual Database, and Comparative Results

Niall A. Fox, Brian A. O'Mullane, and Richard B. Reilly

Dept. of Electronic and Electrical Engineering,
University College Dublin, Belfield, Dublin 4, Ireland
{niall.fox,brian.omullane}@ee.ucd.ie, richard.reilly@ucd.ie
http://wwwdsp.ucd.ie/

Abstract. The performance of deployed audio, face, and multi-modal person recognition systems in non-controlled scenarios, is typically lower than systems developed in highly controlled environments. With the aim to facilitate the development of robust audio, face, and multi-modal person recognition systems, the new large and realistic multi-modal (audio-visual) VALID database was acquired in a noisy "real world" office scenario with no control on illumination or acoustic noise. In this paper we describe the acquisition and content of the VALID database, consisting of five recording sessions of 106 subjects over a period of one month. Speaker identification experiments using visual speech features extracted from the mouth region are reported. The performance based on the uncontrolled VALID database is compared with that of the controlled XM2VTS database. The best VALID and XM2VTS based accuracies are 63.21% and 97.17% respectively. This highlights the degrading effect of an uncontrolled illumination environment and the importance of this database for deploying real world applications. The VALID database is available to the academic community through *http://ee.ucd.ie/validdb/*.

1 Introduction

With biometrics applications becoming more widely accepted and commonplace in high security situations, such as airports and banks, it is not hard to imagine a time in the near future where they will become ubiquitous in every day life. Less critical applications such as online identification (ID) for low cost purchases, for example video rental, will emerge; with the emphasis focusing on convenience and cost rather than false acceptances. Traditional problems with background noise and random lighting will become increasingly critical, because, operational conditions will not be as controlled as in high security applications. To overcome some of these problems, the use of multi-modal biometric systems is becoming more widespread. However, there is a shortage of large multi-modal databases in the research community that can address these critical real world conditions. A brief overview of currently available audio-visual databases is provided, in order to put the VALID database in context.

The M2VTS audio-visual database [1] was recorded when audio-visual speech processing was in its infancy. It consists of recordings of 37 subjects counting ten digits in French, with five recording sessions per subject, spaced weekly. Only one sentence was recorded per session. The recordings conditions were "ideal" with controlled lighting. The extended XM2VTS database [2] addressed some of the limitations of the M2VTS database, in terms of the number of subjects and sentences. It

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 777–786, 2005.
© Springer-Verlag Berlin Heidelberg 2005

consists of 295 subjects and the recording of three sentences. The main problem encountered with the XM2VTS database was the extremely well controlled visual recording conditions and the blue screen background. This does not represent a practical real world scenario. This fact is highlighted by the high visual speech based speaker ID accuracy of 86% reported in previous work by the authors [3]. The recent BANCA database [4] is a large audio-visual database consisting of 208 subjects, recorded under controlled, degraded and adverse scenarios. The 208 subjects are split equally into four different languages groups. This will provide data of a more challenging and practical nature for the testing of audio-visual fusion methodologies. However, the audio-visual recordings of BANCA are text independent (TI), whereas those of XM2VTS are text dependent (TD), meaning that the same text is spoken across all subjects and sessions (see Section 3.2), hence BANCA is more suited to the fusion of the face and speech modalities for person recognition. Also recent, is the BIOMET multi-modal database [5], comprising of five modalities, namely: audio, face, handscan, fingerprint, and signature. Three recording sessions took place with a total participation of 91 subjects. The BIOMET database also consists of video shots of spoken digits, 12 phonetically balanced sentences, and other phrases, recorded in a controlled environment. Head rotation and head profile shots were also included.

The VALID database is a large audio-visual database, consisting of five recording sessions of 106 subjects over a period of one month. The database is designed to be realistic and challenging with four of the sessions recorded in noisy "real world" office scenarios (where computer use is likely) with no control on illumination or background acoustic noise and incorporates variation in accent, skin tone, facial hair, and visual background. The VALID database consists of TD recordings and is suitable for the development of robust TD audio-visual speaker recognition systems.

This paper is organised as follows. Section 2 describes the VALID content and recording conditions. Speaker ID experiments, using mouth region speech features, are carried out. Feature extraction is discussed in Section 3. In Section 4 the experiments are described, results are presented and the relative performances of XM2VTS and VALID are discussed; followed by some conclusions in Section 5.

2 The VALID Database

2.1 The Database Acquisition Hardware

The entire database was recorded using a Canon 3CCD XM1 PAL digital video camcorder, with a sensor resolution of 320k pixels and records in the PAL DV format. The video was captured using a color sampling resolution of 4:2:0 with the audio captured using 16 bit stereo samples at a frequency of 32kHz with PCM encoding. The video frame rate is 25 fps with a pixel resolution of 576 x 720 (rows x columns) and 24bit pixel depth. The DV PAL format employs intraframe lossy compression at a fixed ratio of 5:1. The audio data is not compressed.

2.2 The Database Content

The content of the VALID database was designed to supplement that of the XM2VTS database. Three utterances were recorded per session (in English), namely:

1: "<Subject's full name>"
2: "5 0 6 9 2 8 1 3 7 4"
3: "Joe took father's green shoe bench out"

Utterance 1 was used for subject identify, and will not be publicized. Utterances 2 and 3 are the same as those in XM2VTS. These were chosen so that algorithms tested on XM2VTS could be tested on VALID; and hence the effects of a more practical visual environment could be examined. Utterance 3 is phonetically balanced (i.e. the sentence contains approximately the same number of voiced, unvoiced and plosive sounds). Utterances 2 and 3 also have the advantage of unfamiliarity to subjects; thus, ensuring a slow and intelligible rate of speech. In order to facilitate the training and testing of face recognition systems that are robust to head pose, a head rotation sequence was recorded during Session 1, where the subject was asked to face four targets, placed approximately 15 degrees above, below, left, and right of the camera, resulting in images shots that were slightly off-frontal. The natural environment of Sessions 2-5 also gives varied pose. The database consists of 106 subjects (77 male, 29 female) and has an ethnic composition of 97 Europeans and 9 Asians, comprising of undergraduate/postgraduate students and staff from UCD. Unfortunately, due to the male bias in Engineering, VALID has a high male content.

2.3 Recording Conditions

The five sessions were recorded over a period of one month, allowing for variation in the voice, clothing, facial hair, hairstyle, and cosmetic appearance of the subjects and also variation of the visual background, illumination, and acoustic noise. The first session was recorded under controlled acoustic/illumination conditions, which we refer to as the *controlled environment*. The subject was seated for the recordings with a fixed camera-seat distance and fixed camera height. The illumination was diffuse and constant. For the controlled session, a blue background screen was employed, which will facilitate the task of skin detection. The database is designed to be realistic and challenging, hence the other four sessions were recorded in noisy real world scenarios with no control on illumination or acoustic noise. This is referred to as the *uncontrolled environment*. Acoustic noise/interference includes computer fan noise and speech/movement of third parties. The uncontrolled sessions were recorded indoors, predominantly in offices, but also in more open spaces. The illumination consisted of office and natural lighting. The camera-subject distance was not fixed.

The subjects read each utterance from a paper sheet attached to the camera, with no instruction on speech rate. Due to the difficulty in voicing utterances 2 and 3, many subjects misread them or lost their composure. Consequently, for each session, the utterances were repeated once, with a view to choosing the most suitable utterance in each case for the final database. Fig. 1 shows the variability of the appearance of the subjects across the five recording sessions. The illumination for Session 1 does not vary significantly for the three subjects shown, however, it varies for Sessions 2-5. Session 1 exhibits uniform illumination of the face and limited contrast between the face and background. Sessions 2-5 exhibit non-uniform face illumination, varying background, and in some instances, high contrast.

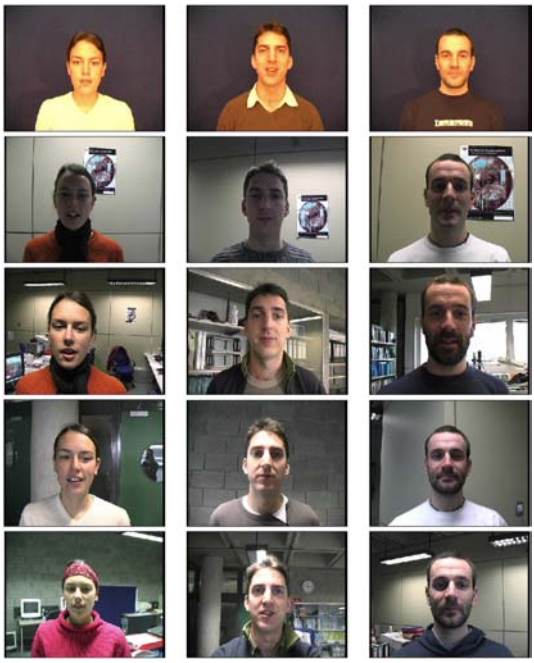


Fig. 1. Three VALID subjects with still images taken from each of the five sessions

3 Visual Speaker Identification

The described VALID database can be employed for audio, face, audio-face, and audio-visual speaker recognition experiments, where, audio-visual refers to the fusion of acoustic and visual speech. Visual speech based speaker recognition differs from face recognition in two major ways. Firstly, face recognition employs the entire face area whereas visual speech speaker recognition employs a region of interest about the speaker's mouth, where most of the speech information is contained. Secondly, for face recognition, a gallery of static face images forms a template, whereas for visual speech speaker recognition, it is attempted to model the visual speech temporal characteristics. The visual speech signal is rarely used as a complete recognition system; rather it is integrated with the acoustic speech signal. Previous work by the authors has shown that audio-visual speaker ID boosts performance, particularly in the presence of acoustic noise [6], [7]. Visual speech based speaker ID is closely related to speechreading (visual speech recognition). Recently, several state of the art reviews have been carried out on audio-visual speech processing [8], [9].

It has been shown in [6], and elsewhere, that late integration is a good strategy to combine audio and visual speech data for speaker ID. This can be carried out for a given test utterance by attaining separate audio and visual scores and combining them using appropriately chosen weights. Indeed, several studies have reported visual speech based speaker recognition results separately [10], [11]; with surprisingly high accuracies. In [10], an accuracy of 97.9% was reported based on shape and intensity features, however, on a subject set of just 12 speakers. In [11] a lip based HMM at-

tained a speaker ID accuracy of 84% on the larger M2VTS database. It is argued that these examples of high performance were based on unrealistic scenarios (controlled recordings), and that in a more realistic environment the performance will be significantly lower.

The foremost distinction between the VALID database and XM2VTS database is the difference in illumination and background. To test the effect of this, speaker ID experiments using mouth region visual speech features were carried out. The VALID database performance is compared with the more controlled XM2VTS database.

3.1 Visual Speech Features

It has been consistently shown in several speechreading studies, that pixel based features outperform geometric features [12], [14]. In [12], it is reported that three pixel based methods outperformed an active appearance model based approach. Geometric features/lip-contours require significantly more sophisticated mouth-tracking techniques compared to just locating the mouth ROI for pixel-based features. This may be difficult, when the illumination conditions are poor. Pixel based features employ linear transforms to map the image ROI into a lower dimensional space, removing the redundant information while retaining the salient speech features. Many types of transforms are examined in the literature, including the *discrete cosine transform* (DCT) [3] [13] [14], *discrete wavelet transform* (DWT) [14], *Hadamard transform* [15], and *principal component analysis* (PCA) [12].

One of the most commonly employed image transforms is the DCT. It is related to the discrete Fourier transform (DFT) and is amenable to fast implementation when the image dimensions are powers of two. It also has good de-correlation and energy compaction properties [16]. The Hadamard transform matrix contains only +1's and −1's, hence it can be implemented "cheaply" using just additions and subtractions. However, it still has good energy compaction performance, but not as good as the DCT [16]. PCA has the disadvantage of requiring an intensive training stage. The DWT has been shown to slightly under perform the DCT for speechreading [12]. In the current study, the DCT and Hadamard features are tested and their relative performance for the task of visual speech based speaker ID is compared under controlled and uncontrolled audio-visual recording conditions. In the literature, various methods have been employed to extract the most important transform coefficients. One popular method consists of applying a mask to the transform coefficient matrix [14]. This mask usually selects the coefficients in a tri-angular fashion (upper-left region of the transform matrix), i.e. in the case of the DCT the high-energy low spatial frequencies. The use of a similar mask is not suitable for the Hadamard transform, because the most informative coefficients are not contained in a local region, as for the 2D-DCT. An alternative method to choosing the features is to select the top L features with the highest energy, as calculated over the complete training set [12]. In [13], for speechreading, selecting features with the *highest energy* was found to outperform two other criteria, namely *highest variance* and *highest variance of mean normalised features*. It was found that by employing either a mask or the *L highest energy* features, a similar set of DCT features were selected.

For the experiments described here, the L highest energy coefficients of the given transform of each video frame ROI are selected. These are referred to as the *static*

visual speech features. For visual speech recognition, it is customary to generate *delta* (first order frame derivatives) and *acceleration* (second order frame derivatives) features also. These were calculated using the available HTK functions, with five adjacent frames employed for both the *delta* and *acceleration* features [17]. The three types of visual features are concatenated to form a *3*L* dimensional feature vector. This is shown in Equation 1 where S, D, A refer to *static*, *delta* and *acceleration* features respectively and o_n refers to the observation feature vector of the n^{th} visual frame; T is the number of frames in the visual speech observation sequence.

$$o_n^{\{S-D-A\}} = [o_n^{\{S\}}, o_n^{\{D\}}, o_n^{\{A\}}],\ 1 \le n \le T. \tag{1}$$

Hence, an entire visual speech utterance will result in an observation sequence, O, of feature vectors denoted by $O=\{o_1, o_2, \ldots, o_n, \ldots, o_T\}$.

3.2 Visual Speech Identification Model

Acoustic speaker modelling techniques are either TD or TI. For TD modelling, the same utterance is used for both training and testing, whereas for TI modelling, the test and training utterances may be different. TD systems require less training data than TI systems, and hence are more suited to audio-visual speech processing, where training data may be scarce. TD modelling has been found to outperform TI modelling [18]. The two databases examined here consist of repetitions of the same speech utterances; for this reason, a TD hidden Markov model (HMM) is used to model the temporal information of the visual speech sequence, i.e. a single HMM models the entire sentence. Three of the available four sessions are used to train the speaker dependent HMMs and the fourth session is used for testing. Left-to-right HMMs with diagonal covariance matrices are trained using HTK. It is difficult to adequately train a HMM using only three observation sequences. Thus, a background HMM is first trained on all speakers and then used to initialise the speaker dependent HMMs parameters.

4 Experiments, Results and Discussion

Visual speaker ID experiments are carried out on both the XM2VTS and the VALID databases. The third recording utterance, *"Joe took father's green shoe bench out"*, from each database is modelled using TD HMMs. 106 subjects from a possible 295 XM2VTS subjects are tested and compared with the 106 VALID subjects. The four uncontrolled VALID sessions are employed. This amounted to a total of 45,523 XM2VTS video frames compared with 36,044 VALID frames. This discrepancy in frame number indicates that the VALID subjects, on average, spoke the third sentence 20% faster than the XM2VTS subjects. This is because the VALID subjects were not instructed to speak slowly, and perhaps, also due to the informal recording settings. The four sessions are cross-validated by choosing three sessions for training and the remaining session for testing; thus, quadrupling the amount of tests that can be carried to *4*106*. As the same utterance and number of subjects are being used, the authors believe that the current study on both the VALID and the XM2VTS databases, offers a fair comparison of a controlled with an uncontrolled environment.

The ROI employed consists of a square window about the centre of the speaker's mouth. The mouth centre is determined manually for every 10^{th} image frame and

interpolated for the other frames. The speaker-camera distance for the XM2VTS recordings is reasonably constant and hence scaling of the face is unnecessary (see Fig. 2b). The speaker-camera distance for the VALID varies significantly, hence, face scale normalization is required. The speaker's inter eye center pixel distance is scaled for each video frame so that it is constant. The extracted 49 x 49 pixel ROI is down sampled to 32 x 32 pixels in order to reduce the data dimensions involved and to allow efficient implementation of the image transforms. To compensate for varying illumination, the gray scale of ROI is histogram equalised followed by mean pixel value subtraction. The image transforms are then applied to the pre-processed ROI. Fig. 2 compares the ROI for three subjects of the VALID with three subjects from XM2VTS. The variation in illumination/scale across the four VALID sessions is evident, whereas XM2VTS exhibits a high level of illumination/scale consistency across the four sessions; this does not represent a real world scenario.

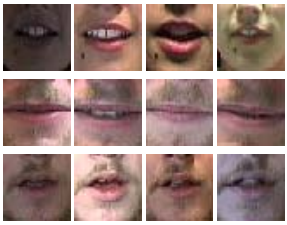


Fig. 2a. Non-scaled ROI frames for three subjects (rows 1 to 3) across the last four sessions of the VALID database

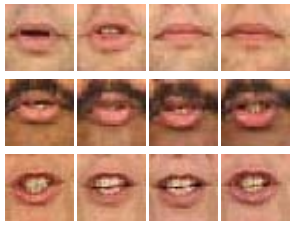


Fig. 2b. Non-scaled ROI frames for three subjects (rows 1 to 3) across the last four sessions of the XM2VTS database

It was found that the best performance was achieved by using just one HMM state and one Gaussian mixture. The speaker ID experimental results for the *static* (S) and *static-delta-acceleration* (SDA) features are presented in Table 1. The SDA feature results are also displayed in Fig. 3. The best XM2VTS accuracy of 97.17% is significantly higher than the best VALID accuracy of 63.21%. This disparity in accuracy across the two databases may seem reasonable when Fig. 2 is considered. The VALID accuracy is significantly lower than the accuracies of 97.9% and 84% reported in [10] and [11] respectively (based on controlled video data, as discussed above). This highlights the effect on performance due to the uncontrolled video recording conditions. From Fig. 3, it can be seen, that the performance rises faster with respect to feature dimension (the L highest static features) for the clean data (XM2VTS) compared to the noisy data (VALID). It should be noted that higher feature dimensions result in increased computational time and that a L value of 100, resulting in a SDA dimension of 300, would be computationally too expensive to implement practically. As expected, the DCT features outperform the Hadamard features for all of the tests. However, the DCT features give only a 1.2% relative performance increase for the XM2VTS SDA feature test with $L=100$; and a larger 5.9% increase for the corresponding VALID test. This suggests that for clean data the DCT features do not significantly outperform the Hadamard features, yet, for nosier data, the DCT features are more important. The use of delta/acceleration features does not yield a significant improvement over the static features alone (XM2VTS: SDA 97.17% versus S 96.46% and for VALID: SDA 63.21% versus S 62.26%).

Table 1. XM2VTS/VALID ID accuracies. The feature dimension is the *L* value used (*L* most energetic features); so the actual SDA dimension is *3*L. HAD* represents Hadamard

XM2VTS Database						
Dimension [S]	10	20	40	60	80	100
DCT Accuracy [S] (%)	69.58	88.21	93.87	94.58	95.99	96.46
HAD Accuracy [S] (%)	69.34	80.42	89.86	92.69	93.63	93.87
DCT Accuracy [S-D-A] (%)	79.25	92.22	96.70	96.46	97.41	97.17
HAD Accuracy [S-D-A] (%)	76.18	86.56	93.40	95.52	94.81	95.99
VALID Database						
Dimension [S]	10	20	40	60	80	100
DCT Accuracy [S] (%)	40.80	50.47	56.37	57.55	60.38	62.26
HAD Accuracy [S] (%)	32.31	45.99	52.59	54.01	57.31	57.55
DCT Accuracy [S-D-A] (%)	42.45	54.72	57.55	59.20	60.38	63.21
HAD Accuracy [S-D-A] (%)	38.44	50.47	55.66	57.08	57.78	59.67

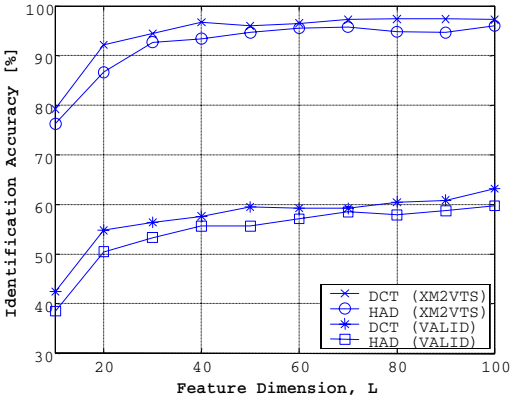


Fig. 3. XM2VTS/VALID ID accuracies based on SDA visual speech features

5 Conclusion and Summary

A new large multi-modal (audio-visual) database has been created and made available to the academic community. Results on this database have been presented. It is hoped that this database will facilitate the testing and comparison of multi-modal algorithms, across the research community, on real world audio-visual data.

The purpose of the study outlined was to compare visual speech speaker ID on controlled audio-visual data with uncontrolled data. High ID accuracies (97.17%) were achieved on the controlled XM2VTS data, however, when the same methodologies were tested on the uncontrolled VALID data using an equivalent sized test set, the performance was significantly poorer (63.21%). This highlights the challenge posed by poor illumination for visual speaker recognition. Manual mouth tracking is employed, so a further drop in performance is expected for automatic tracking. The DCT and Hadamard image transforms are examined. As expected, the DCT performs best, however, despite been computationally less expensive, the Hadamard performance is not significantly lower. Despite the poor VALID performance, it is still expected that for noisy acoustic data, the fusion of VALID mouth ROI information with the acoustic modality should still yield improvements.

In the current study, the visual speech ROI sequence is modeled using HMMs. It was found that when modeling the *static* features, a one state HMM performed best,

which is essentially a Gaussian Mixture Model. It has also been shown that the *static* features are more important for speaker ID than the dynamic *delta/acceleration* features, because, the concatenation of the dynamic features to the *static* features yields only a marginal performance increase. The importance of dynamic features for visual speech recognition, and the importance of static features for visual speaker recognition was shown in [19]. The high performance of a single state HMM and static features indicate that HMMs are not suitable for modeling visual speech, particularly for speaker ID. The temporal information (i.e. the change of the mouth ROI over time) is not modeled. A one state HMM is effectively an average of the mouth ROI over the duration of speech sequence. This can be considered as another feature for face recognition, and indeed, studies have shown that mouth features can be integrated with face recognition systems to yield higher performance [3].

The audio, face, video components, eye/mouth coordinates, and the experimental protocol of the VALID database are available to the academic community through *http://ee.ucd.ie/validdb/*.

Acknowledgements

This work is part of the VALID biometric project, funded by Enterprise Ireland. Andrea Osl and Rosalyn Moran are thanked for the collection/preparation of the VALID database. Finally, a special thank you is given to the database participants.

References

1. S. Pigeon and L. Vandendorpe, "The M2VTS Multimodal Face Database (Release 1.00)," Proc. First International Conf. on Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, pp. 403-409 1997.
2. "The XM2VTS database; http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/."
3. N. A. Fox, R. Gross, P. de Chazal, J. F. Cohn, and R. B. Reilly, "Person Identification Using Automatic Integration of Speech, Lip, and Face Experts," Proceedings of the 2003 ACM SIGMM workshop on Biometrics Methods and Applications, Berkley, California, pp. 25-32, Nov. 2003.
4. E. B. Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J. P. Thiran, "The BANCA Database and Evaluation Protocol," Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Guildford, UK, pp. 625-638, June 2003.
5. S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. L. l. Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacretaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", 2688 ed, 2003.
6. N. A. Fox and R. B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," Proc. of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK, pp. 743-751, June 2003.
7. N. A. Fox and R. Reilly, "Robust Multi-modal Person Identification with Tolerance of Facial Expression," Systems, Man and Cybernetics, 2004 IEEE International Conference on, The Hague, The Netherlands, vol. 1, pp. 580-585, 10-13 Oct 2004.
8. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," Proceedings of the IEEE, vol. 91, pp. 1306-1324, Sept. 2003.

9. C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," IEEE Transactions on Multimedia, vol. 4, pp. 23-35, Mar 2002.
10. J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker Identification by Lipreading," Proceedings of the Fourth International Conference on Spoken Language, ICSLP 96, vol. 1, pp. 62-65, Oct. 1996.
11. T. Wark, S. Sridharan, and V. Chandran, "The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMMs," Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing,. ICASSP '00., vol. 6, pp. 2389-2392 2000.
12. I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A Comparison of Model and Transform-based Visual Features for Audio-Visual LVCSR," IEEE International Conference on Multimedia and Expo, 2001., pp. 825-828 2001.
13. M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-Based Video Features for Audio-visual Speech Recognition.," Proceedings of the 7th ICSLP, Denver, Colorado (USA), vol. 3, pp. 1925-1928 2002.
14. G. Potamianos, H. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," Proceedings of the IEEE International Conference on Image Processing, ICIP 98, Chicago, vol. 3, pp. 173-177, Oct. 1998.
15. P. Scanlon and R. B. Reilly, "Feature Analysis for Automatic Speechreading," Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing, pp. 625-630, Oct. 2001.
16. S. Theodoridis and K. Koutroumbas, "Pattern Recognition": Academic Press, 1999.
17. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.1)". Cambridge University Engineering Department: Microsoft Corporation, 2001.
18. J. Luettin, "Speaker verification experiments on the XM2VTS database," in IDIAP Communication 98-02: IDIAP, Martigny, Switzerland, 1999.
19. S. Lucey, "An Evaluation of Visual Speech Features for the Tasks of Speech and Speaker Recognition," Proc. of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK, pp. 260 – 267, June 2003.

Audio-Visual Speaker Identification via Adaptive Fusion Using Reliability Estimates of Both Modalities

Niall A. Fox, Brian A. O'Mullane, and Richard B. Reilly

Dept. of Electronic and Electrical Engineering,
University College Dublin, Belfield, Dublin 4, Ireland
{niall.fox,brian.omullane}@ee.ucd.ie, richard.reilly@ucd.ie
http://wwwdsp.ucd.ie/

Abstract. An audio-visual speaker identification system is described, where the audio and visual speech modalities are fused by an automatic unsupervised process that adapts to local classifier performance, by taking into account the output score based reliability estimates of both modalities. Previously reported methods do not consider that both the audio and the visual modalities can be degraded. The visual modality uses the speakers lip information. To test the robustness of the system, the audio and visual modalities are degraded to emulate various levels of train/test mismatch; employing additive white Gaussian noise for the audio and JPEG compression for the visual signals. Experiments are carried out on a large augmented data set from the XM2VTS database. The results show improved audio-visual accuracies at all tested levels of audio and visual degradation, compared to the individual audio or visual modality accuracies. For high mismatch levels, the audio, visual, and auto-adapted audio-visual accuracies are 37.1%, 48%, and 71.4% respectively.

1 Introduction

Biometrics is a field of technology devoted to verification or identification of individuals using biological or behavioral traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multi class problem, involves identifying a user from a set of subjects. Speaker identification becomes more difficult as the number of registered subjects increases. Acoustic based speaker identification performs well when the audio signal to noise ratio (SNR) is high. However, the performance degrades quickly as the test SNR decreases [1], which is referred to as an audio train/test mismatch.

The area of audio-visual signal processing has received much attention over the past ten years. Recent state of the art reviews indicate that much of the research carried out focuses on speech recognition [2], [3]. Audio-visual speech processing can also be applied to speaker recognition. However, the same issues remain, such as how to account for the reliability of the modalities and at what level to carry out the fusion. The benefits of audio-visual fusion for the purpose of speaker identification have been shown in [1]. The fusion method employed, used modality weightings found by exhaustive search. Whereas this highlights the potential of audio-visual fusion, it is not useful in a practical scenario. Other audio-visual speaker identification approaches that use more automated fusion techniques [4] do not address the issue of an audio train/test mismatch. In [5], audio-visual speaker verification experiments were carried out on 36 subjects, however, only an audio train/test mismatch was tested, whereas, a visual train/test mismatch was not considered. In [6], robust audio-visual classifier

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 787–796, 2005.
© Springer-Verlag Berlin Heidelberg 2005

fusion under both audio and visual train/test mismatch conditions was described. The adaptive fusion results are encouraging, with improved audio-visual accuracies. However, the test set was small, consisting of only eight subjects. Face and speech information was combined in [7], using post-classifier fusion methods for person verification. The first method was a noise resistant piece-wise linear post-classifier. The decision boundary was constructed such that the error associated with the movement of score vectors under mismatched testing conditions was minimized. The second method was a modified Bayesian post-classifier. Both methods are suitable for person verification and identification. Visual noise was not considered.

The proposed audio-visual fusion method operates in an automatic unsupervised manner that adapts to the performance of each mode in the test environment; by taking into account the output score based reliability estimates of both modalities. We apply the proposed method to the specific problem of closed-set person identification, although the system is not restricted to this application, and can also be applied to the more general problem of open-set person recognition (identification or verification). Results are reported on a large data set of 251 subjects. In this paper, the visual modality refers to a sequence of visual mouth images. Both the audio and visual testing conditions are degraded to give a train/test mismatch. The results show that the accuracy of the visual modality is not adversely degraded by visual lossy compression. These results are important for remote authentication applications, where bandwidth is important and acoustic noise is probable, such as online authentication and video telephony.

This paper is organized as follows. In Sections 2/3 we describe how the audio and visual identification is performed. Section 4 investigates audio-visual fusion methods and describes how fusion is carried out in this study. Results are presented in Section 5. Finally in Section 6, the paper is summarized and some conclusions are drawn.

2 Audio Identification

The XM2VTS audio-visual database [8] was employed, which consists of 295 subjects and four sessions. The first recording per session of the third sentence (*"Joe took fathers green shoe bench out"*) was tested. Only 251 of the 295 subjects were suitable for the modeling methodology employed, due to truncated sentences. The video is at 25Hz and has a resolution of *720x576* which was down-sampled to *360x288* pixels. Audio speaker recognition is a mature topic [9]. We employ standard methods. The signal was divided into 20ms Hamming window frames, with 10ms overlap, giving an audio frame rate of 100Hz. Mel-frequency cepstral coefficients of dimension 16 and the frame energy were extracted from each frame, to give 17 static features. Static features refer to features extracted from the individual frames and do not depend on other frames. 17 first order derivatives or *delta* features were calculated using W_D (the delta window size) adjacent static frames. The *delta* features are appended to the static features, giving a feature vector of dimension 34. These are calculated using HTK [10], with $W_D = 5$. *Cepstral mean normalization* was also performed [10]. A text dependent speaker identification methodology was tested. For text dependent modeling [9] the same phrase is spoken by the subject for both training and testing. It was employed, as opposed to text independence, due to its suitability to the database used; also, it has been found to out-perform text independence [11].

The N subject classes S_i, $i = 1,2,\ldots,N$, are represented by N speaker hidden Markov models (HMMs) denoted by λ_i, $i=1,2,\ldots,N$, where $N = 251$ here. The first three sessions were used for training and the last session for testing. A background HMM was trained for all N subjects and captures the speech variation over the entire train set. Since there are only three training utterances per subject, there was insufficient data to directly train a speaker HMM. For this reason, the background model parameters are used to initialize the training of the speaker models. An utterance is represented by a sequence, O, of speech observation feature vectors denoted by, $O=\{o_1, o_2,\ldots, o_t,\ldots, o_T\}$, where o_t is the observation at time t and T is the number of frames. The HMM output scores are in *log-likelihood* form, denoted by $ll(O|S_i)$.

The audio HMMs were trained on the "clean" speech, which was the original XM2VTS data. Additive white Gaussian noise was applied to the clean audio at SNR levels ranging from 48dB to 21dB in decrements of 3dB. The models were tested on the various SNR levels. This provides for a mismatch between the testing and training conditions. HMM training/testing was carried out using the HTK toolkit.

3 Visual Identification

Visual speech feature analysis has also received much attention recently [12], [13]. Transform based features were used to represent the visual information based on the Discrete Cosine Transform (DCT), which was employed because of its high energy compaction [14]. The visual mouth features were extracted from the mouth region of interested (ROI), which consists of a 49×49 color pixel block. The ROI positions were determined by manual labeling every 10^{th} frame, and interpolating for the other frames. The gray scale values of the ROI are histogram equalized and the mean pixel value was subtracted. This image pre-processing is carried out to account for varying illumination conditions and was found to improve the visual performance. The DCT was applied to the pre-processed ROI. Most of the lip information is contained in the lower spatial frequencies [14]. A mask (see Fig. 1b), which selects the 14 most important coefficients in a tri-angular fashion, forms the visual feature vector.

In [13], an image transform based approach was used to carry out visual word recognition. The system demonstrated robustness to JPEG compression, with no significant drop in performance until JPEG quality factors (QF) levels below 10. For our study, in order to account for practical video conditions, the video frame images were compressed using JPEG compression. Ten levels of JPEG QF were examined; $QF\in\{50,25,18,14,10,8,6,4,3,2\}$, where a QF of 100 represents the original uncompressed images. The compression was applied to each entire video frame. The ROI was then extracted from the compressed images. The manually labeled mouth coordinates were employed, so that any drop in visual performance would be due to mismatched testing conditions only, rather than to poorer mouth tracking. Fig. 1a shows the mouth ROI w.r.t. QF. Blocking artifacts are evident at the lower QF levels.

The visual sentences are modeled using the same HMM methodology as described in Section 2. We refer to these models as the visual speaker models. The number of states employed is adjusted to achieve the best visual accuracy. The *static* features consist of the 14 DCT coefficients. The *delta* features were calculated as in Section 2. The second order frame derivatives or *acceleration* features were also calculated using W_A adjacent *delta* feature frames, where W_A is the size of the acceleration win-

dow. These are calculated using HTK, employing $W_D=W_A=5$. These three types of visual features were also concatenated to form a 42 dimensional feature vector, as shown in Equation 1, where *S*, *D*, *A* refer to the *static*, *delta* and *acceleration* features respectively and o_n refers to the n^{th} visual frame observation.

$$o_n^{\{S-D-A\}} = [o_n^{\{S\}}, o_n^{\{D\}}, o_n^{\{A\}}]. \quad (1)$$

The visual speaker HMMs were trained on the "clean" ROIs and tested on the degraded ROIs; thus providing for a visual train/test mismatch.

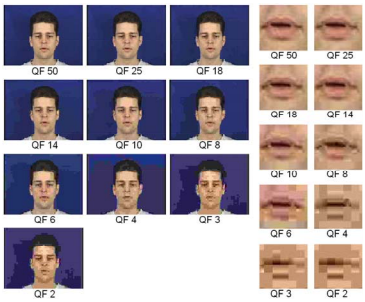


Fig. 1a. Ten levels of JPEG QF and ROIs

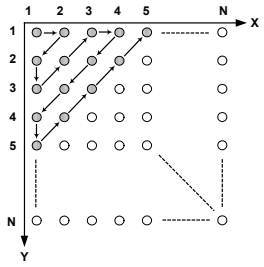


Fig. 1b. The mask for selecting the coefficients

4 Audio-Visual Fusion

The fusion of classifiers is a mature research topic, predating work on audio-visual speech fusion [15]. Due to correlations of the audio and visual modalities, fusion can be carried out at the pre-classification level. The two audio-visual fusion approaches most commonly investigated are *early* (feature fusion) and *late* (score fusion). Feature fusion, while being basic to implement via feature vector concatenation, has several disadvantages. The audio-visual feature vector has a larger dimension, and due to the "curse of dimensionality", this results in making the training of parametric models, such as HMMs, less practical. Also, feature fusion does not take the reliability of either modality into account; if one modality is very noisy, the audio-visual feature vector will be compromised and catastrophic fusion may occur; where the audio-visual accuracy is poorer than either of the single modalities; as demonstrated in [1].

An intermediate modeling approach is also possible, using multistream HMMs [16], [17]. The audio and visual speech streams can be modeled as a coupled time series, which can capture the coupling and the dependence between the two streams, however, as for feature fusion, it is difficult to take modality reliability into account.

Score fusion consists of using the audio and visual classifier outputs to provide an audio-visual classification. The benefit here is that the classifier outputs can be weighted in such a way that takes the reliability of both modalities into account. To date, many automatic audio-visual fusion techniques employ only the audio reliability measure [18] and/or the visual signal is assumed to be of a constant quality. Even if an observation signal is of high quality, the modality may still give a misclassification for two reasons 1) the correct subject class may be indistinguishable for the given expert, and may be consistently misclassified despite favorable test conditions 2) the

model for the correct subject may be a poor representation. If the reliability parameter is determined prior to classification, e.g. measuring the audio SNR, then a modality that performs poorly for a particular speaker can not receive a lower weighting.

Taking these points into account, it is better to calculate a reliability measure based on the classifier score distribution, as this can quantify both the train/test mismatch and classifier performance for a given test utterance. The fusion method employed uses a reliability measure based on the classifier output score distributions and hence takes the reliability of both modalities into account. The log-likelihood scores of the two modalities are normalized and integrated on a per test utterance basis. No prior statistics of the log-likelihood score distributions are employed. The scores are normalization by scaling the top M scores into the range $[0,1]$. Using a low value of M reduces the potential of audio-visual fusion, with the limit of $M=1$ amounting to fusion by method of voting [19] where no reliability information is considered. For a high value of M, the worst scores (outliers) can unfairly skew the normalized score distribution. Tests showed that the system performance degraded for $M<50$ and $M>100$. A value for M of 75 was chosen. Since the top M normalized scores are not the actual log-likelihoods, we use the likelihood $l(O|S_i)$ instead of $ll(O|S_i)$. The audio and visual test utterance observations are denoted by O_A and O_V respectively. The individual scores can be combined using the weighted *sum* or *product* rules. It was shown in [15], both theoretically and empirically, that the sum rule is more robust to classifiers errors than the product rule, thus the *weighted* sum rule is a good choice, particularly in this study where both of the audio and visual modalities may be highly degraded. The weighted sum likelihood that O was produced by the i^{th} speaker S_i, is:

$$l_{AV}(O_A, O_V \mid S_i) = \alpha_A l_A(O_A \mid S_i) + \alpha_V l_V(O_V \mid S_i), \text{ where } \alpha_A + \alpha_V = 1,\ \alpha_A, \alpha_V \in [0,1], \tag{2}$$

where α_A and α_V denote the audio and visual weights.

If the highest ranked class receives a high score and all of the other classes receive relatively low scores, then the confidence level is high. Conversely, if all the classes receive similar scores, the confidence is low. Various metrics exist, which can be used to capture this score confidence information. Examples include, *score entropy* [18], *score dispersion* [18], *score variance* [5], and *score difference* [20]. For a test observation O_m, we have the set of M ranked normalized scores $\{l(O_m|S_i)\}_{i=1...M}$. The score difference, ξ, between the two highest ranked confidence scores is calculated as

$$\xi_m = l_m(O_m \mid S_\alpha) - l_m(O_m \mid S_\beta), \quad \mathrm{m} \in \{\mathrm{A}, \mathrm{V}\}, \tag{3}$$

where S_α and S_β are the speakers achieving the best and second best scores respectively, and m denotes the modality. The difference of the top two best scores was employed for this study because, even though it is computational inexpensive, it performed well across all levels of audio and visual degradation. A high ξ value indicates a confident score whereas a low value indicates a score of poor confidence.

A mapping between the reliability estimate and α_A / α_V is required. A sigmoidal mapping [21], [18], [2], can be used, but the parameters of the sigmoid curve require training. Another option is to form corresponding bins of reliability estimates and α_A / α_V values, effectively a lookup table, but again this requires training. Considering the small amount of audio-visual training data available it was decided to use a non-learned approach to automatically selcet the α_A / α_V values. This was carried out as follows:

1. For each specific identification trial (user interaction), the system is presented with two modality observations, O_A and O_V.
2. The two classifiers each generate a set of N match scores, which are normalized to give the sets of M ranked scores $\{l(O_A|S_i)\}_{i=1...M}$ and $\{l(O_2|S_i)\}_{i=1...M}$, and the reliability estimates ξ_1 and ξ_2 are calculated using Equation 3.
3. α_V is automatically swept from 0 to 1 in steps of 0.05. For each of these α_V values, the modality score lists $\{l(O_A|S_i)\}$ and $\{l(O_V|S_i)\}$ are combined using Equation 2 (with $\alpha_A=1-\alpha_V$), to give the combined score set $\{l(O_A,O_V|S_i)\}_{i=1...N}$.
   The combined score set is subsequently normalized as before, to give $\{l(O_A,O_V|S_i)\}_{i=1...M}$, and the combined reliability estimate, ξ_{AV}, is calculated. Effectively, ξ_{AV} is a weighted combination of the individual reliabilities.
4. α_V is automatically selected to maximizes ξ_{AV} for the given test using Equation 4, to give α_{Vopt} and $\alpha_{Aopt}=1-\alpha_{Vopt}$. The maximum ξ_{AV} value should correspond to the combined scores of highest confidence, i.e. maximizes the score separation between the highest ranked class and the other classes. Finally, we combine $\{l(O_A|S_i)\}$ and $\{l(O_V|S_i)\}$ using α_{Aopt} and α_{Vopt}, to make the final decision.

$$\alpha_{2\,opt} = \arg\max_{\alpha_2 \in [0,1]} \{\xi_{12} | \alpha_2\}. \tag{4}$$

It should be noted that the above procedure is carried out for every identification trial, and thus the fusion weights are determined automatically in an unsupervised manner. For illustration, Fig. 4a gives an example arising from specific audio and visual test observations; the audio and visual score reliability estimate values, ξ_A and ξ_V, are 0.4 and 0.1 respectively. The variation of the combined audio-visual reliability estimate ξ_{AV}, is shown, as α_V is varied from 0 to 1; which reaches a maximum value of 0.64 for an α_V=0.4 For this test, 0.4 and 0.6 (1-0.4) are chosen for α_V and α_A respectively.

This fusion method can take account of a noisy audio and/or visual signal, and also, of either classifier performing poorly, and choose the weights appropriately. The advantage of this method is that, while being adaptive, training of the fusion parameters is not required. No assumption has been made about the type of audio/visual noise present. This is important because learned noise statistics that are used for the reliability mapping have been previously shown to vary with the type of noise degradation [18]. This compromises the mapping, as it must perform well for all types of noise (audio/visual), and not just for one specific type. Our method is similar to that of [17] where the difference of log likelihoods values between the first and other hypotheses is maximized by optimizing the stream weights. The scheme differs from the method described here because it is iterative, operates at the frame level and is applied to speech recognition whereas our scheme operates at the utterance level.

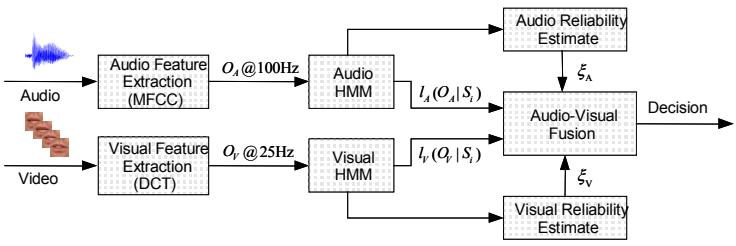


Fig. 2. Separate audio and visual classifiers and fusion based on modality reliabilities

5 Results and Discussion

5.1 Individual Audio and Visual Results

Audio Results: The audio HMMs performed best with 11 states and two Gaussian mixtures per state. The performance w.r.t. audio degradation is given in the second row of Table 2. The accuracy at 48dB is 97.6%. At 21dB the accuracy is 37%. The audio performed very well under "clean" testing conditions, however the roll off w.r.t. SNR is very high (shown in Fig. 4b), highlighting the effect of a train/test mismatch.

Visual Results: We first tested the effect of the number of HMM states on the performance of the four visual feature types described in Section 3, namely *static*, *delta*, *acceleration* and *SDA*. In a HMM, each state is associated with a locally stationary section of the speech signal, whereas the state transitions model the signals' temporal nature It was expected that the dynamic visual features, would perform better using more HMM states compared to the static features. The number of states was increased from one (with one Gaussian mixture per state), until a trend became apparent. The results of this are shown in Fig. 3a. The number of states, that maximized the visual scores for the four feature types, are given in Table 1. The *static* features performed best with just a single state and decreases steadily with increasing number of states. This suggests that the *static* features would be better modeled using a simpler Gaussian mixture model (GMM) approach, as in [20]. The number of states, which maximized the scores for the *delta* and *acceleration* features, are 18 and 15 respectively. The higher number of states required for the *delta/acceleration* features is expected due to their dynamic information. The *SDA* features are modeled best using four states. This suggests a conflict, with the static features performing best with a single state, while dynamic features perform better with a multi-state model.

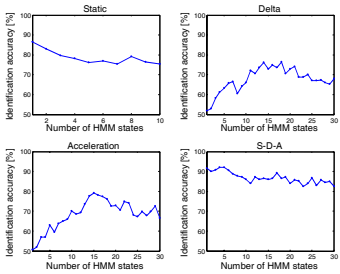


Fig. 3a. Visual accuracy vs. number of HMM states, for the four types of visual features

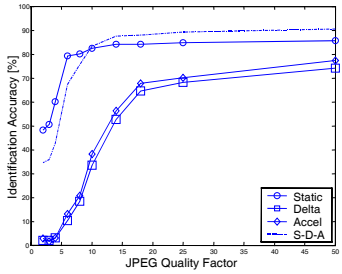


Fig. 3b. Visual accuracy vs. the ten levels of JPEG QF for the four types of visual features

Table 1 and Fig. 3b show how the visual features perform w.r.t. JPEG QF. The best visual performance of 90.4% is surprisingly high, considering that only the lip information is employed. The *static* features show a high level of robustness (48% at a QF=2). This conforms with the high level of speech recognition robustness to JPEG compression reported in [13]. While the *SDA* features outperform the *static* features for high QFs (*SDA* 90.4% versus *static* 85.9% at a QF=50), the performance at a QF=2 is 37.7%, which is poorer than the *static* performance. However, the dynamic

features perform very poorly at low QF levels, both falling to around 2% at a QF=2. It should be noted that in an applied scenario, where the ROI is automatically segmented, rather than manually, poorer robustness would be expected.

Table 1. Visual speaker identification accuracy versus the ten levels of JPEG QF

JPEG QF	50	25	18	14	10	8	6	4	3	2	HMM States
Static	85.9	85.1	84.3	84.3	82.7	80.2	79.4	60.5	50.8	48.0	1
Delta	74.1	68.1	64.5	52.6	33.5	18.3	10.4	3.2	2.0	2.0	18
Accel.	77.3	70.1	67.7	56.2	38.2	20.7	13.1	3.2	1.6	2.8	15
S-D-A	90.4	89.2	88.0	87.6	83.3	75.7	67.3	42.6	35.9	34.7	4

5.2 Audio-Visual Results

The *static* visual feature scores were integrated with the audio scores, because the *static* visual features exhibited the highest robustness to visual degradation. The results for the automatic fusion method described above are shown in Fig. 4b and Table 2. The automatic fusion accuracies are higher than either of the audio and visual modalities, at all degradation levels. This highlights the complementary nature of the audio and visual speech signals and the fusion robustness. At the most severe mismatch levels tested (SNR 21dB, QF 2), the audio, visual, and audio-visual accuracies are 37.1%, 48%, and 71.4% respectively, giving a relative improvement of 92.5% on the audio and 49% on the visual accuracies.

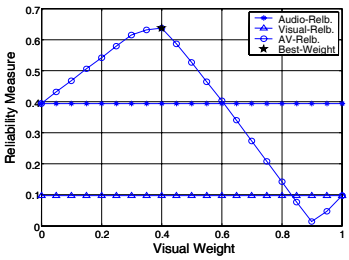


Fig. 4a. Example of how the AV reliability estimate varies w.r.t. α_V for a particular test

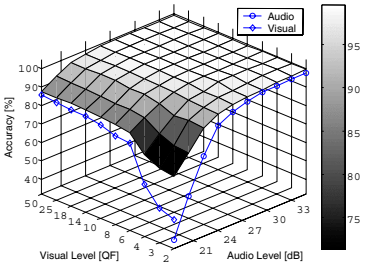


Fig. 4b. Identification accuracies for audio, visual and automatic audio-visual fusion

Table 2. Automatic audio-visual fusion accuracies for ten levels of audio/visual degradation

	dB	48	45	42	39	36	33	30	27	24	21
QF	A / V	**97.6**	**97.6**	**97.2**	**97.2**	**95.6**	**93.1**	**89.1**	**75.8**	**57.3**	**37.1**
50	**85.9**	99.2	99.2	99.2	99.2	99.2	99.2	98.0	96.8	92.7	87.5
25	**85.1**	99.2	99.2	99.2	99.2	99.2	99.2	98.0	96.8	92.3	87.5
18	**84.3**	99.2	99.2	99.2	99.2	99.2	99.2	98.4	97.2	92.3	87.9
14	**84.3**	99.2	99.2	99.2	99.2	99.2	99.2	98.4	97.2	91.5	87.5
10	**82.7**	99.2	99.2	99.2	99.2	99.2	99.2	98.0	96.0	91.1	87.1
8	**80.2**	99.2	99.2	99.2	99.2	99.2	98.8	97.6	96.0	91.1	86.3
6	**79.4**	99.6	99.2	99.2	99.2	99.2	98.8	97.2	94.4	89.9	85.1
4	**60.5**	99.6	99.6	99.6	99.6	99.6	97.6	96.4	91.1	84.7	76.6
3	**50.8**	99.6	99.6	99.6	98.8	98.0	97.6	95.6	91.9	81.5	72.6
2	**48.0**	99.6	99.2	99.2	98.4	98.0	97.2	95.2	91.1	80.6	71.4

6 Conclusions

The lack of large audio-visual databases still remains a major setback for audio-visual research. The XM2VTS database addressed this shortcoming to some extent. How-

ever, due to its extremely well controlled recording conditions, it does not represent real world scenarios. This is highlighted by the high visual speaker identification accuracies achieved in this study, the best been 92% (based on just the mouth ROI). The BANCA database [22] and the new VALID [23] audio-visual database consisting of 106 subjects, recorded under controlled and unconstrained scenarios, will provide data of a more practical nature for audio-visual research.

Experimental results have being presented for automatic audio-visual fusion with application to speaker identification. A large data set of 251 subjects has been tested. Ten levels of train/test mismatch of not only the audio modality but also the visual modality have been examined. The audio-visual fusion methodology uses reliability estimates of both the audio and visual modalities. A benefit of the approach described is that audio-visual training data is not required to tune the fusion process. The results are encouraging with the audio-visual accuracies exceeding both the audio and visual accuracies at all levels of audio and visual degradation. Importantly, integrating a highly mismatched/degraded modality (e.g. audio 37.1%) with a "clean" modality (e.g. visual 85.9%) does not result in catastrophic fusion (audio-visual 87.5%). The fusion method is computationally inexpensive. These results were achieved with the fusion block having no prior knowledge of the level/type of audio/visual degradation.

Further work includes the investigation of automatically fusing the output scores from a face classifier with the audio-visual scores, and the testing of different types of audio and visual noise. The system described has applications in practical scenarios, such as human computer interfaces and multi-modal biometric systems for robust person verification. In conclusion, this paper describes an audio-visual speaker identification system using unsupervised adaptive classifier fusion that demonstrates a high level of robustness to both adverse audio and visual testing conditions.

References

1. N. A. Fox and R. B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," Proc. of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, pp. 743-751, June 2003.
2. G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audiovisual Speech," Proceedings of the IEEE, vol. 91, pp. 1306-1324, Sept. 2003.
3. C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," IEEE Transactions on Multimedia, vol. 4, pp. 23-35, Mar 2002.
4. R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 955-966, Oct. 1995.
5. T. J. Wark, S. Sridharan, and V. Chandran, "The use of Speech and Lip Modalities for Robust Speaker Verification under Adverse Conditions," Proceedings of the IEEE International Conference on Multimedia Computing and Systems, pp. 812-816, June 1999.
6. C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "Adaptive Classifier Integration for Robust Pattern Recognition," IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, vol. 29, pp. 902-907, Dec. 1999.
7. C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," Digital Signal Processing, vol. 14, pp. 449-480, 2004/9 2004.
8. "The XM2VTS database; http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/."
9. J. P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, vol. 85, pp. 1437-1462, Sept. 1997.

10. S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.1)". Cambridge University Engineering Department: Microsoft Corporation, 2001.
11. J. Luettin, "Speaker verification experiments on the XM2VTS database," in IDIAP Communication 98-02: IDIAP, Martigny, Switzerland, 1999.
12. I. Matthews, T. F. Cootes, J. A. Bangham, J. A. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 198-213, Feb. 2002.
13. G. Potamianos, H. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," Proceedings of the IEEE International Conference on Image Processing, ICIP 98, Chicago, vol. 3, pp. 173-177, Oct. 1998.
14. A. N. Netravali and B. G. Haskell, "Digital Pictures": Plenum Press, 1998.
15. J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 226-239, March 1998.
16. S. Bengio, "Multimodal speech processing using asynchronous Hidden Markov Models," Information Fusion, vol. 5, pp. 81-89, June 2004.
17. S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04). vol. 1, pp. 857-860 2004.
18. M. Heckmann, F. Berthommier, and K. Kristian, "Noise Adaptive Stream Weigthing in Audio-Visual Speech Recognition," EURASIP Journal on Applied Signal Processing, vol. 2002, pp. 1260-1273, Nov. 2002.
19. J. Kittler and F. M. Alkoot, "Sum versus Vote Fusion in Multiple Classifier Systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 110-115, Jan. 2003.
20. T. Wark and S. Sridharan, "Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification," Digital Signal Processing, vol. 11, pp. 169-186, July 2001.
21. N. A. Fox, R. Gross, P. de Chazal, J. F. Cohn, and R. B. Reilly, "Person Identification Using Automatic Integration of Speech, Lip, and Face Experts," ACM SIGMM workshop on Biometrics Methods and Applications, Berkley, CA., pp. 25-32, Nov. 2003.
22. "The BANCA Database; http://www.ee.surrey.ac.uk/Research/VSSP/banca/."
23. "The VALID Database; http://ee.ucd.ie/validdb/."

Speaker Identification Using the VQ-Based Discriminative Kernels

Zhenchun Lei, Yingchun Yang, and Zhaohui Wu

College of Computer Science and Technology, Zhejiang University
310027 Hangzhou, P.R. China
{leizhch,yyc,wzh}@zju.edu.cn

Abstract. In this paper, a class of VQ-based discriminative kernel is proposed for speaker identification. Vector quantization is a well known method in speaker recognition, but its performance is not superior. The distortion of an utterance is accumulated, but the distortion source distribution on the codebook is discarded. We map an utterance to a vector by adopting the distribution and the average distortions on every code vector. Then the SVMs are used for classification. A one-versus-rest fashion is used for the problem of multiple classifications. Results on YOHO in text-independent case show that the method can improve the performance greatly and is comparative with the VQ and the basic GMM's performances.

1 Introduction

Support vector machines (SVMs) [1] have got more attention in speaker identification recently for its superior performance. SVM is based on the principle of structural risk minimization. Experimental results indicate that SVMs can achieve a generalization performance that is greater than or equal to other classifiers, while requiring significantly less training data.

The methods using SVMs in speaker verification and identification can be divided into frame-based and utterance-based. In the former, every frame is scored by the SVM and the decision is made based on the accumulated score over the entire utterance [2]. Another is utterance-based approaches which map an utterance into a vector as SVM's input, and the researchers focus on how to construct a better kernel dealing with utterances having different lengths, such as fisher kernel [3] and dynamic time-alignment kernel [4].

Vector quantization (VQ) [5] is a well known method in speaker recognition for its simplicity, but it can only get moderate performance. In VQ method, each speaker is characterized with several prototypes known as code vectors, and the set of code vectors of each speaker is referred to as that speaker's codebook. A speaker's codebook is trained to minimize the quantization error for the training data from that speaker.

In this paper, we combined the SVM and codebook in VQ method for speaker identification. When using VQ method, the distortion of an utterance is accumulated from all frames, but the distortion source distribution on the codebook is discarded. So we constructed a new discriminative kernel by adopting this distribution and the SVM was used to classify. Our experiments were tested on YOHO database and in text independent speaker identification case.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 797–803, 2005.
© Springer-Verlag Berlin Heidelberg 2005

This paper is organized in the following way: section 2 is the VQ model's brief introduction. In section 3 we review the SVMs theory briefly and current apply approaches in speech recognition and speaker recognition using SVMs. The new discriminative kernels will be described in section 4. Section 5 presents the experimental results on YOHO. Finally, section 6 is devoted to the main conclusions.

2 VQ Method

One of the most successful text-independent recognition methods is based on vector quantization (VQ). It provides an effective way to describe the personal speech characters, but it is too simple to get better performance.

In the training phrase, a mathematical model called VQ codebook is constructed for each speaker from their speech samples. VQ codebooks consisting of a small number of representative feature vectors are used as an efficient means of characterizing speaker-specific features. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. Each cluster is represented by a code vector, which is the vectors average of the cluster. There are some clustering algorithms such as k-means, LBG [6], LVQ [7] and so on. In our experiments the LBG algorithm was used.

In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker and the VQ distortion accumulated over the entire input utterance is used to make the recognition decision. Denote the sequence of feature vector extracted from the unknown speaker as $X = \{x_1,...,x_T\}$. The goal is to find the best matching codebook from all codebooks $C = \{c_1,...,c_N\}$. The average quantization distortion $\overline{D}$ that results from coding a verification utterance in codebook is:

$$\overline{D} = \frac{1}{T}\sum_{i=1}^{T} d(x_t, \hat{c}_t) \tag{1}$$

where

$$\hat{c}_t = \arg\min_{c_i \in C} d(x_t, c_i) \tag{2}$$

We use this average quantization distortion in making the decision.

3 Support Vector Machine

3.1 Theory

SVM theory [1, 12] is mainly from the problem of binary classification, and its main idea can be concluded as the following two points: it constructs a nonlinear kernel function to present an inner product of feature space. It implements the structural risk minimization principle in statistical learning theory by generalizing optimal hyperplane with maximum margin between the two classes.

The hyperplane is defined by $x \cdot w + b = 0$ that leaves the maximum margin between the two classes. It can been shown that maximizing the margin is equivalent to

minimizing an upper bound on the generalization error of the classifier, providing a very strong theoretical motivation for the technique. The vector w that maximizes the margin can be shown to have the form:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{3}$$

where the parameters α_i are found by solving the following quadratic programming (QP) problem.

$$\max_{\alpha}(\sum_{i} \alpha_i + \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j) \tag{4}$$

subject to:

$$\begin{aligned} &\sum_{i} \alpha_i y_i = 0 \\ &0 \le \alpha_i \le C \end{aligned} \tag{5}$$

The main feature of the SVM is that its target functions attempts to minimize the number of errors made on the training set while simultaneously maximizing the margin between the individual classes. This is an effective prior for avoiding overfitting, which results in a sparse model dependent only on a subset of kernel functions.

The extension to non-linear boundaries is acquired through the use of kernels that satisfy Mercer's condition. The kernels map the original input vector x into a high dimension space of features and then compute a linear separating surface in this new feature space. In practice, the mapping is achieved by replacing the value of dot production between two data points in input space with the value that results when the same dot product is carried out in the feature space. The following is formations:

$$\max_{\alpha}(\sum_{i} \alpha_i + \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)) \tag{6}$$

The kernel function K defines the type of decision surface that the machines will build. In our experiments, the radial basis function (RBF) kernel and the polynomial kernel are used and they take the following forms:

$$\begin{aligned} k_{rbf}(x_i, x_j) &= \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma}\right)^2 \right] \\ k_{poly}(x_i, x_j) &= (x_i \cdot x_j + 1)^n \end{aligned} \tag{7}$$

where σ is the width of the radial basis function and n is the order of the polynomial. The use of kernels means that an explicit transformation of the data into the feature space is not required.

3.2 Discriminative Kernels for Speaker Recognition

Recent several approaches using SVMs have been proposed in speech applications and speaker recognition. A well-known kernel is the fisher kernel made by Jaakkula and Haussler [3], which has been explored for speech recognition in [8] and speaker

recognition in [10]. Denoting $p(x \mid \theta)$ is a generative model, where θ are its parameters, the mapping function is an analogous quantity to the model's sufficient statistics as following:

$$U_\theta(x) = \nabla_\theta \log(p(x \mid \theta)) \tag{8}$$

Each component of U is a derivative of the log-likelihood score for the input vector x with respect to a particular parameter.

Constructing a superior kernel function for utterances can be difficult and still a challenge. The dynamic time alignment kernel (DTAK) is developed by incorporating an idea of non-linear time alignment into the kernel function [9]. The pair HMM a kernel is similar to DTAK, which may be more suited to a text dependent speaker verification task where the model topology is predefined [10]. Campbell also introduced the sequence kernel derived from generalized linear discriminates in [11]. Like them, we constructed a new kernel function to deal with the variable length utterances.

4 VQ-Based Discriminative Kernels

In VQ method, the decision is made depending on the scores (average distortions of whole utterance) from all models and selects the smallest score as the result. It only considers the score, but where the score is from is ignored. In our method, we map an utterance into a fixed-size vector according the score source from the codebook in VQ algorithm.

The utterance, X, is denoted as a sequence of acoustic feature vectors $X = \{x_1, ..., x_n\}$, and the vector x_i have d components. A codebook $C = \{cb_1, ..., cb_{cn}\}$ has been given from VQ algorithm. Now we construct a new kernel.

Campbell [11] brought a generalized linear discriminate sequence (GLDS) kernel based on Taylor approximate, which defined the mapping $X \to \bar{b}$ as:

$$X \to \frac{1}{n}\sum_{i=1}^{n} b(x_i) \tag{9}$$

We also adopted this idea and extended to combine the codebook. We mapped an utterance to multiple $\bar{b}$ by mapping every frame to its nearest code vector, so the result is a matrix rather than a vector and then be translated to a vector by simply expanding. We simply set $b(x)$ to be the difference vector to the code vector (the polynomial of the difference vector will be test in future like GLDS) in our current experiments.

For x_i , we can find the vector from codebook which has the minimal distance:

$$cb_t = \arg\min_{j=1...cn}\{d(x_i, cb_j)\} \tag{10}$$

then map x_i to an matrix:

$$\mathrm{M}(\mathrm{x}_{\mathrm{i}}) = [\mathrm{m}_1, \ldots, \mathrm{m}_{\mathrm{cn}}] \tag{11}$$

where

$$M_k = \begin{cases} x_i - cb_t, k = t \\ 0, else \end{cases} \tag{12}$$

The matrix M has d*cn size, which is same to the codebook, it reflect the distance between x_i and C at dimension-level. Now the whole utterance X:

$$\Phi(X) = \frac{1}{n}\sum_{i=1}^{n} M(x_i) \tag{13}$$

After expanding $\Phi(X)$ to one-dimension simply, a d*cn size vector can be got, which is the mapping result of an utterance X based on the codebook C. Also we can descript a linear kernel between $\Phi(X)$ and $\Phi(Y)$ directly:

$$K_{linear}(X,Y) = \sum_{i=1}^{d}\sum_{j=1}^{cn} \Phi(X)_{ij} \cdot \Phi(Y)_{ij} \tag{14}$$

And the polynomial and RBF kernel can also be constructed in the same way.

$$K_{poly}(X,Y) = (\sum_{i=1}^{d}\sum_{j=1}^{cn} (\Phi(X)_{ij} \cdot \Phi(Y)_{ij}) + 1)^n$$

$$K_{rbf}(X,Y) = \exp\left[-\frac{1}{2} \cdot \frac{\sum_{i=1}^{d}\sum_{j=1}^{cn} (\Phi(X)_{ij} - \Phi(Y)_{ij})^2}{\sigma^2} \right] \tag{15}$$

where σ is the width of the radial basis function and n is the order of the polynomial.

5 Experiments

Our experiments were performed using the YOHO database. This database consists of 138 speaker prompted to read combination lock phrases, for example, “29_84_47”. Every speaker has four enrollment sessions with 24 phrases per session and 10 verify sessions with 4 phrases per session. The features are derived from the waveforms using 12th order LPC analysis on a 30 millisecond frame every 10 milliseconds and deltas computed making up a twenty four dimensional feature vector. Mean removal, preemphasis and a hamming window were applied. Energy-based end pointing eliminated non-speech frames.

The 50 speakers, labeled 101 to 154, were used in our experiments and 50 SVMs were trained one speaker from all other 49 speakers. For every speaker, a LBG algorithm was used on its 96 enroll utterances, and a codebook could be got, which is the same to VQ model. Then all speakers' utterances were been mapped based on the codebook, and the support vector machines were training using the vectors as inputs.

The SVM is constructed to solve the problem of binary classification. For N-class, the general method is to construct N SVMs [12]. The ith SVM will be trained with all of the examples in the ith class with positive labels, and all other examples with negative labels. We refer to SVMs trained in this way as 1-vs-r (one-versus-rest) SVMs. Another method is 1-vs-1 (one-versus-one) SVMs, which construct $K=N(N-1)/2$ classifiers and each classifier be trained on only two out of N classes. In our experiments, the 1-vs-r method was adopted.

In training phase, 50 SVMs were trained discriminating one speaker from all other 49 speakers. So the positive sample number is 96 and the negative sample mumble is 96*49=4704. Training SVMs rely on quadratic programming optimizers, so it is not easily to large problems. There are some algorithms for this problem and the SMO [14] is used.

In our experiments, we used two codebook size, 64 and 128, which generated the mapped vectors having 24*64=1536 and 24*128=3072 dimensional space respectively. The SVM with linear kernel, polynomial kernel (n=3, C=1) and radial basis function kernel (σ =10, C=1) were be training.

Since the GMM [13] is very popular for speaker recognition, we also tested the GMM with the same data and showed the result here as a reference. The basic GMMs consisted of 64 and 128 component Gaussian mixture model with diagonal covariance matrices. Table 1 shows the result.

Table 1. Performance of each model for text independent speaker identification experiments on YOHO database

Models	Codebook size	Error rate (%)
VQ	64	8.7
SVM/VQ(linear kernel)	64	5.6
SVM/VQ(rbf kernel)	64	4.4
SVM/VQ(polynomial kernel)	64	4.0
Basic GMM	64	5.4
VQ	128	6.1
SVM/VQ(linear kernel)	128	4.2
SVM/VQ(rbf kernel)	128	3.5
SVM/VQ(polynomial kernel)	128	3.1
Basic GMM	128	3.2

Table 1 show that the linear kernel, rbf kernel and polynomial kernel of SVM/VQ improve 35.6%, 49.4% and 54% at 64 codebook size respectively. And when 128 codebook size, they are 31%, 43% and 49%. Comparing to the basic GMM, our approach is superior when codebook size is 64 and is comparative when 128.

6 Conclusions

A class of discriminative kernels for text independent speaker identification was presented. The basic idea is mapping the variable length utterances to the fixed size vectors using a codebook. The mapping procedure is running on the minimum distance, and the score source in VQ model is taken into account. The experiments on YOHO show that our method can improve the VQ model's performance greatly.

The same idea can also be applied in GMM which is used widely in speaker recognition. And the relations between frames should be considered. Like the GLDS kernel, the polynomial will be used for improvement. Those are our future research.

Acknowledgment

This work is supported by National Natural Science Foundation of P.R.China (60273059), Zhejiang Provincial Natural Science Foundation for Young Scientist of P.R.China (RC01058), Zhejiang Provincial Natural Science Foundation (M603229) and National Doctoral Subject Foundation (20020335025).

References

1. V.Vapnik. Statistical Learning Theory. Johm Wiley and Sons, New York, (1998)
2. V.Wan, W.M.Campbell, Support Vector Machines for Speaker Verification and Identification. in Proc. Neural Nerworks for Signal Processing X, (2000) 775-784
3. T.S.Jakkola and D.Haussler. Exploiting generative models in discriminative classifiers. In Advances in Neural Information Processing System 11, M.S.Kearns, S.A.Solla, and D.A.Cohn, Eds. MTT Press, (1999)
4. Nathan Smith, Mark Gales, and Mahesan Niranjan, Data-dependent kernel in SVM classification of speech patterns. Tech.Rep. CUED/F-INFENG/TR.387, Cambridge University Engineering Department, (2001)
5. A.E.Rosenberg and F.K.Soong, Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Comput. Speech Lang., vol 22 (1987) 143-157
6. Y.Linde, A.Buzo, and R.M.Gray, An algorithm for vector quantizer desigh, IEEE Trans. Commun., vol.20, pp.84-95 (1980)
7. T.Kohonen, The self-organizing map, Proceedings of the IEEE, vol. 78, (1990) 1464-1480
8. Shai Fine, Jiri Navratil, and Ramesh A.Gopinath, A hybrid GMM/SVM approach to speaker recognition. in ICASSP (2001)
9. Hiroshi Shimodaira, Kenichi Noma, Mitsuru Nakai and Shigeki Sagayama, Dynamic Time-Alignment Kernel in Support Vector Machine, NIPS, (2001) 921-928
10. Vincent Wan, Steve Renals, Valuaution of Kernel Methods for Speaker and Identification. in Proc. ICASSP, (2002)
11. W.M.Campbell, GENERALIZED LINEAR DISCRIMINANT SEQUENCE KERNEL FOR SPEAKER RECOGNITION, in Proc. ICASSP, (2002) 161-164
12. C.J.C.Burges, A tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, vol.2, no.2, (1998) 1-47
13. D.A.Reynolds and R.C.Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Processing, vol.3, (1995) 72-83
14. J. Platt.Fast training of SVMs using sequential minimal optimisation. Advances in Kernel Methods: Support Vector Learing, MIT press, Cambridge, MA, (1999) 185-208
15. M.Schmidt and H.Gish, Speaker Identification via Support Vector Machies. in Proc.ICASSP, (1996) 105-108
16. Tomi Kinununen and Pasi Franti, Speaker Discriminative Weighting Method for VQ-based Speaker Identification. Proc.3rd International Conference on Audio- and video-based biometric person authentication, Halmstad, Sweden, (2001) 150-156

Exploiting Glottal Information in Speaker Recognition Using Parallel GMMs

Pu Yang, Yingchun Yang, and Zhaohui Wu

College of Computer Science and Technology
Zhejiang University, Hangzhou, 310027, P.R. China
{kuvun,yyc,wzh}@zju.edu.cn

Abstract. The information of the vocal tract and the glottis are two kinds of sources which can characterize speakers. Though the former one has archived quite good performance in automatic speaker recognition (ASR) tasks, the glottal information behaves poorly when used individually. This work explores how to combining vocal tract and glottal information in an efficient and effective way. Taking into account the short-term correlation between them, our improved joint probability function model of the corresponding features is first proposed. Then we present a novel integrating system which uses parallel Gaussian Mixture Models (GMM) grounded on this function. Together with the traditional GMM, it also forms a hybrid model. Both methods were applied to YOHO and SRMC corpus, and experimental works show promising results.

1 Introduction

Vocal tract is considered as the most important part of human phonation system. Information from it has been popular adopted in many speaker identification or verification systems because the envelop shape could provide speaker dependent factors [1]. Glottis, on the other hand, is also a key part of human phonation apparatus. This information reflects the vibration frequency of vocal tract which is also known to carry specific speaker information [2] [3].

There are many speaker recognition systems that make use of either vocal tract information or glottal information. Those based on vocal tract work well when speech is recorded in clean conditions, but their performance drops considerably when speech was recorded in hostility environments [4]. Meanwhile, speaker recognition systems exclusively based on glottal information are more robust to noise and channel distortions, but they can not do well when the number of speakers becomes large. So, incorporating glottal information and vocal tract information in an effective and efficient way has become a major problem.

Various techniques have been investigated for handling this integration at various levels. One kind of them is to fuse the short-term glottal features with short-term vocal tract features at front-end as the input of classifier. Sonmez et al. [5] and Shao et al. [6] are two examples. Another kind of integrations is to model simultaneously the statistical distribution of the short-term vocal tract features and the long-term glottal features. Peskin et al. fall into this category.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 804–812, 2005.
© Springer-Verlag Berlin Heidelberg 2005

They have exploited statistics of long-term glottal information to form certain high-level model which acts as complementary part of vocal information based model. Work of Arcienega et al. [8] also belongs to this kind. In spite of the good performances of above systems, the complex mechanisms involved in speech production imply dependence of glottis and vocal tract. And Ezzaidi et al. [9] have contributed a great work about a rudimental model on such dependence.

In this paper, we have improved their joint probability function model that takes into account the correlation between source and vocal tract information. Some modifications have been added to the model for the sake of precise modelling of glottal information and data consideration. Based on this improved joint probability function, an integrating model which uses parallel GMMs is then presented. And it shows good performance in the automatic speaker recognition task, especially when combined with the traditional GMM model.

The remainder of this paper is organized as the following: section 2 gives our proposed model. Then, the framework of speaker recognition is presented in section 3. Experiments and discussions are done in section 4. Conclusions are given in the final part.

2 Proposed Model

2.1 Motivation

It is easy to think that the speed of glottis vibration would metamorphose the shape of vocal tract more or less. In [9], Ezzaidi and el. have illustrated the role of glottal features when dependence of the glottal source and vocal tract are maintained. They found that short-term glottal information and vocal tract information could be jointly exploited and then established a probability model of feature vectors assuming the priori knowledge of glottal information. That is, each speaker is supposed to be defined by its probability function:

$$f_s(\boldsymbol{x}_i, y_j) = P_s(\hat{x} = \boldsymbol{x}_i, \hat{y} = y_j) = f_s(\boldsymbol{x}_i / y_j) f_s(y_j). \tag{1}$$

where $\hat{x}_n$ are l-dimension MFCC [11] which are extracted from a windowed signal centered at time $n\Delta t$. And $\hat{y}_n$ is the one-dimension fundamental frequency which is extracted simultaneously. And $f_s(y_j)$ is a priori probability of a fundamental frequency equal to y_j, and $f_s(\boldsymbol{x}_i/y_j)$ is a posteriori probability of observing a MFCC vector equal to $\boldsymbol{x}_i$, given knowledge of the fundamental frequency y_j.

However, their method treats every subspace of glottal information contributes equally in speaker recognition and only considers the glottal distribution within each interval. In practice, this means not only ignores the different contributions each subspace may provide in recognition, but also loses sight of the holistic distribution of glottal information which could better character speaker's identity than separateness' does. What is more, such a linear division of glottal information would lead to problem of data sparsity whatever in training or testing process.

Due to the shortcomings described above, we should modify this joint function model in the following aspects:

(1) Models trained from different subspaces of glottal information contributes differently to speaker identity; the farer the subspace is from the average of speaker's glottal distribution, the less this model contributes to the final decision.

(2) The division of subspaces of glottal information should be based on its distribution of the speaker instead of linear partition which holds no speaker dependent information.

2.2 Subspace Divisions

In order to solve the $f_s(\boldsymbol{x}_i, y_j)$,many efforts have been made on estimation and integration of $f_s(\boldsymbol{x}_i/y_j)$. In previous work, the space $(\boldsymbol{x}, y)$ was divided linearly into subspaces according to the fundamental frequency scale. That is, each subspace is continuous in scale and has an equal length of fundamental frequency range. Section 2 has pointed out the problems of this method. So, as what is mentioned early, we propose a new division way in which the speaker's glottal distribution could be utilized.

It is supposed that the fundamental frequency features $\hat{y}$ of one speaker satisfy a normal distribution $N(\mu, \sigma)$, with its mean value being μ and covariances being σ. Its probability function can be written as:

$$f_s(y) = \frac{1}{\sqrt{2\pi}\sigma} exp[-\frac{(y-\mu)^2}{2\sigma^2}]. \tag{2}$$

Then we define $I_k, k = 1, 2, \ldots, N$ as sub-intervals of the fundamental frequency set with total range being G. N is the number of intervals with

$$I_1 \cup I_2 \cup \ldots \cup I_N = G. \tag{3}$$

For each $k \in \{1, 2, \ldots, N\}$, the subspace I_k can be written as:

$$I_k = [\mu - \beta_k, \mu - \beta_{k-1}) \cup [\mu + \beta_{k-1}, \mu + \beta_k), \tag{4}$$

where

$$\int_{\mu-\beta_k}^{\mu+\beta_k} f_s(y) = \frac{k}{N}. \tag{5}$$

As described in Figure 1, each subspace H_k in the space $(\boldsymbol{x}, y)$ is associated with a fundamental interval I_k. It can be inferred from the definition of I_k that our division of H_k takes advantage of fundamental frequency's distribution and got equal features in every subspace. For each H_k, we suppose that the probability function $f_s(\boldsymbol{x}_i/y_j)$ is independent of the fundamental frequency inside the interval I_k. That is, inside an interval I_k the fundamental frequency is supposed to act the same way for our model. So,

$$f_s(\boldsymbol{x}_i/y_j) = P(\hat{x} = \boldsymbol{x}_i/I_k, speaker = s \ \ with \ \ y_j \in I_k). \tag{6}$$

And we suppose that the priori probability of fundamental frequency equal to y_j can be viewed as the sum of probability of the fundamental frequency within the interval which y_j belongs to:

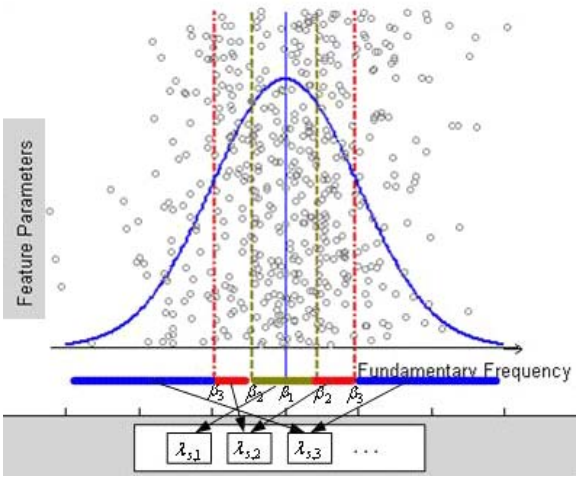


Fig. 1. Subspace division based on the distribution of fundamental frequency

$$f_s(y_j) = P(\hat{y} = y_j \ \ with \ \ y_j \in I_k) \cong \sum_{y_v \in I_k} P(\hat{y} = y_v). \tag{7}$$

According to the definition of I_k in Equation. (4) (5), we can get:

$$\begin{aligned} f_s(\boldsymbol{x}_i, y_j) &= f_s(\boldsymbol{x}_i / y_j) f_s(y_j) \\ &\cong f_s(\boldsymbol{x}_i / y_j) \sum_{y_v \in I_k} P(\hat{y} = y_v) = \frac{1}{N} f_s(\boldsymbol{x}_i / y_j) \end{aligned} \tag{8}$$

So in each subspace I_k, we can use the observing $f_s(\boldsymbol{x}_i/y_j)$ to train one model $\lambda_{s,k}$, and the number of models of one speaker would be equal to N.

2.3 Fusion of Subspace Models

We fuse N models of subspaces using weighted-sum rule [12] at decision level. Suppose the score subspace model $\lambda_{s,k}$ of speaker s is $score_{s,k}$. In order to distinguish the different contribution of different subspace, we put different weight $w_k$on them. The farther the subspace is from the mean of speaker's distribution, the less weight we give to it. So the final score of speaker s is:

$$score_s = \sum_{k=1}^{N} (w_k \cdot score_{s,k}). \tag{9}$$

where $0 \leq w_N \leq w_{N-1} \leq \ldots \leq w_1 \leq 1$.

There can be many realizations of w_k that could satisfy the above restriction. In the experiments, we used $w_k = 1/\sqrt{k}, k = 1, 2, \ldots, N$.

3 Framework of Speaker Recognition

3.1 Parallel GMMs for SR

Based on the proposed model, we describe the framework of our parallel GMMs speaker recognition system in this section.

As most automatic speaker recognition systems do, a Gaussian Mixture Model (GMM) with mixture number of 32 is used as λ in our system [10]. At the training stage, each speaker s gets parallel N GMM models $(\lambda_{s,1}, \ldots, \lambda_{s,N})$ from every subspace of speech partitions of that speaker. These GMMs and his fundamental frequency distribution parameters (μ_s, σ_s) compose the whole model of this speaker s; And for a certain speaker k, the test speech Z is also divided based on (μ_s, σ_s). Then resulting speech partitions are put into one corresponding GMM, $\lambda_{s,3}$ for example, of N parallel GMMs to get N scores, which are finally fused by Equation. (9).

3.2 Combination with Traditional GMM

In order to explore the complementary performance between proposed parallel GMMs and traditional GMM, we plan to combine their output scores. Suppose λ_s is the traditional GMM of speaker s which is trained from all MFCC features of whole fundamental frequency range. And its output score for a certain test speech Z is written as t_score_s. Meanwhile, parallel GMMs' output score for this speech file Z is written as p_score_s. To combine traditional GMM with parallel GMM, we use a trade of τ in getting the final testing score $score_s$,

$$score_s = \tau \cdot p_score_s + (1 - \tau) \cdot t_score_s. \tag{10}$$

4 Experiments and Discussions

4.1 Speech Database

We use two speech corpus to evaluate our methods in speaker recognition. One is the YOHO corpus [13] which has 138 speakers For each speaker in it, there are 4 ENROLL sessions with 24 sentences each, and 10 VERIFY sessions with 4 sentences each. The other corpus SRMC (Speaker Recognition in Multi-Channel Environment) [14]. It is a large multi-channel speech database which contains two rounds speech data of 303 speakers. Its data was recorded from mobile phone, PDA, telephone and microphone simultaneously. Each speaker in this corpus has Personal Information (PI)session of 3 sentences, Paragraph (PR) session of 1 sentence, and Mandarin Digit (MD), Dialect Digit (DD), Englisth Digit (ED), Province Phase (PP), Free Talking (FT) session of 10 sentences each. And here, we select all 138 speakers of YOHO corpus and all SRMC's 303 speaker's telephone channel data of the first round for our experiments.

Our speech was first passed through a pre-emphasize filter $H(z) = 1 - 0.97z^{-1}$; then a sliding Hamming window of 32ms and a shift of 16ms was positioned on the signal, in which the features was extracted. Our vocal tract

features vectors were composed by 16-dimension Mel Frequency Cepstral Coefficients (MFCC). At the same time, the fundamental frequency was extracted using the method which is based on Subharmonic-to-Harmonic Ratio (SHR) provided by Sun [15]. Only the problem of text-independent speaker identification of close set was considered here.

4.2 Experiments' Results

Two groups of experiments were performed. We firstly investigate the performance of our parallel GMMs for speaker identification. Then, the experiments of combination of parallel GMMs and traditional GMM were made to get the best combination performance by variance with the tradeoff τ.

In the experiments, the number of subspace N of parallel GMMs is set to 3. With the aim of overcoming the fundamental frequency estimation, we choose an overlap of 10Hz between the subspaces.

Speaker Identification Using Parallel GMMs. In order to find out if the parallel GMMs is robust under different number of speakers, we made experiments on some subsets of two corpus. And in baseline method, we use 32-mixture GMM as the classifier.

For YOHO corpus, one ENROLL session (24 sentences in total) was used for training and all 10 VERIFY sessions (40 sentences in total) were used for testing. We adopted first 30, first 50 and total 138 speakers subsets of YOHO in experiments. TABLE 1 shows the results.

Table 1. Speaker Identification on *YOHO*

Method	First30(%)	First50(%)	Total138(%)
GMM	94.1	93.0	85.9
Pal. GMMs	94.9	94.0	87.6

For SRMC corpus, our evaluations were carried out using first 50, first 100 and total 303 speakers of it. We select all sentences from PR session and PI sessions (more than 1 minute) for training. And the remaining 50 sentences of each speaker were all used for testing. The results of YOHO corpus are listed in Table 2.

Table 2. Speaker Identification on *SRMC*

Method	First50(%)	First100(%)	Total303(%)
GMM	85.0	81.1	73.2
Pal. GMMs	85.7	83.3	78.7

There are two important points to be noted. Firstly, parallel GMMs obtains higher identification rate than the traditional GMM in all cases. It is obvious that taking into account the correlation of MFCC and fundamental frequency will bring us improvements. Though the identification rate falls as the number of speakers increase, traditional GMM decreases more rapidly than parallel GMMs. We can find in Table 2 that, when identification rate descends from 85.0% to 73.2% in the traditional GMM, only 7.0% percent descends is found in parallel GMMs. The similar cases can be found in all other sets of our experiments.

Secondly, the performances in two corpus differ slightly. In YOHO corpus, performance gained by parallel GMMs increase stably in different sets. But things were different in SRMC corpus. This can be explained by the effects of different environments. YOHO corpus was constructed under ideal lab condition; however, SRMC corpus was recorded under daily communication condition, which is prone to be affected by environments. So parallel GMMs, utilizing the fundamental frequency information, would act better.

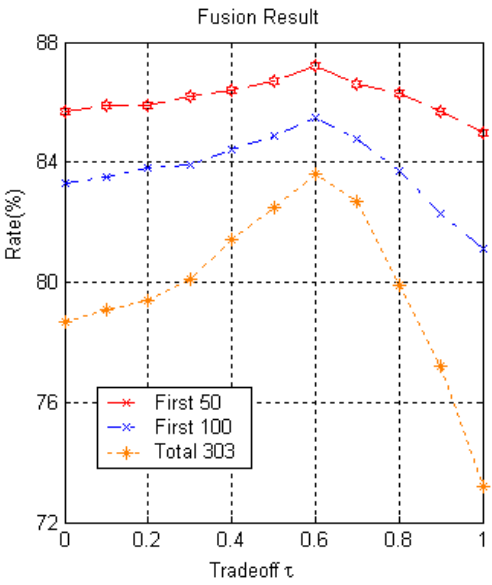


Fig. 2. Speaker identification rate with variance of the tradeoff τ

Combine Parallel GMMs with GMM. Performances of the combined model according to Equation. (10) with variance of the tradeoff τ in speaker identification task are demonstrated in Fig. 2. Experiments were done on SRMC corpus.

In fact, when $\tau = 0$ or $\tau = 1$, it degraded to parallel GMMs and traditional GMM. From Figure 2, we can find that the identification rates are ascending along with the increase of τ for all cases. And the best performances are all achieved when $\tau = 0.6$. The first 50 speakers' subset get 87.2%, 85.5% for first 100 speakers and 83.6% for all 303 speakers. Then, there is a drop for each case.

5 Conclusions

In this paper, we improved the joint probability function model, which takes into account the correlation between glottal source and vocal tract information, due to its flaws. And the parallel Gaussian Mixture Models was proposed in speaker recognition task for sake of precise modelling of glottal information and data consideration. Promising results of experiments on YOHO and SRMC corpus are achieved, especially when applying the combination of parallel GMMs and traditional GMM.

Acknowledgment

This work is supported by National Natural Science Foundation of P. R. China (No.60273059), Zhejiang Provincial Natural Science Foundation for Young Scientist of P. R. China (No.RC01058), Zhejiang Provincial Natural Science Foundation (No.M603229) and National Doctoral Subject Foundation of P. R. China (20020335025).

References

1. Atal B. S.: Automatic recogntion of speakers from their voices. Proc. IEEE. **64** (1976) 460-475
2. Kermal Sonmez, Elizabeth Shriberg, Larry Heck and Mitchel Weintraub: Modeling Dynamic Prosodic Variation for Speaker Verification. Proc. Intl. Conf. on Spoken Language Processing. **7** (1998) 3189-3192
3. H. Mizuno et al.: Pitch dependent phone modeling for HMM-based speech recognition. J. Acoust. Soc. Jpn(E). **15** (1994) 77-84
4. A. Adami, R.Mihaescu, D. Reynolds and J. Godfrey: Modeling Prosodic Dynamics for Speaker Recognition. IEEE ICASSP'03. **4** (2003) 788-791
5. Reynolds Douglas A.: The effects of handset variability on speaker recognition performance: Experiments on Switchboard corpus. IEEE ICASSP'96. **1** (1996) 113-116
6. X Shao, B. Milner and S. Cox.: Integrated Pitch and MFCC Extraction for Speech Reconstruction and Speech Recognition Applications. Eurospeech'03. (2003) 1725-1728
7. B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Reynolds, B. Xiang: Using Prosodic and Conversational Features for High-performance Speaker Recognition: Report from JHU WS'02. ICASSP'03. (2003) **4** 792-795
8. Arcienega M. and Drygajlo A.: Pitch-dependent GMMs for Text-Independent Speaker Recognition Systems. Eurospeech'01. Scandinavia (2001) 2821-2824
9. Hassan Ezzaidi, Jean Rouat and Douglas Shaughnessy: Towards combining pitch and MFCC for speaker identification systems. Proceedings of Eurospeech. (2001) 2825-2828
10. Campbell J. Jr.: Speaker Recognition: A Tutorial. Proceedings of the IEEE. **85** (1997) 1436-1462
11. B. A. Dautrich, L. R. Rabiner and T. B. Martin: On the effects of varying filter bank parameters on isolated word recognition. IEEE Trans. Acoust., Speech, Signal Processing. **31** (1983) 793-807

12. A. K. Jain and A. Ross: Learning User-specific Parameters in a Multibiometric System. Proc. Intl. Conf. on Image Processing. (2002) 57-60
13. Campbell, J. Jr.: Testing with the YOHO CD-ROM Voice Verification Corpus. ICASSP'95. (1995) 341-345
14. Lifeng Sang, Zhaohui Wu and Yingchun Yang: Speaker Recognition System in Multi-Channel Environment. IEEE International Conference on System, Man & Cybernetics. (2003) 3116-3121
15. X. Sun: A Pitch Determination Algorithm Based on Subharmonic-to-harmonic ratio. The 6th International Conferernce of Spoken Language Processing. Beijing China. **4** (2000) 676-679

Biometric Recognition Using Feature Selection and Combination

Ajay Kumar[1,2] and David Zhang[2]

[1] Department of Electrical Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India
ajaykr@ieee.org
[2] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong
csdzhang@comp.polyu.edu.hk

Abstract. Most of the prior work in biometric literature has only emphasized on the issue of feature extraction and classification. However, the critical issue of examining the usefulness of extracted biometric features has been largely ignored. Feature evaluation/selection helps to identify and remove much of the irrelevant and redundant features. The small dimension of feature set reduces the hypothesis space, which is critical for the success of online implementation in personal recognition. This paper focuses on the issue of feature subset selection and its effectiveness in a typical bimodal biometric system. The feature level fusion has not received adequate attention in the literature and therefore the performance improvement in feature level fusion using feature subset selection is also investigated. Our experimental results demonstrate that while majority of biometric features are useful in predicting the subjects identity, only a small subset of these features are necessary in practice for building an accurate model for identification. The comparison and combination of features extracted from hand images is evaluated on the diverse classification schemes; naive Bayes (normal, estimated, multinomial), decision trees (C4.5, LMT), k-NN, SVM, and FFN.

1 Introduction

Feature evaluation is critical while designing a biometric based recognition system under the framework of supervised learning. The existing research in biometrics has not made any attempt to evaluate the usefulness of the features that have been proposed in the literature [1]-[4]. Feature subset selection helps to identify and remove much of the irrelevant and redundant features. The small dimension of feature set reduces the hypothesis space, which is critical for the success of online implementation in personal recognition. Furthermore, researchers have shown [5]-[8] that the irrelevant and redundant training features adversely effects the classifier performance. Table 1 summarizes the effect of redundant training information on some common machine learning algorithms. These observations provide us the motivation to perform the experiments to evaluate the advantages of the feature subset selection and combination of some common biometric modalities.

Most of the prior work in the biometric fusion literature has examined the fusion of modalities at score and decision level. However it is generally believed that a fusion scheme applied as early as possible in the recognition system is more effective [9]. Therefore, the fusion at feature level typically results in a better improvement than at

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 813 822, 2005.
Springer-Verlag Berlin Heidelberg 2005

decision level. This is because the feature representation conveys the richest information as compared to the matching scores or abstract labels. Furthermore, the feature level fusion has not received adequate attention in the biometrics. Therefore we have also investigated the performance improvement using feature level fusion in the context of feature subset selection.

Table 1. Effect of redundant biometric features

Machine Learning Algorithm	*Effect of Irrelevant Features*	*References*	*Typical Example*
Nearest Neighbor	Training complexity grows exponentially	[5],[6]	[1],[2]
Naive Bayes	Invalidation of assumption that features are independent in a given class	[7]	[3]
Decision Trees	Overfitting of training data, Large tree complexity	[8]	[4]

2 Feature Evaluation and Selection

Feature selection is used to identify the useful features and remove the redundant information. The usage of small size feature vector results in reduced computational complexity which is critical for online personal recognition. The selection of effective features may also result in increased accuracy. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been traditionally used to reduce the large dimension of feature vectors. However, PCA or LDA *transforms* the feature vectors to reduced dimension rather than actually selecting a subset. Several feature subset selection algorithm have been proposed in the literature. The feature subset evaluation and selection algorithm consists of three basic modules as shown in Figure 1; feature subset-generation and evaluation and the stopping criteria. Let N be the total number of potential biometric features in the biometric training dataset. The exhaustive search through the total number of 2^N candidate feature subsets in infeasible even with moderate N. Therefore various search strategies (*e.g.* starting point, search direction, *etc*.) have been studied in the literature. The goodness of each of the candidate feature subset is evaluated with a feature evaluation criterion. The goodness index of current feature subset is compared against those from the previous best feature subset and replaced if this index from current feature subset is better than those from the previous best feature subset. There are two commonly used feature evaluation criterion; wrapper-based and filter-based. The wrapper is one of the most commonly used algorithms, which evaluates and selects feature subset by repeated use of a particular classification algorithm. However, it is highly time consuming and prohibitive when the dimension of feature vectors is large (such as those from palmprints evaluated in this work). Therefore we employed filter-based algorithm for the feature evaluation which is detailed in next section. The feature selection process usually stops with a suitable stopping criteria; *e.g.* predefined number of features, iterations or goodness index, addition or deletion of features does not increase goodness index, *etc*.. In this work we used Correlation based Feature Selection (CFS) algorithm which has been shown [10] to be quite effective in feature subset selection. The CFS is a classifier independent algorithm and its usefulness is illustrated from our experimental results.

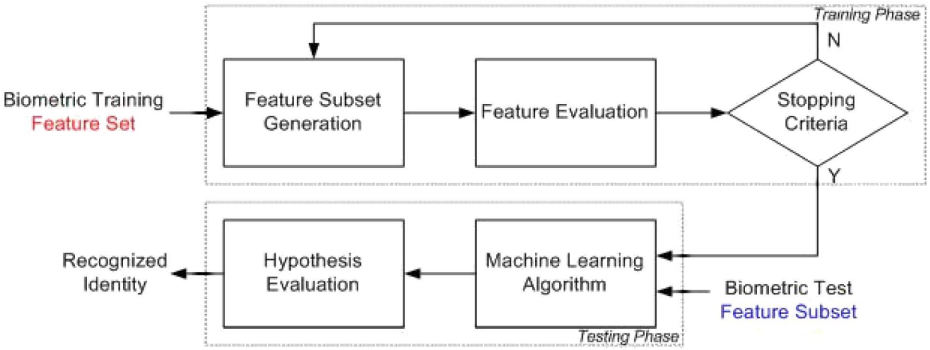


Fig. 1. Building a biometric recognition system using feature subset selection

3 Correlation Based Feature Selection Algorithm

The CFS algorithm uses a correlation based objective function to evaluate the usefulness of the features. The objective function $J_{cfs}(\lambda)$, also know as Pearson's correlation coefficient, is based on the heuristic that a good feature subset will have high correlation with the class label but will remain uncorrelated among themselves.

$$J_{cfs}(\lambda) = \frac{\lambda \psi_{cr}}{\sqrt{\lambda + \lambda(\lambda - 1)\psi_{rr}}} \tag{1}$$

Above equation illustrates the merit of λ features subset where ψ_{cr} is the average feature to class correlation and ψ_{rr} is average feature to feature correlation within the class. The CFS based feature selection algorithm uses $J_{cfs}(\lambda)$ to search the feature subsets using the best first search [11]. The best search starts with the evaluation of all the individual features considering them as a separate subset. The feature subset with highest objective function is retained. The new feature subset is further expanded by adding all possible combinations of new single features and all the resulting combinations are evaluated in the similar manner. If the addition of a new feature does not show any improvement then the search process returns to next unexpended subset and repeats in the similar manner. The search is aborted if the addition of new features does not show any improvement in the last 5 consecutive expanded combinations. This stopping criteria is the same (default) used in *MLC++* machine learning library [12] but for the wrapper feature selector. The expanded feature subset before the termination of search process is assumed to be the best feature subset. However there may be some locally predictive features in the unselected feature set which may be useful in some classification schemes [11]. Therefore the average correlation of every unselected feature with its corresponding class is also examined. If this correlation is higher than the highest average correlation between any of the already selected features, then this feature is included in the list of best feature subset. We examined the usefulness of CFS scheme by evaluating recognition accuracy and the size of best feature subset with those from original feature set.

4 Classification Schemes

The personal recognition will use the feature vectors from the training images to train or learn the classification algorithm. In this work, we investigated a number of classification algorithms to evaluate the benefits of feature subset selection. These algorithms are quite popular and well known in pattern recognition literature. However, with few notable exceptions, their usefulness for hand recognition is yet to be evaluated. The simplified version of Bayes rule, known as *naive Bayes*, which assumes that the feature vectors within a class are independent, was firstly evaluated. The *naive Bayes* has shown to work well with real data samples and it traditionally makes the assumption that the feature values are normally distributed. However, this assumption may be violated in some domains and our experiments were not restricted to the normality assumption. The distribution of features was also estimated using nonparametric kernel density estimation [14] and employed in the *naive Bayes* classifier. The *multinomial* model has been shown to outperform [15] other alternative models on the real data and was therefore also investigated for the performance.

The *k*-Nearest Neighbor (*k-NN*) classifier employed minimum Euclidean distance between the query feature vector and all the prototype training data. The Support Vector Machine (*SVM*) classifier employed polynomial kernel as it gave us the best results. The execution speed of multi-layer Feed-Forward Neural Network (*FFN*) is among the fastest of all models currently in use. Therefore this network may be the only practical choice for online personal recognition. A linear activation function was selected for the last layer of *FFN* while the sigmoid activation function was employed for other layers. The training weights were updated by using resilient backpropagation, which achieves faster convergence and conserves memory [19].

The decision tree algorithms use training data to build a logical tree and have been proved popular in practice. The *C4.5* [20] is the most used algorithm and uses entropy criteria to select the most informative features for the branching during the training stage. The feature that gives the most information is selected to be at the root of the tree. Another extension of *C4.5*, also known as logistic model tree (*LMT*), uses a combination of tree structure and logistic regression model to build the decision tree. The different logistic regression functions at tree leaves are built using *LogitBoost* algorithm [22]. The construction of *LMT* is detailed in [21] and was evaluated in the experiments as it achieved much higher accuracy than *C4.5*.

5 Experiments

In order to examine the goals of our experiments the biometric image database from 100 subjects was employed. The dataset consisted of 1000 hand images, 10 images per subject, which were obtained from digital camera using unconstrained peg-free setup in indoor environment. These hand images were collected during two sessions with an average interval of three months, as the focus of experiments was to investigate the performance of biometric modalities instead of their stability with time. The volunteers were in the age group of 16-55 years but not too cooperative and were not paid for the data collection. During the image acquisition, the users were only required to make sure that (i) their fingers do not touch each other and (ii) most of their hand (back side) touches the imaging table. The automated segmentation of handshape and palmprint image was achieved as detailed in reference [23].

5.1 Palmprint and Hand-Shape Features

Each of the 300 × 300 pixels segmented palmprint images were further divided into 24 × 24 pixel blocks with an overlapping of 6 pixels. The palmprint feature vector of dimension 1 × 144 is extracted from the standard deviation of significant discrete cosine coefficients in each of the image sub-blocks as detailed in [24]. The hand-shape features of dimension 1 × 23 features were also extracted from the hand-shape image; perimeter (f_1), solidity (f_2), extent (f_3), eccentricity (f_4), *x*-*y* position of centroid relative to shape boundary $(f_5 - f_6)$, convex area (f_7), 4 finger length $(f_8 - f_{11})$, 8 finger width $(f_{12} - f_{19})$, palm width (f_{20}), palm length (f_{21}), hand area (f_{22}), and hand length (f_{23}). Further details on these seven new shape features $(f_1 - f_7)$ can be seen in [24]-[25]. The signature analysis on the hand-shape boundary image is used to extract the image reference points, *i.e*. four fingertips, four inter-finger points and hand-base. We employed five image samples from every user collected during the first session for training and the rest for testing. In order to allow fair selection and combination of features, same training and testing splits are used to generate the results.

5.2 Classifier Parameters

The parameters of *SVM* and *FFN* employed in the experiments were empirically selected. The *SVM* using polynomial kernel gave much better results than those from radial basis function. Therefore to conserve the space only results from polynomial kernel are reported. The *SVM* training was achieved with *C-SVM*, a commonly used *SVM* classification algorithm [26]. The training parameter γ and ε were empirically fixed at 1 and 0.001 respectively. Similarly the number of input nodes in *FFN* were also empirically selected for the best performance; 100 (80) for palmprint, 50 (50) for hand-shape and 125 (75) for the combined feature set. The entries in the brackets represent the numbers when corresponding feature subset is employed for the performance evaluation. The *FFN* neuron weights were updated using resilient back-propagation algorithm and the training was aborted if the maximum number of training steps reached to 1000. The *C4.5* decision tree was pruned with a confidence factor of 0.25. The splitting criteria for *LMT* was the same as the one used for *C4.5*, *i.e*. information gain. The minimum number of feature vectors at which a node can be considered for splitting was fixed to 15.

6 Results

The experimental results for the palmprint recognition are summarized in Table 2. This table also shows the performance of corresponding classifier with and without the feature subset selection. The evaluation of 144 palmprint features from the training set, using the CFS algorithm described in section 3, has revealed 75 redundant and irrelevant features. This suggests that the feature selection has been aggressively pursued in the palmprint domain. The performance of 69 relevant palmprint features, or feature subset, is also illustrated in Table 2. It can be seen from this table that the

kernel density *estimation* has managed to improve *naive Bayes* performance but the performance improvement is significant when *multinomial* event model is employed. The best performance for palmprint recognition is achieved with *SVM* classifiers when the second order polynomial kernel is used. However, the achieved performance of nearest neighbor classifier suggest that it may be preferred in some applications as it is inherently simple and does not require training phase. The performance of *FFN* is better than *naive Bayes* but quite similar to that from *SVM* or *k-NN*. The performance of decision tree *C4.5* has been worst and this may be due to the large number of features that make the repeated portioning of data difficult. However the performance of *LMT* is also promising and similar to that of *k-NN*. The average tree size for the decision tree build using 144(69) features for *LMT* and *C4.5* was 16 (12) and 285 (281) respectively. This is not surprising as *LMT* algorithm has shown [21] to be often more accurate than *C4.5* and always resulting in a tree of small size than those from *C4.5*.

Table 2. Comparative performance evaluation for the palmprint recognition

	Naive Bayes			KNN	SVM			FFN	Decision Tree	
	Normal	*Estimated*	*Multinomial*		*d*=1	*d*=2	*d*=3		*C4.5*	*LMT*
144 Feature	69.4	74.4	91.8	94.4	95.2	95.8	95.8	95	50.6	94.6
69 features	68.8	75.4	92.8	95	94.4	95.6	95.5	94.8	50.2	93.8

Table 3. Comparative performance evaluation for the hand-shape recognition

	Naive Bayes			KNN	SVM			FFN	Decision Tree	
	Normal	*Estimated*	*Multinomial*		*d*=1	*d*=2	*d*=3		*C4.5*	*LMT*
23 Feature	73.9	79	29	84.6	89	88.4	88.6	86.4	67	89.6
15 Features	78.6	80	51.8	84.2	87	85.4	83.2	85	66.6	87.8

One of the important conclusions from the Table 2 is that the usage of feature selection has effectively reduced the number of features by 52.08% while improving or maintaining similar performance in most cases. *This suggest that while majority of palmprint features are useful in predicting the subjects identity, only a small subset of these features are necessary in practice for building an accurate model for identification*.

Table 3 summarizes the experimental results for the hand-shape identification. The evaluation of 23 hand-shape features from the training data has selected 15 most informative features; $f_1, f_7, f_8 - f_{11}, f_{14}, f_{16} - f_{19}, f_{20} - f_{23}$. The decision tree using *LMT* achieved the best performance while those from the *multinomial naive Bayes* is the worst. The usage of *multinomial* event model in *naive Bayes* has resulted in significant performance improvement from the palmprint features (Table 2) while those from hand-shape features has been degraded (Table 3). This can be attributed to the inappropriate estimation of the term probabilities resulting from the small size hand-shape feature vectors. The average size of decision tree build using 23 (15) features using *LMT* and *C4.5* was 81 (69) and 251 (255) respectively.

The experimental results for the combined hand-shape and palmprint features are shown in Table 4. The CFS algorithm selected 75 features subset from the combined list of 167 features. The combined feature subset had 13 hand-shape features,

Table 4. Comparative performance evaluation for the combined features

	Naive Bayes			KNN	SVM			FFN	Decision Tree	
	Normal	*Estimated*	*Multinomial*		*d*=1	*d*=2	*d*=3		*C4.5*	*LMT*
167 Feature	77	80	80.2	97.4	97.6	98	97.8	97.2	68.4	96
75 Features	78.2	78.6	94.2	97.8	97.8	97.8	98	96.8	68.2	96.4

i.e. $f_3, f_7, f_8, f_{10} - f_{14}, f_{17} - f_{18}, f_{20} - f_{23}$, and 62 palmprint features. It may be noted that the reduced feature subset obtained from the combined feature set is not the addition or sum of reduced feature subset individually obtained from palmprint and hand-shape feature set. *This suggests that only a certain combination of features, rather than the combination of individual feature subset carrying the discriminatory information, is useful in the feature level fusion*. The new hand-shape features selected in the individual and combined feature subsets, *i.e.* f_3, f_3, f_7, justify their usefulness. However, other new examined hand-shape features, *i.e.* f_2, f_4, f_5, f_6, could not establish their significance. As shown in Table 4, the *SVM* classifier achieved the best performance, which is closely followed by *k-NN*. It can be noticed that the combination of hand-shape and palmprint features has been useful in improving the performance for all the classifiers except for the case from *naive Bayes* classifier. The performances of combined features from the *multinomial naive Bayes* classifier using feature subset selection suggest that the *multinomial* event model is most sensitive to irrelevant and redundant features. The size of decision tree build using 167 (75) features using *LMT* and *C4.5* was 16 (12) and 285 (231) respectively.

It is prudent to examine how the performance of various classifiers that are adversely effected by the irrelevant and redundant features. The performance improvement of these classifiers with the availability of more features, using a fixed number of training samples, is investigated. In this set of experiments, all the available features from the training samples were ranked in the order of their merit using CFS objective function (1). The feature vectors in the test data set were also ranked in the same order of ranking generated from the training data. The performance of these classifiers starting from first 10 features was computed and the next 10 features were added at every successive iterations. The number of input nodes for *FFN* classifier was empirically fixed to 75, irrespective of number of features. Figure 2(a) shows the performance variation for *k-NN*, *SVM* and *FFN* classifiers with the increase in number of features. The *SVM* classifier does not show any appreciable increase in the performance with the addition of irrelevant features (say beyond 75) and its performance is generally the best of all the classifiers evaluated in this paper. *It is interesting to note that the feature selection strategy has been able to find 20 (10) best features that give* $\approx$ *96% (89%) accuracy using SVM classifier; This 20(10) feature subset consists of 15(6) palmprint and 5 (4) hand-shape features*.

The performance of *LMT* classifier in Figure 2(b) show an initial increase in performance with the increase in informative features but the performance stabilizes with the addition of non-informative and redundant features (beyond 70-75). Thus the performance of *LMT* suggests that it is insensitive to the redundant and irrelevant features and this is due to the fact that the *LMT* is built using the stagewise fitting process to construct the logistic regression models which select only relevant features

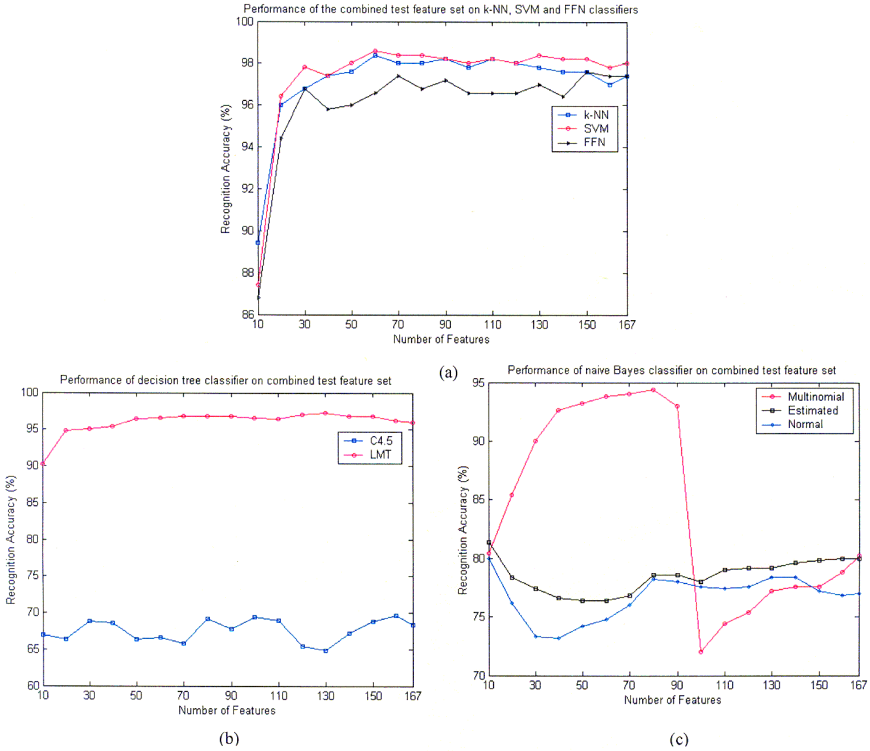


Fig. 2. The performance analysis of classifiers with the number of features; *k-NN*, *SVM* and *FFN* in (a), decision trees in (b), and naïve Bayes in (c)

from the training data. The *C4.5* decision tree continues to maintain worse performance and the feature selection strategy do not have any appreciable effect on the performance. Figure 2(c) shows the results for the performance of *naive Bayes* classifier. The performance estimates of *naive Bayes multinomial* classifier shows a tendency of exponential increase with small number of features before abrupt decrease in performance. The performance of *naive Bayes* with nonparametric kernel *estimation* is marginally better than those from with *normal* assumption but is quite poor.

7 Conclusions

The evaluation and selection of useful biometric features can improve the accuracy and reduce the complexity of classifier. This can also have high impact on user convenience and economics in the acquisition of online biometric data. It is not possible to locate the relevant features from the real biometrics data in advance and therefore the performance of feature selection strategy must be measured indirectly. The best way to do this is to compare the classifier performance with and without feature subset selection. The effectiveness of feature subset selection and combination was evaluated on the diverse classification schemes; probabilistic classifier (naive Bayes),

decision tree classifier (*C4.5, LMT*), and instance based classifier (*k*-NN) and learned classifiers (*SVM, FFN*).

Experimental studies in this paper further suggest that while a majority of features extracted from the hand images are useful in biometric recognition, only a small subset of these features are actually needed in practice for building an accurate model for subject recognition. This is important as the prior studies in biometric literature have not focused on the issue of feature subset selection. The usage of small size feature vectors results in reduced computational complexity, which is critical for online personal recognition. The analysis of experimental results in Table 2-4 indicate that the correlation-based feature subset selection is capable of effectively selecting the relevant palmprint and hand-shape features. We are currently working to examine the issue of feature subset selection for other biometric modalities.

References

1. C. Liu and H. Wechsler, "Independent component analysis of Gabor features for face recognition," IEEE Trans. Neural Networks, vol. 14, pp. 919-928, Jul. 2003.
2. D. Zhang, W.K. Kong, J. You, and M. Wong, "On-line palmprint identification," IEEE Trans. Patt. Anal. Machine Intell., vol. 25, pp. 1041-1050, Sep. 2003.
3. R. Sanchez-Reillo, C. Sanchez-Avila, A. Gonzales-Marcos, "Biometric identification through hand geometry measurements," IEEE Trans. Patt. Anal. Machine Intell., vol. 22, pp. 1168-1171, Oct. 2000.
4. A. Ross and A. K. Jain, "Information fusion in Biometrics", Pattern Recognition Lett., vol. 24, pp. 2115-2125, Sep. 2003.
5. P. Langley and S. Sage, "Scaling to domains with irrelevant features," Computational Learning Theory and Neural Learning Systems, R. Greiner (Editor), vol. 4, MIT Press 1994.
6. D. W. Aha, D. Kibler, and M. K. Albert, "Instance based learning algorithms," Machine Learning, vol. 6, pp. 37-66, 1991.
7. P. Langley and S. Sage, "Induction of selective Bayesian classifiers," Proc. 10th Intl. Conf. Uncertainty in Artificial Intelligence, Seattle, W. A., Morgan Kaufmann, 1994.
8. J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, Los Altos, California, 1993.
9. A.K. Jain, A. Ross and S. Prabhakar, " An Introduction to Biometric Recognition", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics, Vol. 14, No. 1, pp. 4-20, January 2004.
10. M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," Proc. 21st Australian Computer Science Conference, Springer-Verlag, pp. 181-191, 1998.
11. M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," Proc. 7th Intl. Conf. Machine Learning, Stanford University, CA. Morgan Kaufmann Publishers., 2000.
12. R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger, MLC++: A machine learning library in C++, available on http://www.sgi.com/tech/mlc/docs.html
13. C. Oden, A. Ercil, and B. Buke, "Combining implicit polynomials and geometric features for hand recognition," Pattern Recognition Letters, vol. 24, pp. 2145-2152, 2003.
14. G. H. John and P. Langley, "Estimating continious distribution in Bayesian classifiers," Proc. 11th Conf. on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, 1995.
15. S. Eyheramendy, D. Lewis, and D. Madigan, "On the naive Bayes model for text classification," to appear in Artificial Intelligence & Statistics, 2003.

16. A. McCallum and K. Nigam, "A comparison of event model for naive Bayes Text Classification," Proc. AAAI-98 Workshop on Learning for Text Categorization, 1998.
17. D. W. Aha, D. Kibler, and K. Albert, "Instance based learning algorithms," Machine Learning, vol. 6, pp. 37-66, 1991.
18. V. Vapnik, Statistical Learning Theory, Wiley & Sons, Inc., New York, 1998.
19. M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The PROP algorithm," Proc. Intl. Conf. Neural Networks, vol. 1, pp. 586-591, Apr. 1993
20. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, 1993.
21. N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," Proc 14th European Conf. Machine Learning, Cavtat-Dubrovnik, Croatia, Springer-Verlag, Vol. *2837,* pp 241-252, 2003.
22. J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting" The Annals of Statistic, 38(2), 337-374, 2000.
23. A. Kumar, D. C. M. Wong, H. Shen, and A. K. Jain, "Personal verification using palmprint and hand geometry biometric," Proc. AVBPA, pp. 668-675, June 2003.
24. Kumar and D. Zhang, "Integrating shape and texture for hand verification," Proc. ICIG 2004, pp. 326-329, Hong Kong, Dec. 2004.
25. John C. Russ, The Image Processing Handbook, 3rd ed., CRC Press, Boca Eaton, Florida, 1999.
26. N. Cristianini and J. S-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2001.

Evaluation of Biometric Identification in Open Systems

Michael Gibbons, Sungsoo Yoon, Sung-Hyuk Cha, and Charles Tappert

Computer Science Department, Pace University
861 Bedford Road, Pleasantville, New York, 10570 USA
mikegibb@us.ibm.com, {scha,ctappert}@pace.edu

Abstract. This paper concerns the generalizability of biometric identification results from small-sized closed systems to larger open systems. Many researchers have claimed high identification accuracies on closed system consisting of a few hundred or thousand members. Here, we consider what happens to these closed identification systems as they are opened to non-members. We claim that these systems do not generalize well as the non-member population increases. To support this claim, we present experimental results on writer and iris biometric databases using Support Vector Machine (SVM) and Nearest Neighbor (NN) classifiers. We find that system security (1-FAR) decreases rapidly for closed systems when they are tested in open-system mode as the number of non members tested increases. We also find that, although systems can be trained for greater closed-system security using SVM rather than NN classifiers, the NN classifiers are better for generalizing to open systems due to their superior capability of rejecting non-members.

1 Introduction

Biometric applications are becoming more common and acceptable in today's society. Technology continues to improve, providing faster processors, smaller sensors and cheaper materials, all of which are contributing to reliable, affordable biometric applications. The most common use of biometrics is for verification. In biometric verification systems, a user is identified by an ID or smart card and is verified by their biometric, i.e., a person's biological or behavior characteristic such as their fingerprint, voice, iris, or signature. This is analogous to a user at an ATM machine using a bank card to identify and a PIN to verify. Another use of biometrics is for identification, which is the focus of this paper. Identification can be applied in a closed system such as employee positive identification for building access, or in an open system such as a national ID system. Positive biometric identification, a 1-to-many problem, is more challenging than verification, a 1-to-1 problem. As stated in [1], "positive identification is perhaps the most ambitious use of biometrics technology".

There have been many promising results reported for closed identification systems. Although high accuracies have been reported in writer, iris and hand geometry studies [4, 10, 12, 15, 17], these accuracies may lead to a false impression of security. One may ask if there really are any situations that correspond to closed worlds[1]. For example, take an employee identification system. Can it be guaranteed that a biometric of a guest (a non-member) visiting the facility does not match one of an employee (a member)? The answer is no. This paper will investigate the generalizability of biometric identification as it pertains to the security of a system. Our hypothesis is that the accuracies reported for closed systems are relevant only to those systems and do not generalize well to larger, open systems containing non-members.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 823–831, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Since it is impractical to test a true population, we use a reverse approach to support the hypothesis. We will work with a database *M* of *m* members, but assume a closed system of m' members, where $m' < m$, and train the system on the m' members. We then have $m - m'$ members to test how well the system holds up when non-members attempt to enter the system. This approach is used on two biometric databases, one consisting of writer data and the other of iris data.

In section 2 of this paper, positive identification and the associated error rates will be explained. In section 3, the biometric databases and the pattern classification techniques used in this paper will be described. In section 4, the statistical experiments to support the hypothesis are described and observations presented. Section 5 concludes with a summary and considerations for future work.

2 Error Rate Types in Biometric Identification

Consider the positive identification model. Positive identification refers to determining that a given individual is in a member database [1]. This is an m-class classification problem – that is, given data from *m* subjects in a biometric database $M = \{s_1, s_2, \quad, s_m\}$, the problem is to identify an unknown biometric sample *d* from a person, s_q, where $s_q \in M$. In this model, a classifier can be trained based on all exemplars in *M* to find decision boundaries, e.g., support vector machine. If a similarity-based classifier such as a nearest neighbor is used, an unknown sample *d* is compared to each d_i of *M*. The error rate in this model is simply a number of misclassified instances divided by the testing set size. We claim that a classifier with the lowest error rate is not necessarily the best for security, and that classifier designers might consider the following three types of error rates.

Consider an unknown biometric sample *d* from a person, s_q, where $s_q \notin M$. If this biometric data of a non member enters directly to the above model, can it be classified correctly? If the classifier has no reject capability, the unknown will be classified into one of the decision regions or as the closest matching subject. However, if the classifier has a reject capability, the number of classes in *M* becomes $m+1$, i.e., *m* member classes + 1 reject class. Therefore, if the questioned instance is in a reject area in SVM or the closest match is outside the nearest neighbor thresholds, the unknown will be classified as none of the members. This study investigates the reject capability of two classifiers: support vector machine and nearest neighbor.

In the later scenario with members and non-members, there are three kinds of error. A 'false reject', FR, error occurs when a classifier identifies an unknown biometric sample *d* from a person, s_q, where $s_q \in M$, as a reject. The other errors are 'false accepts', FA, of which there are two types – those that can occur between members of the system, FA (1), and those that can occur as non-members enter the system, FA (2).

FA(1) occurs when a classifier identifies an unknown biometric sample *d* from a person, s_q to s_i where $s_q, s_i \in M$ and $s_q \neq s_i$. FA(2) occurs when a classifier identifies an unknown biometric sample *d* from a person, s_q to s_i, where $s_q \notin M$ and $s_i \in M$.

The frequencies at which the false accepts and false rejects occur are known as the False Accept Rate (FAR) and the False Reject Rate (FRR), respectively. These two error rates are used to determine the two key performance measurements of a biometric system: convenience and security [1]:

$$\begin{aligned} Convenience &= 1 - FRR \\ Security &= 1 - FAR \end{aligned} \tag{1}$$

In this paper, we will pay close attention to the Security measurement as we test our hypothesis.

3 Biometric Databases and Classifiers

Two biometric databases are used to support our claims in this study: the writer and iris biometric databases. Although there are many classifiers to choose from in the field of pattern classification, we used two pattern classification techniques: Support Vector Machines (SVM) and Nearest Neighbor.

3.1 Databases

In a previous study, Cha et al. [3] studied the individuality of handwriting using a database of handwriting samples from 841 subjects' representative of the United States population. Each subject copied a source document three times. Each document was digitized and features were extracted at the document, word, and character level. For the purposes of this study, we used the same database but focus only on the document features: entropy, threshold, number of black pixels, number of exterior contours, number of interior contours, slant, height, horizontal slope, vertical slope, negative slope, and positive slope. A detailed explanation of these features can be found in [4].

From the iris biometric image database [9], we selected 10 left bare eye samples of 52 subjects. In comparison to the writer database, the iris database has many fewer subjects, but a much larger number of samples per subject. This will allow for more samples to be trained.

After the images are acquired, they are segmented to provide a normalized rectangular sample of the iris. Features are extracted using 2-D multi-level wavelet transforms. For this experiment, 3 levels are used producing a total of 12 parts. The 12 parts produce 12 feature vectors consisting of the coefficients from the wavelet transform. The mean and variance of each vector are obtained to produce a total of 24 features for each sample. See [16] for more information on the 2-D wavelet transforms used.

3.2 Classifiers

In recent years, the SVM classifier has gained considerable popularity among the possible classifiers. The objective of the SVM classifier is to separate data with a maximal margin, which tends to result in a better generalization of the data. Generalization helps with the common classification problem of over-fitting.

The points that lie on the planes that separate the data are the support vectors. Finding the support vectors requires solving the following optimization problem (details of this method can be found in [2, 13]):

$$\min_{w,b,\xi} \tfrac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$
$$\text{subject to: } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i > 0 \tag{2}$$

The geometric representation of the SVM is easily visualized when the data falls into the linear separable and linear non-separable cases. However, real world data tends to fall into the non-linear separable case. For more information on the different cases, please refer to [11] which devotes a chapter to SVMs. To solve the non-linear separable problem, the SVM relies on pre-processing the data to represent patterns in a higher dimension than the original data set. The functions that provide the mapping to higher dimensions are known as phi functions or kernels. Common kernels include Radial Basis Function (RBF), linear, polynomial, and sigmoid. The RBF kernel is used in this study and additional information on this kernel follows in section 4.

The other classifier we consider is the Nearest Neighbor classifier, which computes distances from a test subject d to each member d_i of the database, and classifies the test subject as the subject that has the closest distance. The distances can be computed using various methods such city-block distance or Euclidean distance.

A reject threshold can be introduced into the Nearest Neighbor classification. If the distance between test subject d and its' nearest neighbor d_i is within the threshold, the classification is that of the closest member. However, if the distance is greater than the threshold, the subject is rejected and classified as a non-member. In this study, we used a reject threshold.

4 Statistical Experiments

Our hypothesis is that biometric identification on closed systems does not generalize well to larger, open systems containing non-members. In order to investigate this hypothesis, experiments were conducted on subset database $M' \subset M$ from both the writer and iris databases described in section 3.

4.1 Experiment Setup

For each of the databases, training sets were created. Training sets for the writer data consisted of $m' = 50$, 100, 200 and 400 members. Training sets for the iris data consisted of $m' = 5$, 15, 25 and 35 members. These sets included all instances per member, i.e., 3 per member for writer and 10 per member for iris.

For the first part of the experiment, an SVM was trained on the members. Parameter tuning, or SVM optimization, was performed prior to training. The first parameter tuned is the penalty parameter C from equation (2), and depending on the kernel used, there are additional parameters to tune. For this experiment we used an RBF kernel of the form:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \gamma > 0 \tag{3}$$

The γ parameter in equation (3) is the only kernel parameter requiring tuning. A grid-search method as defined in [8] was used to optimize these two parameters.

Tuning the parameters gives 100% accuracy on each of the training sets. Therefore, we have 0% FAR and FRR, or equivalently, 100% security and convenience. The next step is to test non-members to determine the true security of the trained SVM. For each training set we created a combined evaluation set consisting of the trained members plus an increasing number of non-members. The evaluation sets for the 50-writer trained SVM consisted of 50, 100, 200, 400, 700 and 841 subjects, where the first 50 subjects are the members and the remaining subjects are non-members. Similarly, the evaluation sets for the 25-iris trained SVM consisted of 25, 35, 45 and 52 subjects, where the first 25 subjects are the members and remaining subjects are non-members.

In the second part of the experiment, the Nearest Neighbor classifier was used. For this classifier, threshold tuning was required. The threshold has to be large enough to allow identification for the known members, but small enough not to allow non-members to be classified as members. As the threshold increases, the FAR increases and FRR decreases.

The Receiver Operating Characteristic (ROC) curve for the Nearest Neighbor classifier is presented in figure 1. The ROC curve is a plot of FAR against FRR for various thresholds. When designing an identification system, there is a trade off between the convenience (FRR) and security (FAR) of the system. For this experiment, we have chosen an operating threshold that is close to equal error rate, but leaning towards a higher security system.

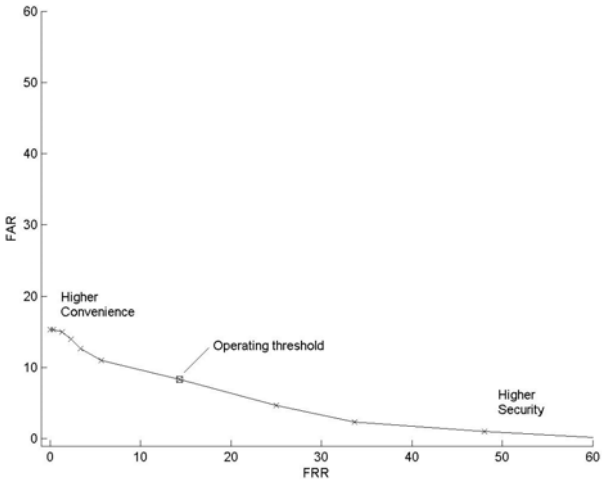


Fig. 1. The ROC curve for the Nearest Neighbor classification of 100 members. The operating threshold was chosen close to equal error rate, but favoring FAR or security

4.2 Results and Analysis

In the positive identification model we consider two errors: false accepts and false rejects. Since the SVM was able to train the members to 100% accuracy, we eliminate

the false accepts and false rejects for members. The remaining tests are non-members and therefore can only produce false accepts. The false accepts correlate to the security measurement of the system, a measure of extreme importance to the system. In figure 2, the security results are shown for the writer data.

As hypothesized, for each curve, as the number of non-members increases, the security monotonically decreases (or equivalently, the FAR monotonically increases). It might also be noted that the final rates to which the security curves decrease appear to converge – that is, to approach asymptotes. To ensure that this is not an artifact of the particular handwriting data used, we obtained similar experiment results on the iris data as presented in figure 3. The iris data in figure 3 follows the same pattern as the writer data in figure 2, although convergence is not as evident for these data.

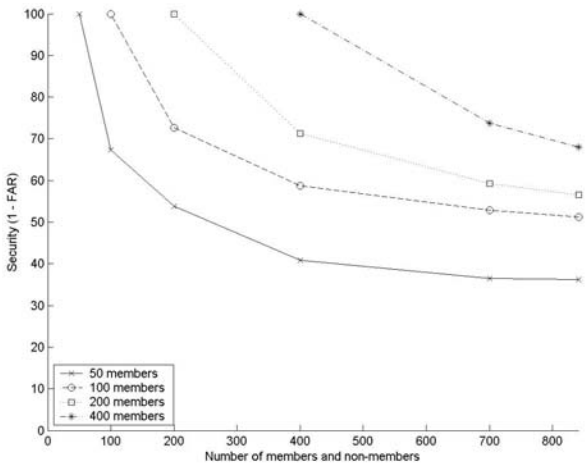


Fig. 2. Security results for writer data using SVM as non-members are introduced to the system

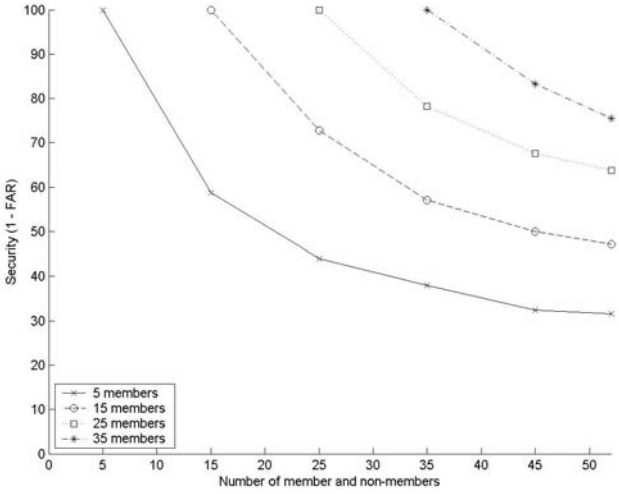


Fig. 3. Security results for iris data using SVM as non-members are introduced to the system

Next, we present the results for the Nearest Neighbor classifier. As can be seen in figure 4, the same pattern emerges, although in this experiment we did not obtain 0% FAR for the members. When using the Nearest Neighbor approach, a one-versus-all method was used to obtain the accuracy for the closed environment.

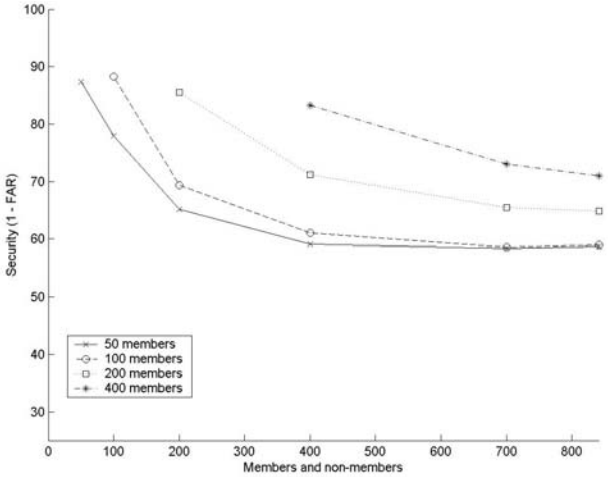


Fig. 4. Security results for writer data using Nearest Neighbor as non-members are introduced to the system

Last, we present a comparison of the results from the two classifiers used in this experiment. Figure 5 illustrates the security performance for 100 members of the writer database. Notice, although Nearest Neighbor does not perform as well on the closed environment, it eventually meets and surpasses the performance of the SVM as non-members enter the system.

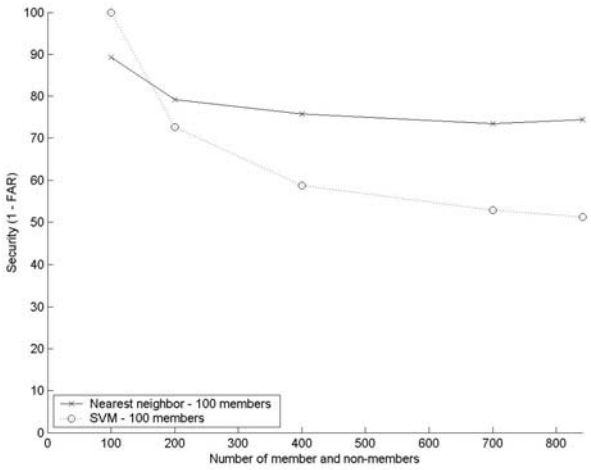


Fig. 5. A comparison of the performance of the Nearest Neighbor and SVM classifiers on the writer data consisting of 100 members

5 Conclusions

In this paper, we found that system security (1-FAR) decreases rapidly for closed systems when they are tested in open-system mode as the number of non members tested increases. Thus, the high accuracy rates often obtained for closed biometric identification problems do not appear to generalize well to the open system problem. This is important because we believe that no system can be guaranteed to remain closed. This hypothesis was validated by experiments on both writer and iris biometric databases.

We also found that, although systems can be trained for greater closed-system security using SVM rather than NN classifiers, the NN systems are better for generalizing to open systems through their capability of rejecting non members. Thus, it appears that the reject thresholds of NN classifiers do a better job of rejecting non members than the reject regions of SVM classifiers.

As commented by a reviewer of this paper, most complex biometrics systems use more complex classifiers. Given that performance on closed systems is not sufficient for applications where security is essential, we feel even the most complex classifiers should be tested in an open environment. For a more in depth analysis of various classifiers on open environments, refer to [6].

In summary, we demonstrated that the generalization capability of closed biometric systems in open environments is poor, and that the significantly larger error rates should be taken into account when designing biometric systems for positive identification. For increased security, for example, multi-modal biometrics might be considered [14].

5.1 Future Work

When designing an identification system, there is a trade off between the convenience and security of the system. Most systems would choose security over convenience. However, in our implementation of SVM for the writer data, we imply choosing convenience over security (guarantee 0 false rejects). In our study on writer data, there just are not enough samples per member to put into the testing set. It would therefore be beneficial to run further experiments against larger biometric databases.

We think that it may be advantageous to develop open identification systems by using a verification model. Also, it would be beneficial to explore the features for the given biometrics since security results could improve with features that better identify a member. Additional forms of biometrics, such as fingerprint, face, voice, hand geometry or a combination of biometrics might also be tested to further test the hypothesis.

References

1. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: Guide to Biometrics. Springer (2004)
2. Burges, C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. 2:121-167 (1998)

3. Cha, S.-H., Srihari, S.N.: Writer Identification: Statistical Analysis and Dichotomizer. Proceedings International Workshop on Structural and Syntactical Pattern Recognition (SSPR 2000), Alicante, Spain. pp. 123 – 132 (2000)
4. Cha, S.-H., Srihari, S.N.: Handwritten Document Image DatabaseConstruction and Retrieval System. Proceedings of SPIE Document Recognition and Retrieval VIII, Vol. 4307 San Jose (2001)
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons (2001)
6. Gibbons, M.: On Evaluating Open Biometric Identification Systems. Master's Thesis, CSIS Department, Pace University (2005)
7. Grother, P., Phillips, P.J.: Models of Large Population Recognition Performance. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04) Vol. 2. Washington, D.C. pp. 68-77 (2004)
8. Hsu, C.-W., Chang, C.-C., Len, C.-J.: A Practical Guide to Support Vector Classification.
9. Kee, G., Byun, Y., Lee, K., Lee, Y.: Improved Techniques for an Iris Recognition System with High Performance. Lecture Notes Artificial Intelligence (2001)
10. Krichen, E., Mellakh, M.A., Garcia-Salicetti, S., Dorizzi, B.: Iris Identification Using Wavelet Packets. Pattern Recognition, 17th International Conference on (ICPR'04) Vol. 4. pp. 335-338 (2004)
11. Kung, S.Y., Mak, M.W., Lin, S.H.: Biometric Authentication: A Machine Learning Approach. Pearson Education Company (2004)
12. Ma, Y., Pollick, F., Hewitt, W.T.: Using B-Spline Curves for Hand Recognition. Pattern Recognition, 17th International Conference on (ICPR'04) Vol. 3. Cambridge UK. pp. 274-277 (2004)
13. Osuna, E., Freund, R., Girosi, F.: Support Vector Machines: Training and Applications. MIT Artificial Intelligence Laboratory and Center for Biological and Computational Learning Department of Brain and Cognitive Sciences. A.I. Memo No 1602, C.B.C.L. Paper No 144 (1997)
14. Prabhakar, S., Pankanti, S., Jain, A.K.: Biometric Recognition: Security & Privacy Concerns. IEEE Security and Privacy Magazine. Vol. 1, No. 2. pp. 33-42 (2003)
15. Schlapbach, A., Bunke, H.: Off-line Handwriting Identification Using HMM Based Recognizers. Pattern Recognition, 17th International Conference on (ICPR'04) Vol. 2. Cambridge UK. pp. 654-658 (2004)
16. Woodford, B.J., Deng, D., Benwell, G.L.: A Wavelet-based Neuro-fuzzy System forData Mining Small Image Sets.
17. Zhang, D., Kong, W.-K., You, J., Wong, M.: Online Palmprint Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 25, No. 9. pp. 1041-1050 (2003)

Exploring Similarity Measures for Biometric Databases

Praveer Mansukhani and Venu Govindaraju

Center for Unified Biometrics and Sensors (CUBS), University at Buffalo
{pdm5,govind}@buffalo.edu

Abstract. Currently biometric system performance is evaluated in terms of its FAR and FRR. The accuracy expressed in such a manner depends on the characteristics of the dataset on which the system has been tested. Using different datasets for system evaluation makes a true comparison of such systems difficult, more so in cases where the systems are designed to work on different biometrics, such as fingerprint and signature. We propose a similarity metric, calculated for a pair of fingerprint templates, which will quantize the "confusion" of a fingerprint matcher in evaluating them. We show how such a metric, can be calculated for an entire biometric database, to give us the amount of difficulty a matcher has, when evaluating fingerprints from that database. This similarity metric can be calculated in linear time, making it suitable for large datasets.

1 Introduction

Biometric-based systems, in their various forms, have become popular over the last few years. Unlike various token or knowledge based systems, they cannot be easily compromised and are used in a variety of low or high security applications. Most systems in place use one biometric (such as facial image, fingerprint data, etc) to correctly identify or authenticate a candidate, though multi-biometric systems have also been developed in which a combination of two or more user characteristics can be used, either in serial or parallel.

The existing biometric based systems can be classified into one of two classes – identification based or authentication based systems. In identification based systems, a biometric template is provided to the system, which must then identify its owner based on the templates of various candidates stored in its database. This can be likened to a 1:N matching. The system outputs the predicted identity of the owner of the template, and in most cases a score, which indicates the confidence level of the match. A simpler form of matching is of 1:1 type, which is carried in authentication based systems, where the claimed identity of the user is given to the system as input, along with the template data. The system must only retrieve the corresponding stored template of the claimed identity, and compare it with the given template. The score outputted by the system can be compared with a set threshold, to authenticate the user.

Biometric based authentication / identification systems are designed to work efficiently and accurately for a large number of users, typically a few thousands or even in some cases, up to a million users. Such systems have to handle large databases, which may also store more than one template per user. This is especially true in the case of some behavioral based biometrics such as signatures, where two or more templates taken from a single user may significantly vary. Typically such systems require three to five templates per user. Even for a physical biometric such as fingerprint or

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 832–840, 2005.
© Springer-Verlag Berlin Heidelberg 2005

facial image, where there will be lesser variation across templates (provided enrolled within a short time interval of one another), multiple templates, usually two or three are taken. It becomes imperative that any operations carried out, over the entire database, must be scalable to efficiently run over such large data.

Section 2 describes the existence of measures for estimating the complexity of speech databases and word lexicons. We have also explored currently used methods for estimating the performance of biometric systems and quality of biometric databases. In section 3, we have defined a similarity measure for fingerprint databases, which can serve as an indicator for estimating the difficulty / 'confusion' of a recognizer on a particular database. Section 4 shows how, once the similarity measure has been first defined for a pair of fingerprint templates, and the values computed over pairs of templates in the database, they can be used calculate the similarity metric for the entire database. Section 5 describes the results of various experiments performed to support our theory.

2 Current Work

2.1 Existing Evaluation Measures

Measures to evaluate performance of speech and handwriting recognition systems, on the basis of complexity of the language models, have been researched extensively. Perplexity of a language model can be expressed as the average number of words that appear after a given word in that model. It has been used as a measure of complexity of the model. Nakagawa et al [5] have explored the relation between the size of the vocabulary of a model and the recognition rate of a speech recognition system. Futhermore, they have shown how, given prior knowledge of the performance of the recognizer and information about the perplexity of a language model, they can predict the recognition rate of that recognizer on that language model.

Marti et al [4][8] have used a HMM based handwritten text recognizer on different language models (simple, unigram and bigram) and shown that using a model with a smaller perplexity results in a better recognition rate.

A method to evaluate the difficulty of word recognition can be measured on the basis of the length of the words in the language model. (Grandidier et al [1]). It has been observed that a system is able to correctly recognize larger words better than smaller words. However, this measure cannot be used to predict the performance of the system.

Govindaraju et al [2] have defined 'lexicon density' as a measure of the distance between two handwritten words. It is an average measure of the total cost of edit operations required to convert one word in the lexicon to another. They have shown that a measure based on lexicon density is a better indication of the complexity of a word lexicon than measures such as lexicon size or edit distance.

2.2 Evaluation of Biometric System Performance

Evaluation of biometric based systems is done in terms of their False Accept Ratios (FARs) and False Reject Ratios (FRRs). [7] The FAR is the probability of a system evaluating two biometric templates of different people, to be of the same person.

Correspondingly the FRR can be defined as the probability of the system to recognize two templates of the same person to belong to different people. The accuracy of the system could also be expressed in terms of the Equal Error Rate (EER), which is the value of the FRR, at a particular threshold, when it is equal to the FAR. These evaluation measures do not take into account other factors which could affect the performance of the system, such as the similarity of biometric templates or the quality of the template data.

Most biometric systems are tested using databases, of varying sizes, having templates belonging to different users. This makes it impossible to compare the performance of a system, having results on a particular database, with the performance of another, which has been tested on a different database. Moreover, there is no method to compare the performance of systems that use different biometrics to identify / authenticate a user. These two systems will be using totally different types of databases, for example one system may use signature data and the other may use face images.

Bolle et al [3] have attempted to classify fingerprint databases based on the quality of images present. They generate a similarity score using a one-to-one matching between fingerprints of the same user and have used a statistic based on the mean and standard deviation of the scores to separate them into three classes: good, bad and ugly. However, visual inspection of the plot is needed to accurately separate the classes and classify the database. Moreover, this technique is primarily a measure of the quality of templates in the database, and does not give any indication of the difficulty of the recognizer in distinguishing one fingerprint image from another.

3 Calculating Similarity Measure for a Pair of Fingerprints

Currently used fingerprint based biometric systems store a user's fingerprint data not as an image but in the form of a template which is its compact form. The template is considered to be an accurate representation of the users biometric. Matching two fingerprints do not involve the actual comparison of the fingerprint images, but a matching procedure between the two corresponding templates. Typically fingerprint templates contain spatial information about the minutiae detected in the images, and are represented as a list of the extracted minutiae. Fingerprint images typically contain between 30 – 50 minutiae, although prints with up to 100 minutiae have also been observed. For a match between two fingerprints a match of 6-8 minutiae points are usually considered sufficient; however up to 12 matching points are needed for some particular law-enforcement applications. [10]

For calculating a similarity metric for a biometric database, we first define a similarity measure between two fingerprint images, and then average it out over all the images in the database. The similarity measure does not indicate whether the two templates belong to the same person, rather it is an indicator of the level of difficulty of a recognizer in comparing the two templates. Hence, though two templates which are very similar to one another will most likely belong to the same user, such may not always be the case. Similarly, two very dissimilar templates will more often than not, be of different persons, but not every time. A pair of templates having an extreme similarity value ie. either too high or too low should be correctly classified by the biometric system with ease, and pairs with intermediate values are going to take greater computation from a matcher to classify.

Consider a fingerprint template $M = \{m_1, m_2, \ldots\}$ where m_i is the i^{th} detected minutiae point. Each point m_i can be expressed as a triple (x_i, y_i, θ_i) denoting the (x,y) pixel location of the point and its orientation respectively. Here we have assumed each template to be extracted from a complete fingerprint image. The similarity measure between two fingerprint templates (M_a and M_b) can be calculated as follows

1. For each m_i in M_a calculate the corresponding m_j in M_b such that for all m_k in M_b, where $j \neq k$, $dist(m_i,m_j) < dist(m_i,m_k)$
2. Similarly, for each m_j in M_b calculate the corresponding m_i in M_a such that for all m_k in M_a, where $i \neq k$, $dist(m_j,m_i) < dist(m_j,m_k)$
3. Select $\{m^a_1,m^a_2,\ldots m^a_n\}$ from M_a such that for all m^a_x in $\{m^a_1,m^a_2,\ldots m^a_n\}$ exists corresponding m_j in M_b such that $(d^a_x = dist(m^a_x, m_j)) < dist(m_i, m_k)$ where m_i in M_a, but m_i not in $\{m^a_1,m^a_2,\ldots m^a_n\}$ and m_k in M_b such that m_k corresponds to m_i from step 1.
4. Similarly, select $\{m^b_1,m^b_2,\ldots m^b_n\}$ from M_b such that for all m^b_x in $\{m^b_1,m^b_2,\ldots m^b_n\}$ exists corresponding m_i in M_a such that $(d^b_x = dist(m^b_x, m_i)) < dist(m_j, m_k)$ where m_j in M_b, but m_j not in $\{m^b_1,m^b_2,\ldots m^b_n\}$ and m_k in M_a such that m_k corresponds to m_j from step 2.
5. Similarity $s(Ma, Mb) = (\sigma(d^a_1, d^a_2,\ldots d^a_n) + \sigma(d^b_1, d^b_2,\ldots d^b_n)) / 2$

It could be noted that a large value of the similarity score implies that the two templates are less similar to one another and thus less likely to confuse the recognizer. Conversely, a very small score implies that the two templates are highly similar, most probably belonging to the same person. Intermediate values indicate some possible confusion while matching. A perfectly similar template, which in practice is only going to occur when a template is matched against itself, will give us the similarity score 0.

4 Estimating the Similarity Metric for a Fingerprint Database

We can now estimate a similarity measure for a pair of fingerprints. This value indicates the difficulty of a fingerprint matching system in evaluating the two templates. A matcher finds it more difficult to identify the owner of a template from a dataset having a large number of similar prints. Conversely, if the dataset is such that the fingerprint templates stored are not very similar to one another, then the work of the matcher becomes easier. To quantize this level of difficulty, we define a similarity metric for a fingerprint database, calculated from the individual similarity measures for the stored templates, as follows:

$$S(\{M_1,M_2\ldots M_n\}) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s(Mi, Mj)}{n(n-1)/2}$$

Thus the similarity score for a database of fingerprints is defined as the average of scores calculated for all pairs of templates in the database.

4.1 Similarity Measures for Large Fingerprint Databases

We have established a metric for calculating the similarity of the templates in a fingerprint database. However, for a database of size n, calculation of this value will be

done in time $O(n^2)$. Though it can easily be calculated for smaller datasets, it does not scale well to databases with a large value of n. Most commercially available matching systems have been designed to run on databases having thousands or even a few million templates. Calculating the similarity measure for such a large database is not feasible, and certainly not efficient.

We propose a randomization method by which a similarity metric can be efficiently estimated for a database in linear time. Consider a dataset $\{M_1,M_2\ldots M_n\}$ having n fingerprint templates. The similarity measure can be calculated as follows:

1. Randomly select N fingerprint templates, where $N = \lfloor \sqrt{n} \rfloor$
2. For each of the N selected templates do:
   a. Randomly select N different fingerprint templates
   b. Calculate similarity measure between the template in (2) and each of the templates selected in (2a)
3. Calculate the mean of all such similarity measures as $S(\{M_1,M_2\ldots M_n\})$

On comparing the distribution functions of the similarity of all the pairs of fingerprint templates and the distribution function for the similarity calculated as above, using a reduced number of computations, it is observed that they are almost equal, and thus have the same mean. This implies that such a method of calculating similarity of a large biometric database, while it runs in significantly less time than the 'brute-force' method, it gives us the same output.

5 Results

5.1 Description of Fingerprint Datasets and Matcher

We have used 2 datasets for our experiments, both obtained from CUBS. Dataset 1 (D1) is a smaller dataset consisting of 20 users, each user providing us with 3 templates. Dataset 2 is a larger dataset having 50 users, again 3 templates per user. Each template stores the spatial information (x, y, θ) of each extracted minutiae point of the user.

The fingerprint matcher used is the standard matcher used at CUBS, and has comparable performance to any fingerprint matching software available commercially today. The output given by the matcher is a score between 0 and 1, where a 1 indicates a match between the input templates.

5.2 Analysis of Template Similarity Metric Based on Matcher Output

For each user, we have computed the similarity score for each pair of his templates (taking number of minutiae points compared (n) = 10). (Fig 1.) We have also given those templates to the CUBS fingerprint matcher to obtain a matching score. We have also computed similarity scores for all pairs of templates belonging to different users, and obtained the corresponding matching scores from the fingerprint matcher. Graphs of the similarity score of each template pair vs. the corresponding matcher output are plotted for both datasets.

Points closer to (0,1) represent those pairs of templates which have a high similarity and have been identified by the CUBS matcher as belonging to the same person with a high confidence ratio. This is because fingerprints which are very similar more often than not belong to the same person. Points which have a high x value, are those which have been identified as dissimilar pairs of prints, and have been classified by the matcher as belonging to different people. However most points have a similarity score between 2 to 8, and these pairs are those which would not have been classified by the fingerprint matcher so easily.

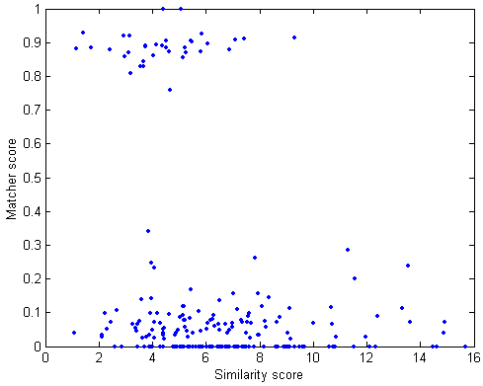


Fig. 1a. Similarity score vs. Matcher score for individual templates (Dataset 1)

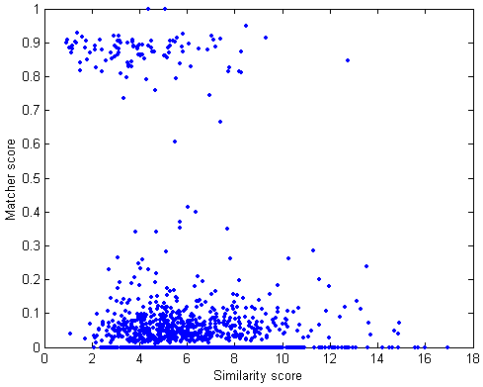


Fig. 1b. Similarity score vs. Matcher score for individual templates (Dataset 2)

5.3 Applying the Fingerprint Similarity Metric to Datasets

We have estimated the similarity metric for both our datasets D1 and D2, as can be seen in Table 1. Recall that a larger value of the similarity measure means that the two templates are more dissimilar to one another. We can see that the database D1 has a greater similarity measure than D2, which implies that, a recognizer will find it more difficult to distinguish between two random templates belonging to D1 than for two templates belonging to D2. However, it may be noted that as D2 has a greater

number of fingerprint templates, the size of the dataset will add to the confusion for the matcher. Thus, the similarity score, in conjunction with the size of the dataset (number of users) will be the most accurate way of representing the complexity of the dataset.

Table 1. Comparision of similarity metrics calculated for dataset D1 and D2

Database characteristic	Dataset D1	Dataset D2
Number of users	20	50
Total number of templates	60	150
Similarity measure	6.71	6.32

5.4 Scaling Similarity Metrics to Large Databases

We have estimated the similarity metric for two datasets (D1 and D2) having templates from 20 and 50 different users respectively, using a reduced number of computations as described in section 4.1. Table 2 shows that the values of the similarity metric, so estimated, are very close to the actual similarity metric calculated, using a 'brute force' method, involving all pairs of templates. However, this estimation requires a significantly less amount of time due to the reduced computations involved. A running time of O(n) indicates a large speedup over the previously required time of $O(n^2)$, making this method scalable to large databases.

Table 2. Comaparision of similarity metric calculations for datasets D1 and D2 using whoeldatasets and reduced data

Database characteristic	Database D1	Database D2
Number of users	20	50
Similarity measure calculated using whole database	6.71	6.32
Similarity measure on reduced data	6.58	6.43
Number of similarity comparisons for whole database	190	1225
Number of similarity comparisons using reduced data	~20	~50

Figures 2(a) and 2(b) show the similarity score distribution so calculated using the whole set of templates, and the reduced dataset, for datasets D1 and D2 respectively. It is observed that for a particular dataset, the similarity scores, whether calculated for all pairs of templates or a reduced set have the same distribution, and hence the same mean, which is finally used to estimate the similarity metric for the entire database.

6 Conclusions and Future Work

The similarity measure so computed for two templates gives an indication of the difficulty of the recognizer in evaluating them. When computed for the whole database, we can estimate the average "confusion" of the matcher in evaluating two templates from that database. This information, along with the size of the database, used in

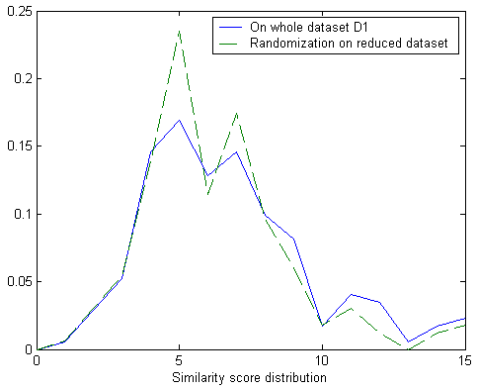


Fig. 2a. Plot of similarity score distribution on whole dataset and score distribution calculated by using a reduced number of computations – Dataset 1

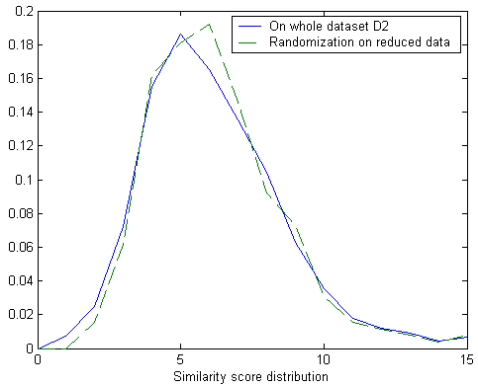


Fig. 2b. Plot of similarity score distribution on whole dataset and score distribution calculated by using a reduced number of computations – Dataset 2

conjunction with the FAR and FRR values, can give us a better idea of the performance of a fingerprint matcher. Moreover this method can be used to evaluate the complexity of fingerprint datasets generated by artificial means, such as a synthetic fingerprint database generator [9]. An accurate estimation of such a similarity metric can be done in linear time, so as to enable scalability across very large datasets.

We need to develop and evaluate similar metrics for various other biometric systems, such as signature verification systems and hand geometry based biometric systems, among others. Just as we can evaluate the complexity of fingerprint databases, we can do so for other biometrics also. This may also enable us to compare the performance of biometric systems which work on different biometrics. Similarity measures could also be computed for multi-biometric based systems, where a combination of two or more biometrics are used.

Another direction of interest to us, is to further reduce the number of computations required to estimate the similarity score of a fingerprint database. Currently we have proved that we can accurately estimate such a template in linear time. It may be pos-

sible to compute the same in a lesser polynomial time, or even in logarithmic time. It remains to be seen how much, if any, loss of accuracy in the computations is caused due to the reduction in running time.

References

1. F. Grandidier, R. Sabourin, A. E. Yacoubi, M. Gilloux, C.Y. Suen: Influence of Word Length on Handwriting Recognition. Proc. Fifth International Conference on Document Analysis and Recognition, Sept 1999
2. H. Xue, V. Govindaraju: On the dependence of Handwritten Word Recognisers on Lexicons. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, No 12, December 2002
3. Ruud M. Bolle, Sharat Pankanti, Nalini K. Ratha: Evaluation techniques for biometric-based authentication systems(FRR)
4. U.-V. Marti, H. Bunke: On the influence of Vocabulary Size and Language Models in Unconstrained Handwritten Text Recognition. IEEE Proceedings 0-7695-1263-1/01
5. Seiichi Nagakawa, Isao Murase: Relationship between Phoneme Recognition Rate, Perplexity and Sentence Recognition and Comparision of Language Models. IEEE Proceedings 0-7803-0532-9/92
6. Ruud M. Bolle, Nalini K. Ratha, Sharat Pankanti: Evaluating Authentication Systems Using Bootstrap Confidence Intervals
7. Rajiv Khanna, Weicheng Shen: Automated Fingerprint Identification System (AFIS) Benchmarking Using National Institute of Standards and Technology (NIST) Special Database 4. IEEE Proceedings 0-7803 – 1479-4/94
8. U.-V. Marti, H. Bunke: Unconstrained Handwriting Recognition: Language Models, Perplexity, and System Performance. L. R. B. Schomaker and L. G. Vuurpijl (Eds.), Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 2000, ISBN 90-76942-01-3
9. R. Cappelli, D. Maio, D. Maltoni: Synthetic Fingerprint-Database Generation. IEEE Proceedings 1051-4651/02
10. S. Pankanti, S. Prabhakar, A. K. Jain: On the Individuality of Fingerprints. IEEE Transactions on PAMI, Vol 24, No 8, 2002

Indexing Biometric Databases Using Pyramid Technique

Amit Mhatre, Sharat Chikkerur, and Venu Govindaraju

Center for Unified Biometrics and Sensors (CUBS),
University at Buffalo, New York, USA
{ajmhatre,ssc5,govind}@buffalo.edu
http://www.cubs.buffalo.edu

Abstract. Biometric identification has emerged as a reliable means of controlling access to both physical and virtual spaces. In spite of the rapid proliferation of large-scale databases, the research has thus far been focused only on accuracy within small databases. However, as the size of the database increases, not only does the response time deteriorate, but so does the accuracy of the system. Thus for larger applications it is essential to prune the database to a smaller fraction which would not only ensure higher speeds, but also aid in achieving higher accuracy. Unlike structured information such as text or numeric data that can be sorted, biometric data does not have a natural sorting order making indexing the biometric database a challenging problem. In this paper we show the efficacy of indexing hand geometry biometric using the Pyramid Technique, to reduce the search space to just 8.86% of the entire database, while maintaining a 0% FRR.

1 Introduction

In an increasingly digital world, reliable personal authentication is an important human computer interface activity. Biometrics such as fingerprints, face and voice verification are gaining industrial, government and citizen acceptance. The US-VISIT program uses biometric systems to enforce homeland and border security. Governments around the world are adopting biometric authentication to implement National ID and voter registration schemes [1]. FBI maintains national criminal and civilian biometric databases for law enforcement. In spite of the rapid proliferation of large-scale databases, the research community has thus far focused only on accuracy with small databases while neglecting the scalability and speed issues important to large database applications. If biometric systems have to truly replace the existing authentication systems, the response time, search and retrieval efficiency become important factors in addition to accuracy.

Traditional databases index the records in an alphabetical or numeric order for efficient retrieval. In biometric templates, there is no natural order by which one can sort the biometric records, making indexing a challenging problem. We propose guiding the search in biometric databases by using the Pyramid technique for indexing biometric databases.

2 Problem Definition

It has been shown that the number of false-positives in a biometric identification system grows geometrically with the size of the database [2]. If FRR and FAR indi-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 841 849, 2005.
Springer-Verlag Berlin Heidelberg 2005

cate the false accept and reject rates during verification, then FRR_N and FAR_N, the rates of false rejects and accepts in the identification mode are given by

$$FAR_N = 1 - (1 - FAR)^N \quad (1)$$

$$\approx N \times FAR \quad (2)$$

$$FRR_N = FRR \quad (3)$$

$$\text{No. of false accepts} = N \times (FAR_N) \approx N^2 \times FAR \quad (4)$$

There are two approaches in which we can attempt to reduce the error of such an identification system: (i) By reducing the FAR of the matching algorithm (ii) By reducing the search space (N) during identification. The FAR of a modality is limited by the recognition algorithm and cannot be reduced indefinitely. Thus, the accuracy and speed of a biometric identification system can be improved by reducing the number of records against which matching is performed. The effects of reducing the search space during identification are obtained by mathematical analysis. Assume that we are able to reduce the search space to a fraction (P_{SYS}) of the entire database. Then the resulting FAR, FRR values and total number of false accepts is given by

$$FAR_{PSYS\,N} = 1 - (1 - FAR)^{P_{SYS} \times N} \quad (5)$$

$$\approx (P_{SYS} \times N) \times FAR \quad (6)$$

$$FRR_{PSYSN} = FRR \quad (7)$$

To reduce the search space, we require a certain classification, partitioning or indexing of the database. There exist well-established procedures such as Henry classification system [3] to classify fingerprint records based on the ridge patterns such as “whorl”, “loop”, “arches” and “tented arch”. However, the problem with the existing Henry Classification system is that it is often found that the distribution of the fingerprints within each class is not uniform. It has been observed [4] that 2 of the classes constitute nearly 65% of the population. Also, binning a fingerprint into one of these classes is a non-trivial task with the best classification system having FRR of 5%. We would want to have a 0% FRR for an identification task, to prevent imposters sneaking through the system. Besides, apart from fingerprint, no such natural classification exists for the other biometrics.

Thus in presence of such a scenario, it is imperative that we go in for a sophisticated technique of partitioning or indexing the database to reduce the search space for not only improving the accuracy of the system, but also to ensure an efficient deployment of biometrics for real-time applications.

3 Previous Work

Classifying fingerprints into the Henry classes has been tried by Jain et al in [5], yielding a system with 12.4% FRR. A similar work by Ratha et al in [6] yielded a FRR of 10% with search space pruned to 25% of the original database. For a real-time application such as on an airport, with a rejection rate of 10%, 1 out of every 10 persons would be falsely rejected and would have to be manually inspected. In an experiment conducted by Cappelli et al [4] on NIST Special Database 4, it was shown that the distribution of Fingerprint population was non-uniform with 2 of the 5 Henry classes they considered holding nearly 65% of the population. Also, such natural classifications are not evident in other biometrics such as hand geometry.

The study of effect of binning was performed in [8] and it was demonstrated that the search space could be reduced to approximately 5% of the original database while keeping the FRR at 0% by using a parallel combination of hand geometry and signature biometrics. The data was clustered using the K-means clustering algorithm. The test template is then associated with the clusters it is closest to, with the new search space being the templates within these closest clusters. However, the binning approach has the shortcoming of having to re-partition the entire database on addition of new samples to it, which is an extremely time intensive process. Thus binning/clustering systems are useful only in cases of non-changing or static databases.

4 Indexing Techniques

Indexing the database implies a logical partitioning of the data space. In this approach of search-space reduction we make use of a Tree structure for organizing the data, such that each of the leaf nodes store one or more biometric templates. Thus given a test template, only the templates having similar index values would be considered as the search space, or in the case of range search, only the fraction of the database lying within the range of the indexes of the test template would be the new search space. The reduced search space could then be passed on for actual identification using a slower but a more reliable biometric such as fingerprint. With Biometric data being inherently multi-dimensional, it is indispensable that the indexing technique support multi-dimensional data. Besides supporting high-dimensional data, there are several additional requirements that the indexing technique must satisfy for lending itself to the biometric domain, such as the following:

1. Tree should be approximately balanced
   A balanced Tree will ensure that the time needed to reach to every leaf node is nearly the same among all the possible paths from the root to the leaf.
2. Tree should be dynamic
   The Tree structure used must be dynamic with respect to frequent insertions and not so frequent deletions. Since we refer to an online real-time application, such as an airport deployment, the number of new enrollments into the database can be expected to be ever increasing. The Tree should not lose its original properties of being approximately height balanced or being non-skewed, with these online operations of insertions and deletions.
3. Tree should support range queries
   Since the feature values measured for a biometric template is dependent on the data acquisition device and the feature extraction algorithm, we would want to provide a tolerance to each of the features measured, based on the intra-user variance observed for the feature. Thus in effect we would place the template into a bounding box, such that the extent along each dimension signifies the tolerance we provide for that dimension/feature. In such cases, the effective search space would be the candidates that lie within this hyper-rectangle. Indexing techniques for biometrics must hence be able to support range query searches.

4. Tree should not have overlapping intermediate nodes
   Overlapping intermediate nodes imply that while parsing through the tree to reach a certain leaf node, we might have to check different paths and thus search in several different leaf nodes.
5. Scalability
   Since we consider real-time applications, we can expect the database size to scale to millions of records. The indexing technique we use should be able to handle such a large amount of high-dimensional data.

Several methods such as KD Trees, R Trees and its variants, X Trees and the Pyramid technique were studied for this purpose.

KD Trees [9]. It is one of the most popular multi-dimensional indexing methods. KD Trees are a form of a Binary search tree, which is formed by a recursive sub-division of the universe by using a (d-1) dimensional hyper-plane at each node, where d is the dimension of the data.

However, KD Trees does not lend to our biometric indexing problem since the tree structure depends heavily upon the order of insertion of data. Its structure is not consistent and may result into a skewed structure with a certain order of insertion. Besides, it is also not a very dynamic structure, since frequent insertions may skew the tree. Deletions also pose a non-trivial problem of reorganization of the entire tree.

R Trees [10]. It makes use of the concept of bounding hyper-rectangles, where each leaf node is a bounding rectangle and encloses all the child nodes it contains within it. It is similar to a b-tree by keeping each of its non-root nodes at least half-full. It results in a height-balanced tree in which all the leaf nodes are at the same level.

The R Trees present us with obvious problem of having overlapping intermediate nodes, which poses the problem of having to search through multiple paths. It has been proved experimentally that with increasing dimensionality the problem of overlap only worsens [14]. Thus with our 24 dimension hand geometry data, using R Trees for indexing is not a viable option.

R* Trees [11]. R* Trees are similar to R Trees, however they try to go beyond the R-Tree heuristic of trying to minimize the area of each rectangle, by optimizing the area of overlap and consider other factors such as the margin of each rectangle. However, since the overlaps in intermediate nodes still exists the problem of following multiple paths would be significant in the high dimensional biometric data.

R+ Trees [12]. R+ Tree is similar in structure to the R Trees, however they completely avoid the overlaps in intermediate nodes. However they result in nodes with poor space utilization and longer tree structures. Also, R Tree and all of its variants are insertion order dependent structures. Thus the use of R+ Trees for indexing biometric databases is rejected too.

X Trees [13]. They are based on the R* Trees. They circumvent the problem of overlapping intermediate nodes by introducing "super nodes" having greater capacity than the other normal nodes. Besides being insertion-order dependent, they have a problem faced by most high-dimensional indexing techniques in terms of trying to use the 50% quantile when splitting a page to fulfill storage requirements. Such techniques result in pages that have an access probability close to 100% as pointed out by [14].

Pyramid Technique [14]. This currently seems to be the single most feasible indexing technique for indexing biometric databases. The concept is based on spatial-hashing the high-dimensional data into a single value. This single value can then be indexed very effectively using the popular B+ Trees.

The pyramid-technique was especially designed for working in higher-dimensional spaces and has been successfully tested on a 100-dimensional data set in [14]. The technique results in a tree structure that is not only insertion order invariant, but also height balanced and dynamic to the frequent insertions and deletions.

5 Methodology

The Pyramid technique requires normalizing the data values to lie between 0 and 1 to further divide the d dimensional data space into 2×d number of pyramids, each having a common tip-point and a (d-1) dimensional base. For instance, for the 2-dimensional case, we have 2×2 = 4 pyramids, as shown in the figure-1 below, with each of those pyramids (p0. .p3) having a common tip-point and a 1-dimensional base. Since we work with 24-dimensional hand geometry data, the number of pyramids formed are 2×24 = 48. To enroll a person in the database, we obtain 5 training samples of the hand geometry template. We average out the features in the 5 training templates and index only the mean-template of each person.

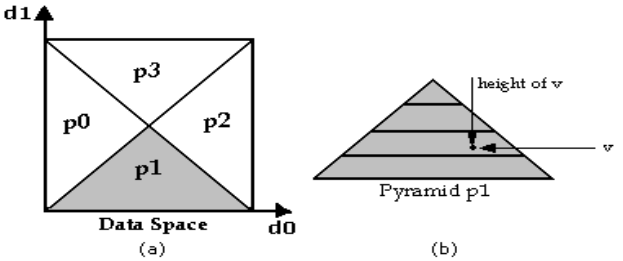


Fig. 1. a. Partitioning 2-d space into 4 Pyramids b. Height of point v in pyramid P1 [14]

The pyramids are numbered in a logical manner, such that a pyramid is numbered i if all the points within the pyramid are farthest from the tip-point along dimension i than any other dimension. Further, we check if the points within the pyramid have their i^{th} coordinate less than or equal to 0.5, in which case, the pyramid is labeled as i, else it is labeled as (i+d), d being the dimensionality of data. For instance, in the figure-1 shown above, all points within pyramids p1 and p3 are farthest from the tip-point along dimension d1 than along dimension d0 and points in p3 have d1 values greater than 0.5.

The height of a point inside pyramid i is defined as the distance from the tip-point in the (i mod d) dimension. For instance, all points lying within pyramids p1 and p3 above, have the height defined as the distance from the tip-point along dimension d1.

The (height + pyramid number) forms the key (pyramid value) for every template and is used as the indexing key in the B+ Tree. On receiving an input template for the identification task, we bound this template by a bounding box to allow a level of

tolerance for the feature values of the template by determining the lower and upper bounds for each feature as follows:

$\text{Lowerbound}_i = F_i - tol \times avgstd_i$, $\text{Upperbound}_i = F_i + tol \times avgstd_i$

F_i: Value of feature-i of the test template

$avgstd_i$: average intra-user standard deviation for feature i

tol: Tolerance-scale factor, $tol \times avgstd_i$ determines the tolerance for each feature.

On performing similar operation on each of the dimensions, we obtain a hyper-rectangle for the test template. To find the candidate templates, we first determine the pyramids that this bounding box intersects. For every intersecting pyramid, we obtain the templates that lie within the range of the bounding box within the pyramid, by performing a range query using the method given in [14]. These templates form our candidate set for 1:1matching to be performed by the final stage using another more reliable biometric such as fingerprint.

6 Hand Geometry Features

The algorithm to extract the feature involves the following steps: (1) Image acquisition (2) Converting to binary image (3) Contour extraction and noise elimination and (4) Feature extraction. The various geometric features that we are interested in involve the lengths and widths of the various parts of the fingers, the area under each finger. All together we have identified 24 invariant features [15] of the hand as shown below

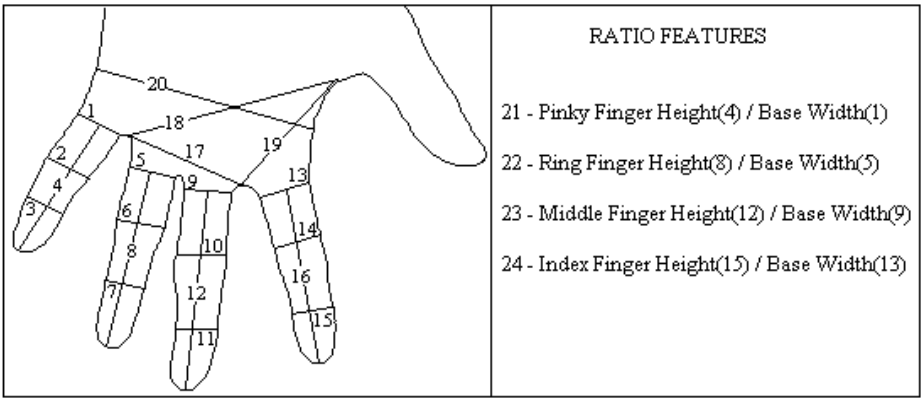


Fig. 2. Features used for Hand Geometry template

7 Analysis

A theoretical analysis of the number of comparisons done to find all candidate templates lying within the bounding box of a given test sample yields the following:

$$\text{Number of Comparisons} = \sum_{i=1}^{D} (2 \times \log_2(N/m)) + \Theta(2 \times D)$$
$$\approx O(\log_2(N/m))$$

where,

D:	dimensionality of the data, 35 in our case
N:	total number of templates in the database
m:	minimum number of templates to be held per intermediate node
$\Theta(2 \times D)$:	comparisons needed to determine the intersecting pyramids
$\log_2(N/m)$:	the height of the B+ Tree
$2 \times \log_2(N/m)$:	number of comparisons to get the leaf nodes representing the lower and upper bounds

Once we have obtained the leaf nodes corresponding to the bounds, we simply have to do a linear scan through the chain of the leaf nodes of the B+ Tree to get the final candidate set for further investigation. Thus we show theoretically that the comparisons needed to determine the candidate set for the final stage, that is, the total time for pruning, is logarithmic in the total number of samples N in the database. Now we experimentally show that the size of the pruned candidate set determined by the indexing system is much smaller than the original size of the database.

8 Experimental Results

We evaluated the algorithm based on the fraction of the database (P_{SYS}) that has to be searched in order to find the given user's record. The percentage was determined by running the algorithm on 5 test cases for each of the 200 users, thus a total of 1000 test cases. The fraction of the database searched P_{SYS} is the average size of the candidate set determined for the 1000 test templates. P_{SYS} can be described by the following equation

$$P_{SYS} = \frac{\sum_{T=1}^{TotalTestCases} \sum_{P=1}^{2 \times d} P_h}{TotalTestCases}$$

where,

P: Pyramid, which varies from 1 to 2×d, d = dimension of data

P_h: Number of points in Pyramid P that lie within the bounding box of template T.

Experiments showed that after indexing the hand geometry data, the penetration rate that it yielded was 8.86% while maintaining 0% FRR, with the *tol*, Tolerance-scale factor, equal to 3.4. Thus the search space for the identification task was reduced to 8.86% of the original size of the database without falsely rejecting even a single template. The penetration rate could be further reduced however at the expense of an increased FRR by reducing the tolerance we provide for each feature.

The effectiveness of indexing can be especially seen from the figures shown below, where the 3(a) refers to the graph of FRR versus Penetration Rate for an Index-

ing system and 3(b) refers to the Binning system described in section 3. The graph for the indexing system was obtained by varying the tolerance-scale factor, *tol*. The graph for the binning system was obtained by varying the number of clusters in which the database was split and averaging the FRR values for same penetration rates.

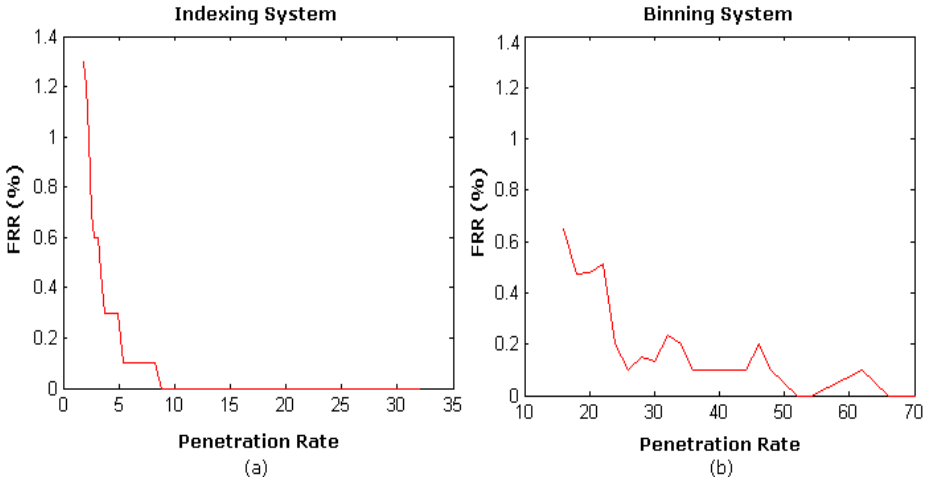


Fig. 3. Comparison of FRR v/s Penetration Rate for Indexing and Binning Systems

The FRR monotonically decreases with increase of the penetration rate in the Indexing system. However, the FRR decreases non-monotonically in the case of a binning system, which implies that there might be a case where even after increasing the penetration, the FRR might worsen instead of improving. Also, as can be seen, the indexing system converges to 0% FRR with much lesser penetration.

9 Conclusion and Future Work

We have presented a framework for implementing indexing in biometric databases in our endeavor to reduce the search space for the identification tasks. We have showed that on using the Pyramid technique for indexing biometric databases, we can prune the database to 8.86% of its original size maintaining a 0% FRR. Our future work involves using a parallel combination of hand geometry and signature indexed systems, which will definitely reduce the search space further. We also intend to evaluate the proposed scheme on a very large data set, by generating synthetic data.

References

1. Ruud Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. Guide to Biometrics, Springer Professional Computing, ISBN: 0-387-40089-3, 2004.
2. D. Maio, D. Maltoni, A. K. Jain and S. Prabhakar. Handbook of Fingerprint Recognition, Springer Verlag, ISBN: 0-387-95431-7, 2003.
3. International Biometric Group: http://www.biometricgroup.com/.

4. Cappelli R., Maio D., Maltoni D., Nanni L. *A two-stage fingerprint classification system*. ACM SIGMM workshop on Biometrics methods and applications, 2003
5. Anil Jain, Sharath Pankanti. *Fingerprint Classification and Matching*. Handbook for Image and Video Processing, A. Bovik (ed.), Academic Press, April 2000
6. Ratha N., Karu K., Chen S., Jain A.. *A Real-time Matching System for Large Fingerprint Databases*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No. 8, August 1996.
7. Searching in high dimensional spaces. Bohm, Berchtold, Keim. ACM Computing Survey, Volume 33, September 2001
8. Mhatre A., Palla S., Chikkerur S. and Govindaraju V. *Efficient Search and Retrieval in Biometric Databases*. SPIE Defense and Security Symposium, 2004.
9. Multidimensional Binary Search Trees Used for Associative Searching. Communications of the ACM, September 1975, Volume 18, Number 9.
10. R-Trees: A dynamic index structure for spatial searching, ACM SIGMOD international conference on Management of data, 1984
11. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. Beckmann, Kriegel, Schneider, Seeger. ACM SIGMOD, Volume 19, (June 1990)
12. The R+ Tree: A dynamic index for multidimensional objects. *Timos Sellis, Nick Roussopoulos* and *Christos Faloutsos*. VLDB 1987: 507-518
13. The X-tree: An Index Structure for High-Dimensional Data. Berchtold S., Keim D., Kriegel H., Proceedings ot the 22nd VLDB ConferenceMumbai, India, 1996
14. The Pyramid-Technique: Towards Breaking the Curse of Dimensionality. Berchtold, Böhm, Kriegel Proc. Int. Conf. on Manage ment of Data, ACM SIGMOD, Seattle, Washington, 1998
15. Pavan Rudravaram. Prototype pegless hand geometry verification. Technical report, Center for Unified Biometrics and Sensors, University at Buffalo, 2004

Classification Enhancement via Biometric Pattern Perturbation

Terry Riopka[1] and Terrance Boult[2]

[1] Lehigh University, Dept. of Computer Science and Engineering
Bethlehem, PA 18015 USA
riopka@kantbelievemyeyes.com
[2] University of Colorado at Colorado Springs, Computer Science Dept.
Colorado Springs, CO 80933 USA
tboult@cs.uccs.edu

Abstract. This paper presents a novel technique for improving face recognition performance by predicting system failure, and, if necessary, perturbing eye coordinate inputs and repredicting failure as a means of selecting the optimal perturbation for correct classification. This relies on a method that can accurately identify patterns that can lead to more accurate classification, without modifying the classification algorithm itself. To this end, a neural network is used to learn 'good' and 'bad' wavelet transforms of similarity score distributions from an analysis of the gallery. In production, face images with a high likelihood of having been incorrectly matched are reprocessed using perturbed eye coordinate inputs, and the best results used to "correct" the initial results. The overall approach suggest a more general approach involving the use of input perturbations for increasing classifier performance in general. Results for both commercial and research face-based biometrics are presented using both simulated and real data. The statistically significant results show the strong potential for this to improve system performance, especially with uncooperative subjects.

1 Introduction

Face detection is a critical preprocessing step for all face recognition systems. Its ultimate purpose is to localize and extract the face region of an image (which may or may not contain one or more faces) and to prepare it for the recognition stage of a face processing engine. In general, as a face preprocessor, it must achieve this task regardless of illumination, orientation or size of the input face image. As daunting as this task is for computers, it is a task that humans appear to do rather effortlessly.

Face detection approaches can be broadly organized into two categories: feature-based approaches [1], and image-based approaches [2]. The former relies primarily on the extraction of low level features incorporating face knowledge explicitly, while the latter treats the face as a pattern that can be learned from the two-dimensional image array, incorporating face knowledge implicitly. However, regardless of the approach, the result of face detection must enable some method for face registration, in order to maximize the effectiveness of the recognition stage of the face processor. In all cases, this relies on the accurate determination of fiducial marks on the face, ultimately needed for scaling and normalization.

Symmetry of the eyes and their consistent relationship with respect to other fiducial marks on faces make them extremely useful for parameterizing and normalizing

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 850–859, 2005.
© Springer-Verlag Berlin Heidelberg 2005

geometric features of the face. Because eye separation does not change significantly with facial expression, nor with up and down movements of the face, eye separation distance is often used for face normalization. Nose distance, another feature often extracted, is relatively constant with respect to side to side movements of the face and also depends on accurate eye localization. In addition, orientation of the line between the eyes is often used to correct for pose variations. Lastly, eyes are essentially unaffected by other facial features like beards and mustaches, making them invaluable features to most face recognition systems. As a result, eye localization is often the critical thread connecting face detection and face recognition algorithms, regardless of the underlying details of either algorithm.

Previous studies have emphasized the critical importance of eye localization and have demonstrated the dramatic effect poor eye localization can have on face recognition [3][4]. Given that the accuracy of eye localization has an effect on face recognition performance, this paper seeks to address the following research question: **Can we observe the effect that input eye perturbations have on an arbitrary recognition algorithm for a given face gallery, and use that information to improve classification performance?** The goal of this paper is to predict classification failure and, in instances in which it is expected to occur, use a failure prediction module to select an alternative eye location (perturbation) that has the greatest chance of yielding a correct classification, thus improving overall system performance.

The paper is organized as follows. A description of the method used to identify candidate face images for eye input perturbation is presented. Next, statistical results of simulated experiments explore the costs/benefits of our technique. The technique is also applied to a set of "real-world" face images to show the utility of the approach. Finally, we conclude with a discussion of the results and comment on the viability of a general approach to improving pattern classification using perturbations of critical input data.

2 Failure in the Context of Face Recognition

All face classifiers ultimately yield some sort of similarity score for an input image against all images in the face gallery. Typically, the scores are ranked to determine the most likely set of matching face images. The definition of "failure" in the context of face recognition typically depends on the application. For example, in identity verification, a serious failure occurs whenever a face not in the database is matched by the system, *i.e.* there is a false positive. In this case, the input face image matches an image in the database with a similarity score that is above a certain threshold. The decision of the system is based entirely on a comparison between two images, to determine whether the person is who the person claims to be.

In identification, the application of interest in this paper, a known or unknown individual is matched against all of the face images in the database, and a set of ranked potential matches is returned. In this case, the definition of failure is more complex. If the person is in the database, failure occurs if too many face images different from that person are ranked higher than the face image of that person in the database. Here, "too many" depends on the criteria of the system and how the results are interpreted. If the person is not in the database, it becomes problematical to determine whether or not the face is in the database based on ranking alone.

We postulate that the relationship between the similarity scores of the matched images (more specifically, the shape of their distribution) contains valuable information that can yield insight into the likelihood that a given match will lead to a correct classification. For example, intuitively, if all top ranked images have very close similarity scores, we might tend to believe there is a low probability that the top ranked image is the correct match. On the other hand, if the top ranked image has a similarity score that is significantly higher than all of the rest, we might tend to believe there is a high probability that the top ranked image is the correct match. In the former case, the distribution of sorted scores may be broad and flat, while in the latter case, narrow and peaked. Note that the criteria for "closeness" of similarity scores also depends on the characteristics of the particular recognition algorithm, since (usually) similarity score is not a metric.

In this paper, we use a machine learning approach to learn the characteristics of "good"' and "bad" similarity score distributions, given a particular recognition algorithm, a specific gallery of images, and various degrees of eye location error. "Good" similarity score distributions are those that result in a correct ID match (rank 1), where each individual (regardless of the number of images in the gallery) has a unique ID. "Bad" distributions are all others.

We make the general assumption that input eye locations are primarily responsible for classification failure as supported by [3]. Using our failure prediction model, we identify images that are likely to be classified incorrectly and then re-process those images using a limited set of perturbed input eye coordinates to yield new similarity score distributions. For each such image, the distribution most likely to yield a correct classification is identified and used to obtain a modified classification.

3 Face Algorithms

Two different face recognition algorithms were used in all of the following experiments: Elastic Bunch Graph Matching (EBGM)[5] and FaceIt, a commercial application based on an LFA algorithm [6]. The EBGM algorithm was provided by the Colorado State University (CSU) Face Identification Evaluation System (Version 5.0) [7]. FaceIt was implemented using programs built from a software development kit licensed from Identix Inc. The reader is referred to the relevant publications for details.

4 Learning Similarity Score Distributions

In order to learn similarity score distributions, a sample of "good" and "bad" similarity score distributions was required. If the intent were to learn "good" and "bad" similarity score distributions for face images *in general*, one might be inclined to train on similarity score distributions from a large set of "real" images of individuals in a given gallery. From an operational perspective and excluding synthetically altered gallery images, this would require considerable data collection and ground truth. However, the very specific intent here is to predict the behavior of a given algorithm on a given gallery with respect to input eye perturbations and to enable the recognition of potential instances where incorrect eye localization can result in misclassification. Generating the perturbation data is quite straightforward. Given some basic

training/testing sets, one simply forces the eye locations to different positions and reprocesses the images.

As was shown in previous research, the behavior with respect to input eye perturbations of a number of face recognition algorithms on degraded images, seems to be quite similar to their behavior on clean, gallery images [3], only slightly smoother. Consequently, the training set in this instance involved only the similarity score distributions obtained by perturbing input eye coordinates of gallery images. The prediction module therefore learns the sensitivity of the algorithm to eye localization error in the context of the gallery for which classification improvement is desired, which we later apply, with good success, to images in the field.

4.1 Preprocessing

The images used to obtain training data consisted of a gallery of 256 individuals, each with four different frontal view poses (for a total of 1024 images) and obtained from the FERET database. The exact set of images can be obtained from the authors.

It was hypothesized that the number of poses of a given individual would affect the relevant characteristics of similarity score distributions. For example, if an individual had ten different poses in a given database, it is conceivable that all ten poses might cluster very closely in the top ranks of the similarity score distribution. On the other hand, with only one pose in the database, an individual's score might be distinctly different from all others, resulting in a similarity score distribution that is much more peaked. This suggested that a multi-resolution approach might be beneficial to extract relevant detail, which might depend on the number of poses each individual has in the database.

Recall that a wavelet basis is described by two functions (the scaling and the mother wavelet function), and a basis is obtained by translating and resizing these functions. Any signal can be represented uniquely and exactly by a linear combination of these basis functions, provided the basis functions are orthornormal. Wavelet basis functions also have a characteristic called compact support, meaning the basis functions are non-zero only on a finite interval. In contrast, the sinusoidal basis functions of the Fourier transform are infinite in extent. The compact support of the wavelet basis functions allows the wavelet transformation to efficiently represent functions or signals which have localized features.

In this application, a 4 point discrete Daubechies wavelet transform [8] was used to process the top 2k sorted similarity scores, where k is the number of poses for each individual. In this case, k=4, resulting in a total of 8 wavelet coefficents,. Reflection generated the necessary points for the function boundary. A Daubechies wavelet transform was used due to its (coarse) similarity to the distributions as well as its overlapping iterations, enabling it to pick up detail that might be missed by, say, the Haar transform.

Two additional features were also computed. The first was the next highest rank of the same ID as the top ranked image. Since only the top 2k similarity scores were observed, this number was clamped at a rank of 2k+1. Very high numbers for ranks are known, from previous experience, to be relatively unstable as predictive features. The intuition here is that the likelihood of the winner being correct is higher if the image of one of its other poses is also highly ranked.

The second feature was the number of pairs of identical Ids in the top 2k similarity scores that have a different ID from the winner. In this case, it was hypothesized that the presence of two (or more) same-ID highly ranked images in the top ranks might also have some bearing on the possibility of classification failure.

4.2 Training

Gallery images were run through each algorithm using all combinations of input eye offsets shown in figure 1, resulting in 9x9=81 runs per algorithm. Note, the same pair of eye offsets was applied to *all* of the gallery images for any given run. Random eye offsets for each individual image were not trained on, since any feasible method used in production would have to apply the same pair of offsets to the entire probe set (see section 4.3).

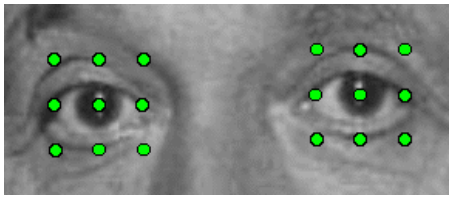


Fig. 1. Eye offsets used for training

The distance between points in the images tested was six pixels. In general, this perturbation depends on the scale of the imaged face, with the goal to select points to span the extent of the eyelids and the whites of the eyes. Similarity scores of the 8 top ranked images were stored along with the other two features discussed previously for all images. Feature vectors were generated and organized into two datasets, one for images whose rank was one (correct matches) and all others (incorrect matches).

A random sampling of 5000 out of 1024x81=82944 feature vectors was used to train a backpropogation neural network [11]. All other 77944 feature vectors were used for testing. This was done for both the FACEIT algorithm and the EBGM algorithm. Thresholds that maximized performance on the test set were fixed for all subsequent experiments and are shown in table 1, along with network architectures and performance. The neural net trained in approximately one day on a G4 Macintosh, and due to the small size of the network, and the relatively small wavelet transform, processed inputs very quickly. Behavior was also observed to be relatively smooth around the peak threshold and relatively stable. Overall performance of the neural net resulted in good generalization, with rates for testing showing only a small loss over training set accuracy.

Table 1. Backpropogation network architecture and performance

Face	Number of Nodes			Constants			Percent Correct		Fixed
Algorithm	Input	Hidden	Output	Learning	Momentum	Sigmoid	Training	Test	Threshold
FACEIT	10	5	1	0.05	0.5	0.05	95.7	94.5	0.4
EBGM	10	5	1	0.05	0.5	0.5	95.2	92.4	0.45

4.3 Random Eye Perturbation Experiments

To study the effectiveness of our approach, we first analyze our prediction ability with respect to controlled simulation experiments. The images used in this experiment are from one session of outdoor data arbitrarily selected from our larger data set collected as follows. Each session consists of the same 1024 FERET images used for training, but displayed on an outdoor LCD monitor and re-acquired under varying time and weather conditions. Images are projected on a 15" LCD monitor and acquired asynchronously by two cameras at high speed from a distance of approximately 100 and 200 ft. Images are zoom adjusted so that facial images have approximately 50-100 pixels between the eyes. Eye coordinates for all images are computed, using the known location of the eyes from the gallery image and a pair of easily identifiable markers located in the projected image.

A series of random Gaussian offsets were applied to the eye coordinates of all images to create a series of probe sets with varying degrees of eye localization error. For this set of experiments, we selected offsets with a mean of zero and four different standard deviations: 2, 4, 6 and 8 pixels radially from the center of the known location of the eye. Note that different random perturbations were applied to each image, and 30 different random seeds were used for each standard deviation. This resulted in 4x30=120 runs of each algorithm on the same set of 1024 images. The intent of this experiment was to show the effectiveness of our approach as eye localization increases in error.

The data flow for the analysis of a single probe image is shown in figure 2. For each probe, the similarity scores are processed and the feature vector passed through the previously trained neural net. If neural net output exceeds the fixed threshold, the image is determined to have a high probability of being correctly classified and its classification is left intact. However, if the neural net output is below the threshold, the image is assumed to have a low probability of being correctly classified, and is then passed onto to the next stage of processing.

Three different subsets of eye offsets were investigated for their effectiveness.

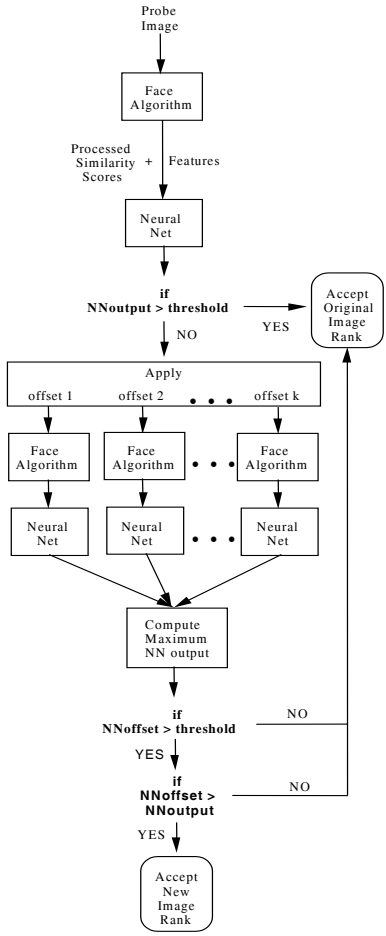


Fig. 2. Flowchart showing data flow for the analysis of a single probe image

In a production setting, it may not be feasible to try all (for example) 81 combinations of offsets (or more) from a resource point of view. It would be beneficial to determine a smaller set of eye perturbations that have a high likelihood of achieving good performance gains versus the cost of reprocessing images. As a result, three subsets of eye offset combinations were tested, referred to as: SCALE (6 offsets), TRANSLATE_SCALE (26 offsets and X_SEP_CONSTANT (8 offsets).

SCALE included those offsets that simply increased or decreased the x separation between the eyes, embodying the implicit hypothesis that scaling is a significant factor affecting face algorithm performance.

X_SEP_CONSTANT included those offsets that simply translated the given x coordinates for both eyes, keeping the distance between them the same.

Finally, TRANSLATE_SCALE included all previous offsets, including scaling in conjunction with translation. No offsets in which one eye was translated in relation to the other were included in the analysis due to the prohibitive cost of post-processing.

Once a probe is identified as having a low probability of being correctly classified, it is then perturbed with an offset, and reprocessed by the face algorithm. This is repeated for all offsets in the subset. The feature vectors each time are input to the neural net, and the largest output (out of all of the offsets applied) is noted. The ranking information for this result supercedes the original classification only if:

1. its neural net output exceeds the fixed threshold
2. its neural net output exceeds that of the original

Results. First, it is instructive to look at how the algorithm behaves with respect to the decisions that are made during processing. As shown in figure 3, the neural net performs extremely well on the initial data, achieving a classification accuracy exceeding 90% over the entire range of initial input eye perturbation. Recall that the eye perturbation in this case is a random Gaussian variable and different for every single image, resulting in a rigorous test for the neural net. Note also, the very low false negative and false positive rates, indicating a relatively high efficiency (at least at this level) of the algorithm.

Not unexpectedly, as the variance of initial eye perturbation increases, performance decreases. However, it is interesting to note that there is a greater *relative* gain as variance increases, and as performance in general decreases. This is shown quite clearly in figure 4. This suggests that such a method might be even more useful as eye localization errors increase since at least one of the perturbations used to try to correct the classification error may be in the direction of the needed change. Changes in and around the correct location may not result in significant benefits. Nevertheless, even in the case of small initial perturbations, significant improvements (albeit small) were noted.

In general, TRANSLATE_SCALE performed slightly better than SCALE, but at a significantly higher cost (see figure 5). With only six offsets, SCALE was able to improve recognition performance significantly with much lower cost. This fact is not very surprising if one considers the importance of scaling in face analysis systems. These results suggest that adjusting factors that affect normalization(specifically eye separation distance) and then re-processing is a prudent approach to improving face recognition. This is consistent with observations made in [4] that eye separation distance seemed to have a greater effect on face recognition performance than the actual location of the eyes themselves.

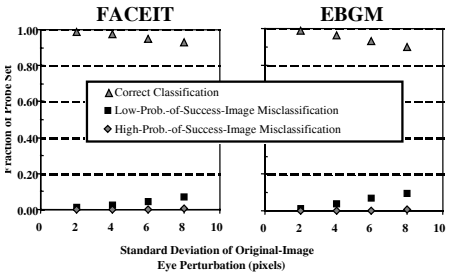


Fig. 3. Classification accuracy of the neural network for FACEIT and EBGM algorithms

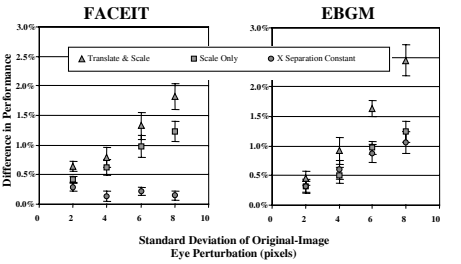


Fig. 4. Classification accuracy of the neural network for FACEIT and EBGM algorithms

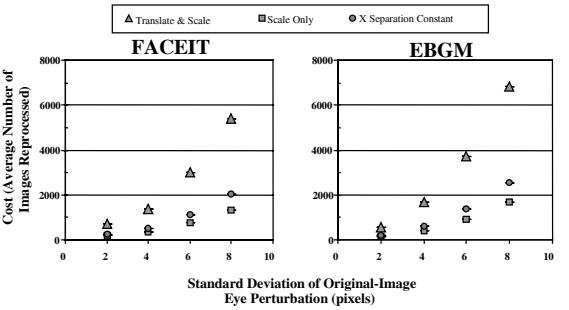


Fig. 5. Maximum number of images reprocessed for each algorithm

Not surprisingly, X_SEP_CONSTANT performed considerably worse, although due to the accuracy of the neural net, performance did not degrade. It is conceivable that bad decisions by the neural net could result in falsely classifying an image as having a high probability of being classified correctly after applying an eye perturbation; however, this was clearly not the case.

With respect to the behavior of the neural network during processing, several important observations can be made. Results only for SCALE are shown in figure 6. First, the fraction of perturbed images that actually resulted in a degraded classification is extremely low, on the order of about 0.1%. Informal observations of the data indicated that even so, the amount of degradation was usually on the order of 1 or 2 ranks (e.g. changing a rank 1 image to a rank 2 or 3). Second, recall that once a probe is initially identified as having a high probability of being incorrectly classified, the image is offset multiple times and the output of the neural net for each re-processing is used to determine what to do with it. If the neural net determines the new result has a low probability of being correctly classified, that result is not considered. As seen in the top of figure 6, the fraction of perturbed images for which this is true is rather high. However, this is to be expected since the likelihood of a given perturbation to actually make things worse is rather high. In fact, the neural net is actually doing quite well, rejecting a large number and accepting only reasonably good possibilities. Of those accepted, *i.e.* when failure is predicted successfully (see the bottom of figure 6), approximately 50% result in an improvement in rank.

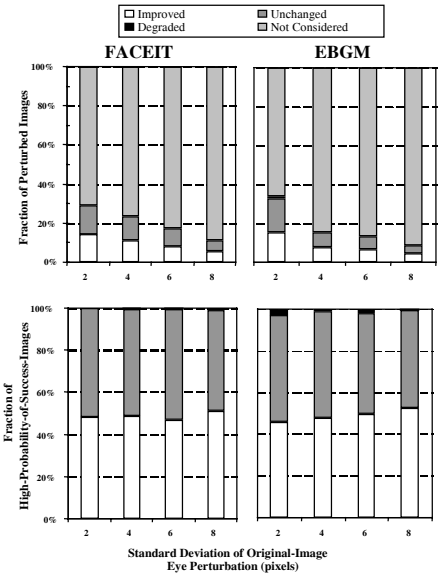


Fig. 6. Breakdown of improved, degraded, unchanged and unconsidered (not detected as failures) images

4.4 Biometric Perturbations of Real Images

Finally, a set of experiments shown in figure 7 clearly shows the benefit of the approach for real images. Four different times of day throughout the month of May were used for this analysis. SCALE perturbations were used to significantly improve face recognition results for the FACEIT algorithm. Note that in this set of experiments, errors in eye localization come from two sources: the eye localization error due to degradation of the input image as a result of atmospheric effects, and the eye localization error due to possible weaknesses in the FACEIT eye localization algorithm. Together, eye localization error is clearly an unknown quantity, but is exploited quite effectively here, to improve overall classification.

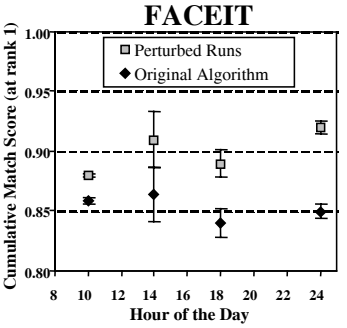


Fig. 7. Performance of FACEIT before and after biometric perturbation. 95% confidence intervals are shown

5 Conclusions

Eye localization has been shown to have a significant impact on face recognition algorithms. This paper uses that fact to show how machine learning and failure prediction can be integrated into a perturbation-based approach for overall system improvement. Our approach was tested on synthetic data using two different face-recognition systems; it showed both good failure prediction performance and, when failure was predicted, corrected for it about 50% of the time. It also managed to do so rather efficiently, requiring only a fraction of the total number of offset combinations, and would be expected to do even better in a production environment.

Using outdoor face data and a commercial face recognition system, the approach was able to predict failures and then predict which perturbations to keep, to achieve a statistically significant 3% to 8% overall improvement beyond the already impressive 85% overall recognition rate of the base commercial face recognition system.

While this paper has focused on face recognition, since the use of "similarity measures" is ubiquitous, this approach should apply across a broad range of pattern recognition problems. In fact, any instance where a weak link exists in a pattern recognition problem, and that also has a limited local perturbation space, is a viable candidate for such an approach.

References

1. Brunelli, R. and Poggio, T. (1993). Face Recognition: Features versus Templates. *IEEE Trans. Pattern Anal. Mach. Intell*. **15**, pp. 1042-1052.
2. Valentin, D., Abdi, D., O'Toole, J., and Cottrell, G. (1994). Connectionist Models of Face Processing: A Survey. *Pattern Recog.*, **27**, pp. 1209-1230.
3. Riopka, T.P. and Boult, T. (2003). The Eyes Have It. *Proceedings of the ACM Biometrics Methods and Applications Workshop*, Berkeley, CA. pp. 33-40.
4. Marques, J., Orlans, N.M., and Piszcz, A.T. (2003). Effects of Eye Position on Eigenface-Based Face Recognition Scoring. *Technical Paper*, Mitre Corp.
5. Okada, K., Steffens, J., Maurer, T., Hong, H., Neven, H. and von der Malsburg, C. (1998). The Bochum/USC Face Recognition System and How It Fared in the FERET Phase III Test. In *Wechsler et al.*, eidtores, *Face Recognition: From Theory to Applic.*, pp. 186-205.
6. Penev ,P. S. and Atick, J. J. (1996). Local feature analysis: A general statistical theory for object representation. *Neural Systems*, 7:477-500.
7. Bolme, D.S., Beveridge, J.R., Teixeira, M. and Draper, B.A. (2003). The CSU Face Identification Eval. System: Its Purpose, Features, and Structure. *ICVS 2003*: 304-313.
8. Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets. *Comm. Pure Appl. Math.*, 41, pp. 909-996.
9. McClelland, J. and Rumelhart, D. (1986). Explorations in Parallel Distributed Processing, Volumes 1 and 2. MIT Press, Cambridge, MA.

Calculation of a Composite DET Curve*

Andy Adler[1] and Michael E. Schuckers[2]

[1] School of Information Technology and Engineering,
University of Ottawa, Ontario, Canada
adler@site.uOttawa.ca
[2] Mathematics, Computer Science and Statistics Department,
St. Lawrence University, Canton, NY, USA and
Center for Identification Technology Research (CITeR)
West Virginia University, Morgantown, WV, USA
schuckers@stlawu.edu

Abstract. The verification performance of biometric systems is normally evaluated using the receiver operating characteristic (ROC) or detection error trade-off (DET) curve. We propose two new ideas for statistical evaluation of biometric systems based on these data. The first is a new way to normalize match score distributions. A normalized match score, $\hat{t}$, is calculated as a function of the angle from a representation of (*FMR*, *FNMR*) values in polar coordinates from some center. This has the advantage that it does not produce counterintuitive results for systems with unusual DET performance. Secondly, building on this normalization we develop a methodology to calculate an average DET curve. Each biometric system is represented in terms of $\hat{t}$ to allow genuine and impostor distributions to be combined, and an average DET is then calulated from these new distributions. We then show that this method is equivalent to direct averaging of DET data along each angle from the center. This procedure is then applied to data from a study of human matchers of facial images.

1 Introduction

One common way to represent the performance of a biometric classification algorithm is the detection error tradeoff (DET) curve. A sample population containing matching (*genuine*) and non-matching (*impostor*) image pairs is presented to the biometric algorithm and the match score, t, calculated to estimate the genuine ($g(t)$) and impostor ($f(t)$) match score distributions. From these distributions, the DET is typically plotted as the false match rate (*FMR*) on the *x-axis* against the false non-match rate (*FNMR*) on the *y-axis*, by varying a threshold τ, and calculating $FMR(\tau) = \int_{\tau}^{\infty} f(x)dx$ and $FNMR(\tau) = \int_{-\infty}^{\tau} g(y)dy$. The DET summarizes the verification performance of the biometric algorithm on the sample population on which it is calculated. Technology evaluations, such as the

* This work is supported by NSERC Canada (Dr. Adler), and by NSF grant CNS-0325640 (Dr. Schuckers) which is cooperatively funded by the National Science Foundation and the United States Department of Homeland Security

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 860–868, 2005.
© Springer-Verlag Berlin Heidelberg 2005

FRVT and FpVTE tests [14][15] use DET curves – or a variant, the Receiver Operating Characteristic (ROC) – to describe their results.

Given its ubiquity, it is perhaps somewhat surprising that few statistical methods have been proposed for analysis and interpretation of DET data in biometric classification. On the other hand, there is a large body of research in the statistical literature, e.g. Zhou et al. [19], and a growing body of work in the machine learning/artificial intelligence literature, e.g. Hernández-Orallo et al. [10]. ROC analysis is used in a wide variety of classification settings including radiography, human perception, and industrial quality control. Zhou et al. ([19]) provide a excellent overview of this work. One limitation of inferential tools for ROC's is the common assumption of Guassian distributions for $g(t)$ and $f(t)$, e.g. Green and Swets [6]. The methodology we propose here does not depend on any distributional assumptions. Another focal area for this research has been the area under the curve or AUC, e.g. Hanley and McNeil [9]. However, biometric authentication has emphasized the equal error rate (EER) as an overall summary of system performance rather than the AUC.

Although most of the literature analyses the ROC, we focus on DET curves since they are more commonly used in biometric identification systems. Here we are motivated to develop methods for a composite DET curve given classification pairs from multiple sources $FMR(\tau)$, $FNMR(\tau)$ in which the original genuine and impostor distributions are either lost, or the match score values, t, are calculated in different spaces. Four types of DET or ROC averaging have been proposed. Bradley [2] suggests using an average based upon the i^{th} ordered threshold in DET space. However, this method leads to difficulties when the number of thresholds tested varies greatly from curve to curve. Vertical averaging (along the FMR) has been suggested by Provost et al. [17], but this method is only appropriate if one of the error rates is more important for some *a priori* reason. When the data to be averaged have very different error rates this method can produce very non-intuitive results, such as if one system reaches $FNMR = 1.0$ at non-zero FMR. Fawcett [5] proposes averaging at the thresholds; however, this method fails when the systems use different match score scales. Finally, Karduan et al. [12] proposed averaging the log-odds transformation of one error rate given the other. In this paper we propose a new method for averaging based on the radial sweep methodology of Macskassy and Provost [13]. This approaches, described below, transforms each curve from the (FMR, $FNMR$) space to polar coordinates.

In this paper we were specifically motivated by how to average the separate DET curves of human volunteers who were asked to perform face recognition [1], by evaluating the whether pairs of images were of the same individual. There are few other reports of comparisons of human face recognition performance to that of automatic systems. Burton and collaborators [3][8] compared PCA based and graph-matching algorithms against human ratings of similarity and distinctiveness, and human memory performance. These studies were focussed on the extent to which automatic algorithms explain features of human performance, rather than as a comparison of recognition performance levels. These

studies did not pursue advanced statistical techniques to synthesize an average masure of human performance. As is typical with data collected from subjective evaluations, assessed values cannot be directly compared between participants. However, in order to compare human face recognition performance levels to each other and to those of automatic software, we wanted a way to calculate the composite human face recognition performance. Because a DET is inherently a two dimensional curve it is difficult to average the curves in a way that properly maintains the importance of both dimensions. In order to address this problem, we develop a technique to calculate an average DET based on regeneration of normalized match scores and distributions. We then show that this is equivalent to a geometrical averaging directly on the DET curves.

The rest of this paper is organized in the following manner. Our method for a composite DET is described in section 2. We then apply this method to data from a group of human subjects (section 3). Finally, in Section 4 and we discuss the applicability of this technique for analysis and interpretation of biometric system verification results.

2 Methods

We use the following notation. A collection of J biometric score distributions are available; each is measured in terms of its own match score $t_i, i = 1 \ldots n_j$. There are no conditions on the match scores other than they be scalar, and increase with match likelihood. The genuine and impostor distributions are represented as $f_j(t_i)$ and $g_j(t_i)$, respectively for $j = 1 \ldots J$. Based on these distributions, the false match rate (FMR_j) and false non-match rate ($FNMR_j$) for biometric system j may be calculated as

$$FMR_j(\tau) = \int_{\tau-}^{\infty} f_j(t)dt = 1 - \int_{-\infty}^{\tau+} f_j(t)dt \tag{1}$$

$$FNMR_j(\tau) = \int_{-\infty}^{\tau-} g_j(t)dt \tag{2}$$

by varying the threshold τ. Clearly, real biometric match score data are not continuous, in which case sums must be used instead of integrals. In this case, it is important that the calculation of either FMR or $FNMR$ but not both, include the distribution value at τ; we include it in the FMR. Implicitly this assumes that the decision process is to accept if the match score is greater than or equal to the threshold, τ. This calculation is illustrated in Fig. 1.

2.1 Normalized Match Scores via Polar Coordinates

In order to perform further analysis on multiple DET curves, it is necessary to calculate a normalized match score common to all curves. In this section, we describe an approach, based on representing the curve in polar coordinates, as illustrated in Fig. 1.

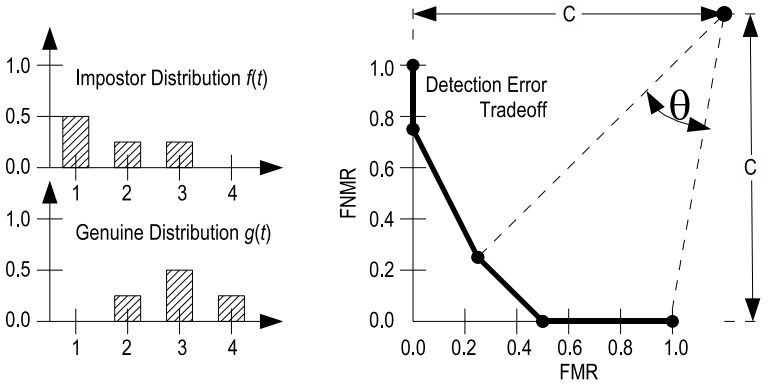


Fig. 1. Calculation of *FMR* and *FNMR* from sample distributions and regeneration of match score t using polar coordinates. Given the discrete *genuine* and *impostor* distributions shown on the left, the DET curve on the right is calculated. From a center at (c, c) an angle θ is calculated to each $FMR, FNMR$ point. A normalized match score t is then calculated from θ. In this example, the distributions are discrete, and the DET curve uses a linear interpolation between points

We have $FMR, FNMR$ coordinate pairs $(x_{ij}, y_{ij}), i = 1, \ldots, n_j; j = 1, \ldots, J$ for a series of J DET curves. By the monotonicity of the DET curves, we know that $x_{1j} \leq x_{2j} \leq \ldots \leq x_{n_j j}$ and $y_{1j} \geq y_{2j} \geq \ldots \geq y_{n_j j}$.

We also assume that no other information is available that would assist us in knowing how the knots in the splines are selected. These points are, as is made clear below, a function of some threshold, τ. Equivalently, we are assuming that no information is available concerning the threshold values. (For example, it would be possible to assume that the thresholds are equally spaced and to derive approximate genuine and imposter distributions following such an assumption.)

Thus, from the DET curve, we calculate an angle

$$\theta_{ij} = tan^{-1}\left(\frac{c - x_{ij}}{c - y_{ij}}\right). \tag{3}$$

We define an angle with respect to the bottom-right of the DET, since at $\tau = -\infty$, $FMR = 1$ and $FNMR = 0$. The DET curve moves left and upward with increasing τ. The limits for θ are

$$\theta_{min} = tan^{-1}\left(\frac{c-1}{c}\right) \tag{4}$$

$$\theta_{max} = tan^{-1}\left(\frac{c}{c-1}\right) \tag{5}$$

Since we wish to calculate a normalized match score $\hat{t}$ in the range $0, \ldots, 1$ from θ, we define

$$\hat{t} = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}} \tag{6}$$

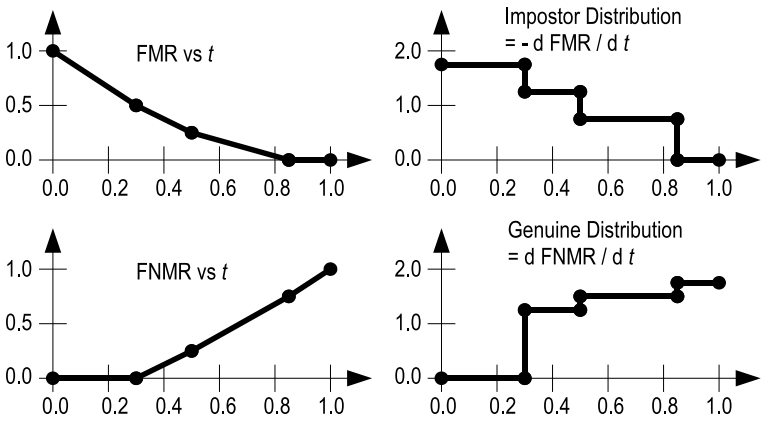


Fig. 2. Reconstructed genuine, $\hat{g}(\hat{t})$, and impostor, $\hat{f}(\hat{t})$, distributions: From the DET curve of Fig. 1 the *FMR* (upper left) and *FNMR* (lower left) are calculated as a function of the normalized match score $\hat{t}$. From these curves, the *impostor* (upper right) and *genuine* (lower right) distributions are calculated as $-\frac{d}{dt}FMR$ and $\frac{d}{dt}FNMR$, respectively

2.2 Distributions from DET Curves

In this section, we use the polar-coordinate representation, to reconstruct candidate genuine, $\hat{g}(\hat{t})$, and impostor, $\hat{f}(\hat{t})$ distributions. Based on the equations 1 and 2, we calculate for each DET curve j.

$$f_j(\hat{t}) = -\frac{dFMR_j}{d\hat{t}} \tag{7}$$

$$g_j(\hat{t}) = \frac{dFNMR_j}{d\hat{t}}. \tag{8}$$

Fig. 2 illustrates the calculations. Since *FMR* and *FNMR* data are not continuous, but are sampled from the DET, the distributions must be defined in terms of discrete approximations to the derivative. One consequence of the discrete derivative is that $\hat{g}$ and $\hat{f}$ are noisy, but this does not matter for this application.

Using this calculation, we now have a collection of distributions $\hat{g}_j, \hat{f}_j$ for $j = 1 \ldots J$, which are all based on the same match scores, $\hat{t}$'s. It is thus possible to combine the distributions, weighted by the number of samples in each (if known). The number of samples in each genuine and impostor distribution are represented as $n_{g,j}$ and $n_{f,j}$, respectively. If the number of samples is unknown, all n values are assumed to be equal. The combined distributions $\bar{f}$ and $\bar{g}$ are

$$\bar{f} = \frac{1}{N_f}\sum_{j=1}^{J} n_{f,j}\hat{f}_j \tag{9}$$

$$\bar{g} = \frac{1}{N_g}\sum_{j=1}^{J} n_{g,j}\hat{g}_j \tag{10}$$

where $N_f = \sum n_{f,j}$ and $N_g = \sum n_{g,j}$.

However, this expression may be shown to be equivalent to a direct averaging of the DET curves in $(FMR, FNMR)$ space, as follows:

$$\hat{FNMR}(\hat{t}) = \int_{-\infty}^{\tau-} \bar{g}(t)dt \tag{11}$$

$$= \int_{-\infty}^{\tau-} \frac{1}{N_g} \sum_{j=1}^{J} \frac{1}{dt} dFNMR_j(t)dt \tag{12}$$

$$= \int_{-\infty}^{\tau-} \frac{1}{N_g} \sum_{j=1}^{J} n_{g,j} \frac{1}{dt} dFNMR_j(t)d\hat{t} \tag{13}$$

$$= \frac{1}{N_g} \sum_{i=1}^{J} n_{g,j}(FNMR_j(\hat{t}) - FNMR_j(-\infty)) \tag{14}$$

$$= \sum_{j=1}^{J} \frac{n_{g,j}}{N_g} FNMR_j(\hat{t}) \tag{15}$$

Similarly,

$$\hat{FMR}(\tau) = \sum_{j=1}^{J} \frac{n_{f,j}}{N_f} FMR_j(\hat{t}) \tag{16}$$

Thus, the average DET at each angle θ can be calculated by a (possibly weighted) average the distance of each curve from (c, c).

3 Results

This paper uses data from a comparison of human and automatic face recognition performance [1]. This study investigated the ability of interested and motivated non-specialist volunteers to perform face identification tasks matched against performance by several commercial face recognition software packages. Images were obtained from the NIST mugshot database [16]. Pairs of frontal pose face images were randomly created from this database. Two-thirds of the pairs were impostors (images of different persons), and one third were genuines (different images of the same person). No special effort was made to select images of the same gender or ethnicity for the impostor pairs.

Twenty one people (16 male, 5 female) participated in the experiments. They were predominantly Caucasian and in the age range 20–40. Participants were asked to log onto a web site, where an application server would present pairs of face images, and the participant was asked whether they were from the same person. Participants were not given any information about the distribution of genuines and impostors, or any feedback about their success. Participants were presented the following options: *same*, *probably same*, *not sure*, *probably different*, or *different*. Each option was converted to a match score value (such that *different*= 1 and *same*= 5).

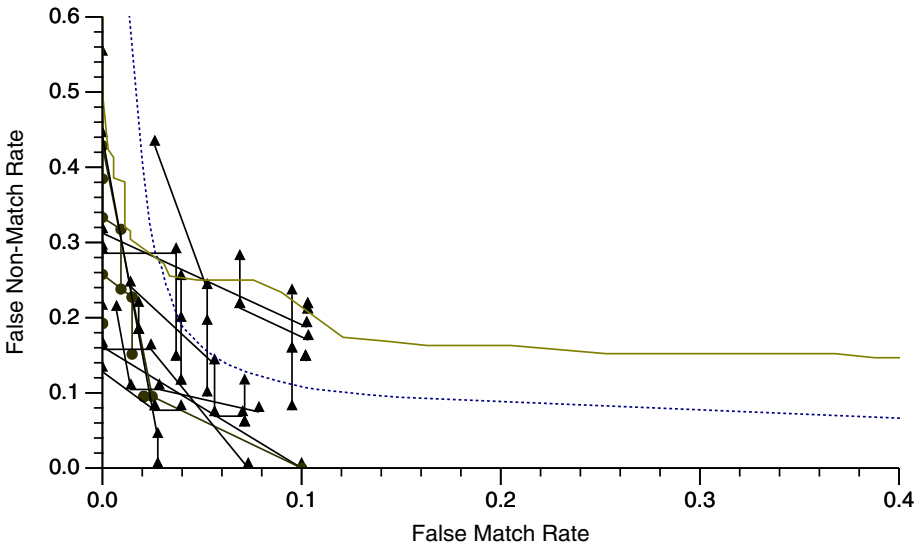


Fig. 3. Calculation of an average DET curve for human face recognizers. Individual human DET curves are shown by symbols (circle=female, triangle=male). The average curve (dotted line) is calculated using the method of this paper. For comparison, the highest performing software available to us in 2003 is also shown (solid line)

4 Discussion

In this paper we have presented a new methodology for combining and averaging DET or ROC curves. This approach was motivated by the need to create a composite DET curve for human evaluators of human faces. This methodology was developed independently of [13]; however, it uses the same basic technique of radially sweeping across the DET curve to create a normalized match score. This permits the creation of normalized distributions for *FMR* and *FNMR* that are a composite of individual DET curves. This normalization is a significant advance in and of itself and adds to a growing body of methods for this purpose [11]. We have used this normalization to to average at normalized radial match scores.

Several issues arise from radial sweeping of DET curves. The first is where to locate the center of the sweeping. Because we would like the averaging to not depend on which error rate is on which axis, we limited possible center points to (c, c) for some constant c. It is immediately clear that choosing a center along the $FMR = FNMR$ line results in an average curve that is independent of the selection of axes. We considered three possible values for c, $0, 1$ and ∞. Choosing $c = 0$, often resulted in composite or average curves that were counter-intuitive because of the acute angles near the axes. This is especially important for biometric systems which are often placed in settings where low *FMR*'s are required. There was little difference between the curves when $c = 1$ and $c = \infty$. However, we prefer $c = 1$ because the radial angles match the typical curvature

of a DET curve and, hence, are more likely to be perpendicular to such curves. The choice of $c = \infty$ results in averaging across parallel 45° lines.

Another issue is the choice of how to "average" the curves. Here we have effectively taken an arithmetic average of the curves. Other choices are possible including a weighted average, to account for database size or importance by varying the weights to be given to each DET. An alternative would be to use a radial median at each angle. This would results in a spline that is not as smooth as the radial mean DET but which may be more robust to "outlying" DET curves.

The question of inferential methods based on the radial mean DET is one that is important for future study. Here we are interested in creating confidence intervals for an individual curve (as in [13]) as well as being create a confidence interval for the difference of two DET curves. Similarly we would like to create tests for significant differences between two or more DET curves. It might also be of interest to test a single observed DET against a hypothetical DET curve. This last case may take the form of a Kolmogorov-Smirnov type test.

References

1. Adler, A., Maclean, J.: "Performance comparison of human and automatic face recognition" *Biometrics Consortium Conference 2004* Sep. 20-22, Washington, DC, USA (2004)
2. Bradley, A. P.: "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* **7**, 1145-1159 (1997).
3. Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., Henderson, Z.: "Human and automatic face recognition: a comparison across image formats" *Vision Research*, 41:3185-3195, 2001.
4. Drummond, C., Holte, R. C.: "What ROC Curves Can't Do (and Cost Curves Can)" In *Proc. 1st Workshop ROC Analysis in AI:ROCAI*, 19-26, (2004).
5. Fawcett, T.: *ROC graphs: Notes and practical considerations for data mining researchers*, Technical Report HPL-2003-4. HP Labs. (2003).
6. Green, D. M., Swets, J. A.: *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York, (1966).
7. Golfarelli, M., Maio, D., Maltoni D.: "On the Error-Reject Trade-Off in Biometric Verification Systems" *IEEE Trans. Pattern Anal. Machine Intel.* **19** 786–796 (1997)
8. Hancock, P. J. B., Bruce, V., Burton, M. A.: "A comparison of two computer-based face identification systems with human perceptions of faces" *Vision Research* 38:2277-2288 (1998).
9. Hanley, J. A., McNeil, B. J.: "The meaning and use of the area under a receiver operating characteristic (ROC) curve" *Radiology* **143** 29–36 (1982).
10. Hernández-Orallo, J., Ferri, C., Lachiche, N. Flach, P.A.,ed.: *ROC Analysis in Artificial Intelligence, 1st Int. Workshop, ROCAI-2004*, Valencia, Spain (2004).
11. Jain, A.K, Nandakumar, K.: Ross, A.: "Score Normalization in Multimodal Biometric Systems", *Pattern Recognition*, (in press, 2005).
12. Karduan, J., Karduan, O.: "Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation" *Methods Inform. Med.* **29** 12-22 (1990).

13. Macskassy, S., Provost, F.: "Confidence Bands for ROC Curves: Methods and an Empirical Study." In *Proc. 1st Workshop ROC Analysis in AI:ROCAI*, 61-70, (2004).
14. NIST: *Face Recognition Vendor test 2002* `http://frvt.org/frvt2002`
15. NIST: *Fingerprint Vendor Technology Evaluation (FpVTE) 2003* `http://fpvte.nist.gov/`
16. NIST: *NIST Special Database 18: Mugshot Identification Database (MID)* `http://www.nist.gov/srd/nistsd18.htm`
17. Provost, F. J., Fawcett, T., Kohavi, R.: "The case against accuracy estimation for comparing induction algorithms" In *Proc. 15th Int. Conf. Machine Learning*, 445-453 (1998).
18. Rukhin, A., Grother, P., Phillips, P.J., Newton, E.: "Dependence characteristics of face recognition algorithms" *Proc. Int. Conf Pattern Recog.* **16** 36–39 (2002)
19. Zhou, X.-H., McClish, D. K., Obuchowski, N. A.: *Statistical Methods in Diagnostic Medicine* John W. Wiley & Sons, (2002).

Privacy Operating Characteristic for Privacy Protection in Surveillance Applications

P. Jonathon Phillips

National Institute of Standards and Technology, Gaithersburg MD 20899, USA
jonathon@nist.gov

Abstract. With the mass deployment of cameras, concern has risen about protecting a person's privacy as he goes about his daily life. Many of the cameras are installed to perform surveillance tasks that do not require the identity of a person. In the context of surveillance applications, we examine the trade-off between privacy and security. The trade-off is accomplished by looking at quantitative measures of privacy and surveillance performance. To provide privacy protection we examine the effect on surveillance performance of a parametric family of privacy function. A privacy function degrades images to make identification more difficult. By varying the parameter, different levels of privacy protection are provided. We introduce the privacy operating characteristic (POC) to quantitatively show the resulting trade-off between privacy and security. From a POC, policy makers can select the appropriate operating point for surveillance systems with regard to privacy.

1 Introduction

With the construction of large databases and mass deployment of devices capable of collecting vast amounts of personnel data, we must deal with the issue of privacy associated with the data collected. Here, privacy means the capability to restrict or control the ability to positively identify a person.

There are many instruments for collecting data and information about a person. Examples are: credit cards, brand loyalty cards, medical records, and mobile phones. For our purposes in this paper, we will only consider cameras and surveillance tasks. An example of a surveillance task is tracking a group of people in a building. Privacy and surveillance tasks raise the following questions: can one develop technology that prevents positive identification of a person, while at the same time, allowing for the completion of a surveillance task that does not require that a person be positively identified? For example, in tracking a person's movements, enough information is to needed to follow a person, but not enough to identify the person.

Clearly there is a trade-off between privacy and completion of a surveillance task. At one end of the spectrum, maximal privacy is assured when surveillance cameras are not installed. However, there will be compromises to needed security processes. At the other end of spectrum, maximal security protection is possible

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 869–878, 2005.
© Springer-Verlag Berlin Heidelberg 2005

when the cameras are installed and the video imagery is not degraded. However, in this case, privacy is the most compromised.

We approach the privacy protection problem by defining a parametric family of privacy functions f_α. A privacy function degrades an image or video sequence with the goal of making identification harder. The more degradation of an image, the greater the privacy afforded an individual, and the greater the difficulty of the surveillance task.

The motivation for the design of our privacy functions are image compression techniques. Compression techniques have been adapted for pattern recognition. Eigenfaces is an adaptation of principal component analysis (PCA) for face recognition [1]. The trade-off between privacy and security replaces the trade-off between fidelity and space in compression.

Defining a family of privacy functions f_α is only part of the process of incorporating privacy protection into surveillance. In order to perform a trade-off between privacy and security, it is necessary to define a quantifiable privacy measure $\mathcal{P}(f_\alpha)$. To define a privacy measure it is necessary to understand and explicitly state the privacy requirements and assumptions. Next we define define the surveillance task and a surveillance performance measure $\mathcal{S}(f_\alpha)$. A surveillance performance quantifies the performance of the surveillance task,

To show trade-off between privacy and security, we introduction the privacy operating characteristic (POC). The POC is produced by computing the privacy and surveillance performance measures for different parameter values α of the privacy functions. Each parameter value will produce a privacy measure and surveillance performance measure. The set of privacy measures and surveillance performance measures are plotted on a POC, which explicitly shows the trade-off between privacy and security. The x-axis plots the privacy measure and the y-axis plots the surveillance performance measure. The POC is analogous to a receiver operating characteristic (ROC) from signal detection theory. From a POC, policy makers can select the appropriate operating point for a surveillance systems with regard to privacy.

In this paper we introduce a four step framework for incorporating privacy concerns into a surveillance task. The steps are summarized below:

1. Define the privacy requirements and a privacy measure.
2. Define the surveillance task and measure of performance.
3. Define a parametric family of privacy functions.
4. Compute the trade-off between privacy levels and surveillance task performance. Plot the trade-off on a privacy operating characteristic (POC).

The incorporation of privacy concerns into surveillance is a new area of research. Newton, Sweeney, and Malin [2] introduced the concept of privacy functions for de-identifying faces. Their privacy measure examines the de-identifying faces and measures performance on closed-set identification, and does not address the impact of privacy on surveillance tasks.

Because of the importance and impact of privacy concerns, Newton, Sweeney, and Malin and this paper will likely be the start of a long discussion on how to best incorporate privacy concerns into surveillance systems. This paper provides

a basis for further research by quantitatively addressing the trade-off between privacy and security. The major contributions of this paper are to:

1. provide a framework for incorporating privacy protection into surveillance tasks,
2. characterize performance in terms of the trade-off between privacy and surveillance,
3. introduce the privacy operating characteristic (POC) to show the trade-off between privacy and surveillance, and
4. provide a detailed example of this framework.

The power of this framework will be shown by an example that models tracking people in a building. The privacy requirement is that the people being tracked cannot be positively identified. The surveillance task is that enough information is maintained so that the people can be tracked as they move around a building.

2 Identification and Verification

The problem formulation of the example requires knowledge of the performance metrics in face recognition and biometrics. The privacy measure is based on the verification task in face recognition and the surveillance measure is based on the closed-set identification task. The face recognition performance measures are taken from the Sep96 FERET evaluation [3].

Some basic terms are introduced to describe how face recognition experiments are conducted and performance is computed. A *gallery* is the set of known individuals. The images used to test the algorithms are called *probes*. The identity of the face in a probe is not known to the algorithm. A probe is either a new image of an individual in the gallery or an image of an individual not in the gallery. A collection of probes is called a *probe set.*

The closed-set identification task returns the top match between a probe and the images in a gallery. Performance for this task is the identification rate which is the fractions of probes that are correctly identified.

In a typical verification task, a subject presents an image to a system and claims to be a person in the system's gallery. The presented biometric image is a probe. The system then compares the probe with the stored image of the subject in the gallery. Based on this comparison the claim is either accepted or rejected.

Verification and false accept rates characterize verification performance. The verification rate (VR) is the fraction of valid claims that are accepted; the false accept rate (FAR) is the fraction of false claims that are accepted. Both these performance rates cannot be maximized simultaneously; there is a trade-off between them. Verification performance is reported by showing the trade-off between the verification and false accept rates on a receiver operator characteristic (ROC). The horizontal axis is the false accept rate and the vertical axis is the verification rate.

3 Problem Formulation

The goal of this paper is to provide a framework for the inclusion of privacy protection into video surveillance tasks. We will demonstrate the framework by an example. The example models the surveillance task of tracking a small group of people around a building. To maintain the track of each person, the system needs to be able to differentiate among the facial images of people in a small group; however, the systems does not have to identify people in the group. In our example, we model the group of people being tracked as a gallery of ten facial still images. On the privacy side, the goal is to make it difficult to determine the absolute identity of a person. The absolute identity means that we can identity the face as belonging to a specific person. For example, the person in an image is John Smith. The surveillance task is to maintain the relative identity of the small group of faces. Determining the relative identity of a face means that given a facial image of person in the group, we can correctly identify which person it is from among the ten faces in the gallery.

We will now address the privacy part. To provide privacy protection, we degrade the observed face image of a person g_i by a privacy function f_α. The privacy function returns a degraded image $g_i^d = f_\alpha(g_i)$. The privacy functions are parameterized by α which allows for varying degrees of privacy. The goal of a privacy function is to produce a degraded image g_i^d that will attenuate the ability to perform the absolute identification. Varying α allows for the trade-off between privacy and surveillance.

The effectiveness of a privacy function f_α at setting α is determined by a privacy measure. The privacy measure presented in this paper models a scenario where a person claims to be the same person as in image g_i^d. The claimant and the person in g_i^d need not be the same person. In fact, the goal is to find a privacy function f_α such that one cannot accurately determine if the claimant and the person in g_i^d are the same. Let p be a facial image of the claimant. The question then becomes, Are the faces in images g_i^d and p of the same person? This is equivalent to the open-set verification problem in biometrics. Since, a ROC is a curve not a number, it cannot be a privacy measure. In this paper, we present a privacy measure based on the ROC. Our privacy measure is not the only possible measure. Future research, experiments, and experiences will guide in the selection of appropriate privacy measures and under which situations a particular privacy measure is applicable.

Our privacy measure $\mathcal{P}(f_\alpha)$ is the FAR that corresponds to a VR = 0.5. A verification rate of 50% means that when g_i^d and p are of the same person, the decision is a random guess as to whether g_i^d and p are the same face. The corresponding FAR is the percentage of the population that is similar to looking to the degraded image g_i^d. Even if FAR is small, 1%-5%, this means that a face recognition algorithm will report 2.5-12.5 million people in the U.S. has being similar to g_i^d.

When computing our privacy measure, g_i^d is compared to p. Another option is to compare g_i^d and p^d. The experiments in this paper use a PCA-based face

recognition algorithm. With PCA-based face recognition algorithms, comparing g_i^d with p or p^d produces the same result.

The second part of the problem is characterizing performance of the surveillance task as a function f_α. The surveillance task is modeled as a small closed-set identification problem, with a gallery $G = \{g_1, \ldots, g_N\}$. The gallery consists of N people with one image per person. All images in the gallery are degraded by the privacy function, which produces a degraded gallery $G^d = \{g_1^d, \ldots, g_N^d\}$. Let $P_d = \{p_1^d, \ldots, p_N^d\}$ be a probe set were the probes have been degraded. All the probes are images of a person in the gallery. The surveillance performance measure $\mathcal{S}(f_\alpha)$ for the surveillance task is the correct identification rate. In tracking a person around a building, the gallery models the set of people that are in the building. The probes model people being tracked, and is one of the inputs into recording a track.

What makes these seemingly impossible goals feasible is the nature of the two parts of problem. The first part concerns privacy, which is an open-universe problem. The second part is the surveillance task, i.e., who is this person from a small closed universe population.

4 Experiments

The experiments we performed show the both the effect of privacy functions on face recognition performance and the plot the trade-off between privacy and security on a POC.

The experiments in this paper are performed with images from the FERET database. The FERET database provides a common database of facial images for both development and testing of face recognition algorithms and has become the de facto standard for face recognition from still images [3].

We report identification scores for two categories of probes. The first probe set was the **FB** set. In this set, the gallery and probe images of a person were collected on the same day under the same conditions. The second set were dup I probes, which consist of duplicate images. A *duplicate* is defined as an image of a person whose corresponding gallery image was taken on a different date or under different conditions; e.g., wearing glasses or with hair pulled back.

The experiments in this paper are performed using a PCA-based face recognition algorithm. PCA is a statistical dimensionality reduction method, which produces the optimal linear least squared decomposition of a training set. Kirby and Sirovich [4] applied PCA to representing faces and Turk and Pentland [1] extended PCA to recognizing faces.

A PCA representation is characterized by a set of N eigenvectors $(\mathbf{e}_1, ..., \mathbf{e}_N)$ and eigenvalues $(\lambda_1, ..., \lambda_N)$. In the face recognition literature, the eigenvectors can be referred to as *eigenfaces* or *eigenfeatures*. We normalize the eigenvectors so that they are orthonormal. The eigenvectors are ordered so that $\lambda_i > \lambda_{i+1}$. One of the factors effecting the accuracy of a PCA-based face recognition algorithm is the number of eigenfeatures in the representation [5]. A general rule of thumb is to include the top 40% of eigenfeatures in the representation. Performance increases as more eigenfeatures are added up to the 40% rule of thumb.

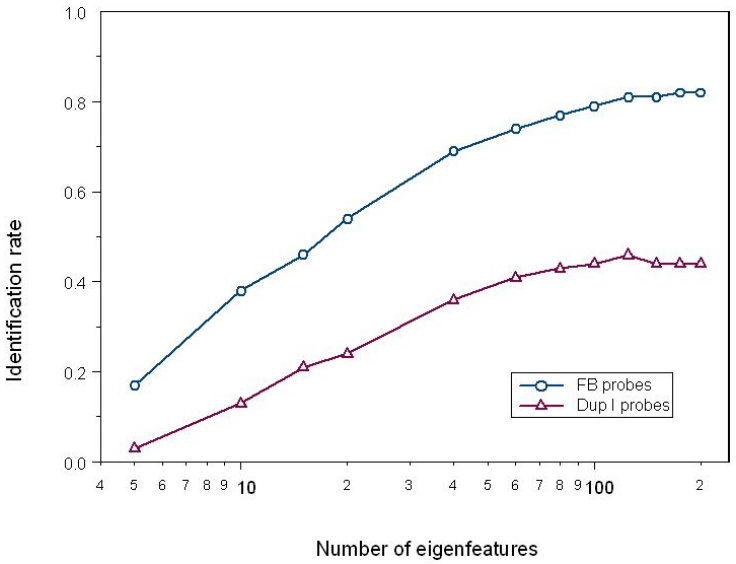


Fig. 1. Closed-set identification performance as a function of the number of eigenfeatures in the representation. Results are given for the Sep96 FERET evaluation **FB** and dup I probe sets. The horizontal axis is on a logarithmic scale

The increase in performance in closed-set identification as a function of the number of eigenfeatures in the representation is shown in figure 1. Performance is reported for the Sep96 FERET evaluation **FB** and dup I probe sets [3]. Figure 1 shows performance increasing for the **FB** probes from 17% for five eigenfeatures to 82% for 200 eigenfeatures and from 3% to 44% for the dup I probes.

In our experiments, the privacy function is the number of eigenfeatures in the representation. The privacy function f_α represents a facial image by eigenfeatures $\mathbf{e}_1, \ldots, \mathbf{e}_\alpha$. Strictly, f_α filters an image through the eigenbasis $\mathbf{e}_1, \ldots, \mathbf{e}_\alpha$. For this privacy function, a smaller α yields a greater level of privacy protection.

In computing a POC, we need two sets of performance figures. We will start with the surveillance performance measure $\mathcal{S}(f_\alpha)$, which is the closed-set identification rate on a gallery of ten people. To provide a more reliable estimate of performance, we report the average performance on four galleries. All four galleries contain different people. Performance is computed for **FB** and dup I probes. (These probe sets only contain probes of people in the gallery. They are not the full Sep96 FERET **FB** and dup I probe sets.) Figure 2 shows $\mathcal{S}(f_\alpha)$ as a function of α, the number of eigenfeatures in the representation. For a gallery of ten, $\mathcal{S}(f_\alpha)$ saturates at 98% for **FB** probes with 15 eigenfeatures, and for dup I probes reaches 90% at 60 eigenfeatures.

The second component of a POC is the privacy measure $\mathcal{P}(f_\alpha)$. Our privacy measure is the false alarm rate that corresponds to a verification rate of 0.50. Figure 3 plots average $\mathcal{P}(f_\alpha)$ for dup I probes as a function of the number of eigenfeatures in the representation. Performance goes from a $\mathcal{P}(f_\alpha) = 11.3\%$ for five eigenfeatures to 2.8% for 15 eigenfeatures and 3.0% for 20 eigenfeatures.

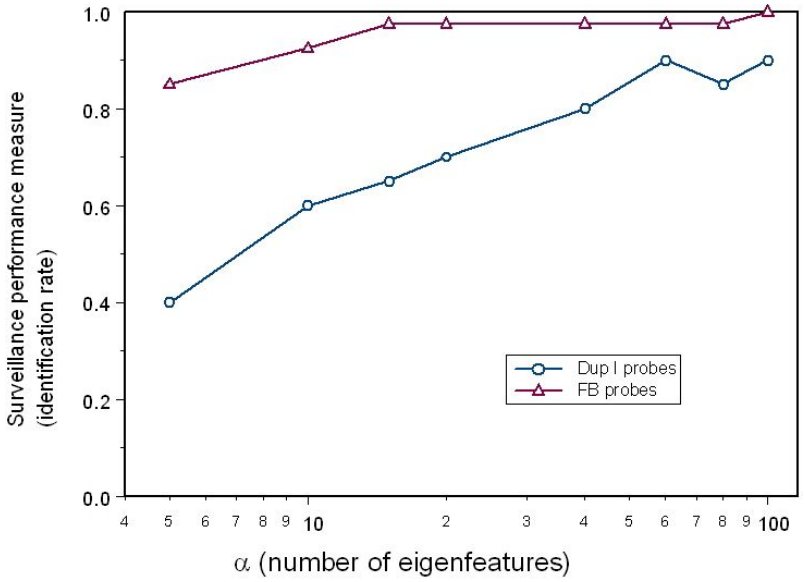


Fig. 2. The average surveillance performance measure $\mathcal{S}(f_\alpha)$ over four galleries of ten people as a function of α, the number of eigenfeatures in the representation. The horizontal axis is on a logarithmic scale

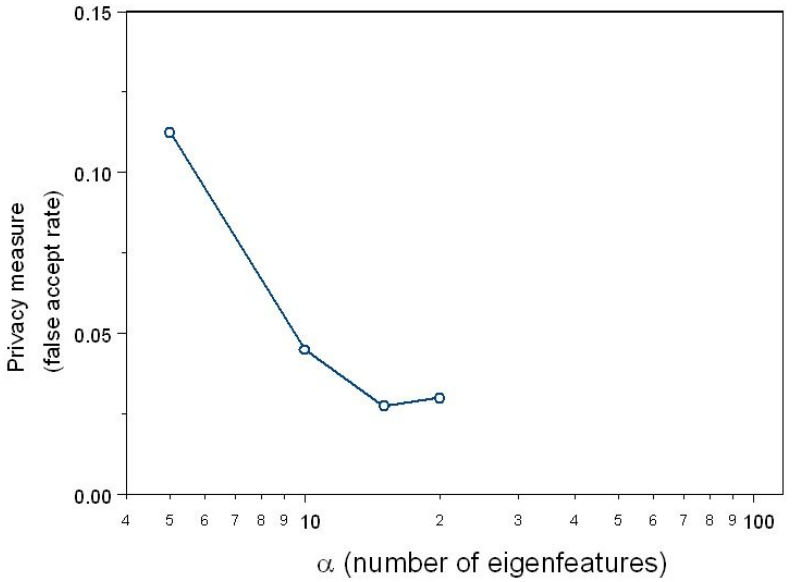


Fig. 3. Average privacy measure $\mathcal{P}(f_\alpha)$ as a function of α, the number eigenfeatures in the representation. The horizontal axis is on a logarithmic scale

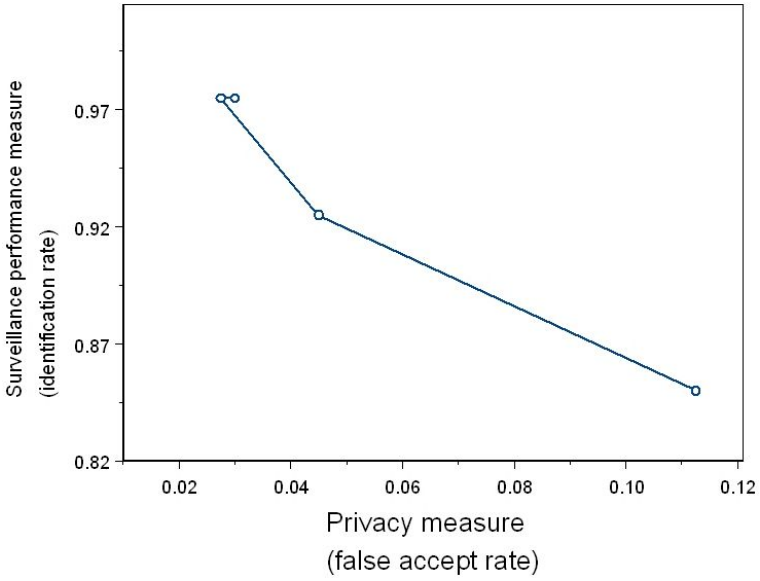


Fig. 4. POC for the identification performance measure computed on **FB** probes and the privacy measure computed on the dup I probes

Figure 4 is the POC for the trade-off between $\mathcal{P}(f_\alpha)$ and $\mathcal{S}(f_\alpha)$ on the dup I probes. The x-axis is the privacy measure $\mathcal{P}(f_\alpha)$ and the y-axis is the surveillance performance measure $\mathcal{S}(f_\alpha)$. Each point on the POC is the $\mathcal{P}(f_\alpha)$ and $\mathcal{S}(f_\alpha)$ for α eigenfeatures in the representation. For example, the point with $\mathcal{P}(f_\alpha) = 0.11$ and $\mathcal{S}(f_\alpha) = 0.85$ was generated from a representation of 15 eignfeatures. The $\mathcal{P}(f_\alpha)$ can be read from figure 3 and the $\mathcal{S}(f_\alpha)$ can read from figure 2. The POCs are a merger of the curves in figures 2 and 3. In our POCs, the number of eigenfeatures is the hidden parameter α.

We ran experiments on two categories of probe sets: **FB** and dup I categories. In computing a POC, the probe set categories for computing the privacy measures and the surveillance measures, do not have to be same. The probe categories are different because they model different operational scenarios. The identification task models tracking were images of a person are acquired within five to ten minutes of each other. The **FB** probe category fits this scenario. The verification performance models the situation were someone claims to be the person in a degraded image. The image of the claimant will most likely be taken on a different day and under conditions than the gallery image. This situation is modeled by using dup I probes. Computing the privacy measure from **FB** probes models the situation were the claimant presents an image that comes from the same video sequences as the gallery images. This would be a highly unusual situation because the claimant would have to have direct access to the surveillance video sequences.

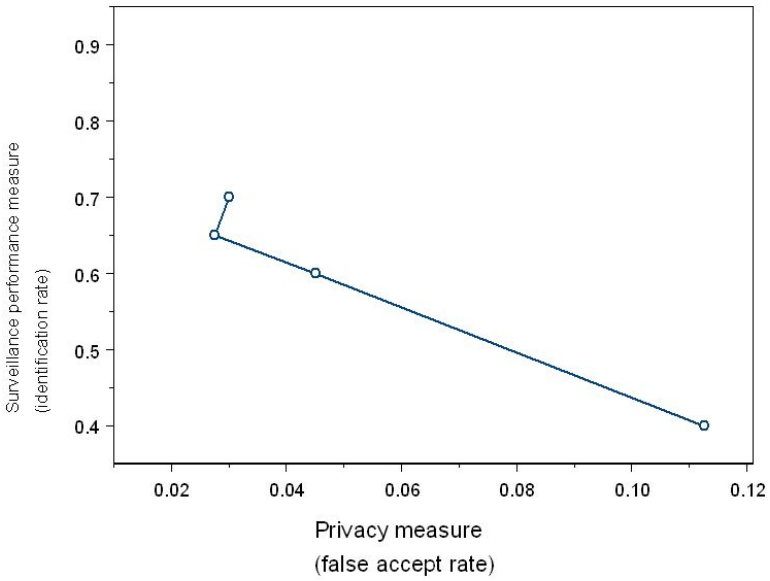


Fig. 5. POC for both identification performance and privacy measures computed on dup I probes

Figure 5 plots the POC for identification performance and privacy measure computed on the dup I probes. In Figure 5, the privacy and surveillance probe sets are both from the same category. It is clear that the privacy-security trade-off is much better for the **FB** probes in the surveillance task than the dup I probes.

5 Conclusions

We have introduced the privacy operating characteristic (POC) to quantitatively show the trade-off between privacy and security. The POC is part of a framework for incorporating privacy protection into surveillance applications. A parametric family of privacy functions degrade images and video sequences to provide varying levels of privacy protection. Changing the parameter produces different levels of privacy. The framework includes a privacy measure and surveillance performance measure. A privacy measure assess the level of privacy protection provided by a privacy function. A surveillance performance measure assess the effectiveness of the surveillance task. Privacy measures and surveillance performance measures quantify the effect of the privacy function on the the privacy-security trade-off. This is plotted on a POC.

Our framework was illustrated by an example with experimental results presented in Section 4. The results in figures 4 and 5 show that using **FB** probes in the surveillance task have a better privacy-security trade-off than dup I probes. This provides very preliminary evidence that surveillance tasks that require

tracking of people over short periods of time can be designed with privacy protection included. The ability to provide privacy protection when tracking people over days or weeks will be a harder problem to solve.

This paper lays the basis for future research in privacy and security. Our experiments report results for a classical face recognition algorithm, Eigenfaces. One avenue of research is to design face recognition algorithms that explicitly incorporate privacy consideration. Another avenue of research is to investigate what consists a good privacy measure.

With the deployment of large number of video cameras and the increasing sophistication of surveillance systems, it necessary that the computer vision community address privacy issues. The computer vision community provides a unique ability to look at technical issues associated with privacy and surveillance and to develop privacy protection technology for incorporation into surveillance systems.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cognitive Neuroscience **3** (1991) 71–86
2. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying facial image. IEEE trans. Knowledge and Data Engineering **17** (2005) 232–243
3. Phillips, P.J., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. PAMI **22** (2000) 1090–1104
4. Kirby, M., Sirovich, L.: Application of the karhunen-loeve procedure for the characterization of human faces. IEEE Trans. PAMI **12** (1990) 103–108
5. Moon, H., Phillips, P.J.: Computational and performance aspects of PCA-based face-recognition algorithms. Perception **30** (2001) 303–321

Headprint – Person Reacquisition Using Visual Features of Hair in Overhead Surveillance Video

Hrishikesh Aradhye, Martin Fischler, Robert Bolles, and Gregory Myers

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

Abstract. In this paper, we present the results of our investigation of the use of the visual characteristics of human hair as a primary recognition attribute for human ID in indoor video imagery. The emerging need for unobtrusive biometrics has led to recent research interest in using the features of the face, gait, voice, and clothes, among others, for human authentication. However, the characteristics of hair have been almost completely excluded as a recognition attribute from state-of-the-art authentication methods. We contend that people often use hair as a principal visual biometric. Furthermore, hair is the part of the human body most likely to be visible to overhead surveillance cameras free of occlusion. Although hair can hardly be trusted to be a reliable *long-term* indicator of human identity, we show that the visual characteristics of hair can be effectively used to *unobtrusively* re-establish human ID in the task of short-term recognition and reacquisition in a video-based multiple-person continuous tracking application. We propose new pixel-based and line-segment-based features designed specifically to characterize hair, and recognition schemes that use just a few training images per subject. Our results demonstrate the feasibility of this approach, which we hope can form a basis for further research in this area.

1 Introduction

The emerging need for *unobtrusive* identification has led to recent research focusing on identification based on face, gait, voice, and clothes, among others. We present the results of our novel use of the visual characteristics of human hair as a primary recognition attribute in midrange still and video indoor imagery. In addition, we discuss scenarios where such use may be feasible and synergistically advantageous in the context of existing unobtrusive biometric tests.

Although often used by people as an important visual biometric, characteristics of hair have been almost completely excluded as a recognition attribute for the computer-based determination of human ID. Such exclusion is understandable in applications where the time interval between enrollment and recognition can be very long, because hair can be removed, substituted, recolored, or rearranged, and can hardly be trusted to be a reliable *long-term* indicator of human identity. Nevertheless, we contend that there are important applications where characteristics of hair can play a critical role in establishing human ID. One such application is video-based multiple-person continuous tracking and reacquisition. In the context of people-tracking systems the term *recognition and reacquisition* corresponds to the objective of reassigning a newly detected

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 879–890, 2005.
© Springer-Verlag Berlin Heidelberg 2005

person or persons to their previously tracked identities as they emerge from blind spots or separate from a group and/or occluding objects. Significant changes in the hair appearance of the people involved are not likely in such a short-term scenario, making the visual characteristics of hair a promising attribute for recognition and reacquisition. Furthermore, the part of the human body that is most likely to be visible to the commonly used overhead surveillance video cameras free of occlusion is the top of the head, and correspondingly, hair. We show in this paper that visual characteristics of hair can be effectively used to re-establish human ID in such an application.

2 Background and Previous Work

An average adult head typically contains 100,000 to 150,000 individual hair strands, each with a diameter of approximately 20–200 μm [1]. This average hair width is too small for each hair strand to be uniquely separable in images captured with consumer grade image capture equipment. For instance, the still images used in this work were captured by a still camera placed about 2 ft above the head of a person of average height. The resolution for still images was 1600×1200, corresponding to approximately 250 μm of head surface per pixel (i.e., 100 dpi) in the image for the average-sized head. Our video imagery, on the other hand, resulted in 720×480 frame sizes at roughly 420 μm of head surface per pixel (i.e., 60 dpi) for the average head. Figures 1(a) and 1(b) clearly bring out the effect of image resolution on the imaged appearance of hair. Both images belong to the same subject. Furthermore, video imagery sometimes contained artifacts due to motion blur.



(a) Hair patch from still image of stationary subject (100 dpi)



(b) Hair patch from video of moving subject (60 dpi)

Fig. 1. Imaged appearance of hair

Due to these and other issues, the use of hair as a biometric has been rare. To the best of our knowledge, the dissertation by Jia [2] was the first effort that reported an explicit attempt to use hair features, among many others, for enhancing face recognition. The author focused mostly on facial and front-view hair and concluded that the use of hair features is largely unreliable for the purposes of face recognition. We found only one other paper [3] with a focus on using hair as one of the primary means of person identification, which also chose an overhead view. However, the authors did

not treat hair in any special way. The overhead view of the head was simply a part of a *blob* delineated by background subtraction and containing shoulders and clothing in addition to hair. The hair boundary was not explicitly located, and the extracted features corresponded to the full extent of the blob and not just hair. The authors of [3] focused solely on still images of stationary subjects, whereas the work presented here has also been demonstrated successfully on low-resolution digital videos of subjects walking below an overhead video camera mounted on a doorway. Our techniques are therefore robust to the translational and rotational motion artifacts of the subject relative to the light source(s) and the camera within the framework of the described scenario. Another major difference is the availability and use of training images. The authors of [3] employed 60 training images for each of 12 subjects. For the given scenario, it is difficult, in general, to justify obtaining more than a few images of each person at the enrollment stage. In this work, we propose a scheme that works with just a single training image for each person and demonstrate reasonable accuracy. Subsequently, we present another scheme that averaged fewer than four training images for each of 30 subjects. For unconstrained head rotation in still imagery, even under these much more difficult conditions, our recognition rates were similar to those reported in [3].

To assist location of hair area and characterize hair features, the proposed approach is partly based on previously published work on texture analysis demonstrated successfully on other domains such as satellite imagery. Texture analysis for hair needs to be rotation invariant and multiresolution, since in a top-view scenario the distance (and therefore, magnification) of the hair from the camera changes according to the person's height. Therefore, of the several dozen texture characterization methods available in the published literature, we chose the multiresolution Rotation-Invariant Simultaneous AutoRegression (MR-RISAR) method proposed by Mao and Jain [4]. We also make use of a line-texture operator [5][6][7] that generates a binary mask to assert whether or not each pixel in the input image depicts a *line point*.

3 Approach

3.1 Hair Boundary Delineation

The number of skin pixels in the image is a measure of the degree of baldness and allows us to reliably detect subjects that are almost completely bald and to delineate their head areas. It is well known [8] that, regardless of race, human skin pixels in a color image taken under normal indoor/outdoor illumination fall into a relatively tight cluster in 3D RGB color space. We found that reasonably good skin detection usually results for the following values: $< \frac{I_R}{I_R+I_G+I_B} = 0.45 \pm 0.09 > < \frac{I_G}{I_R+I_G+I_B} = 0.33 \pm 0.05 >$, where I_R, I_G, I_B are intensities in the R, G, B color channels. Since skin is smooth, bright, and specular under illumination, we require that a skin pixel exceed a conservative intensity threshold of 100 (for intensity range 0 – 255) and that adjacent skin pixels not differ in intensity by more than 10%.

The hair region is seen to exhibit a very high density response to a line-texture operator previously discussed in [5][6][7]. The binary image generated as a result of the operator is then consolidated by the eight-connected grow/shrink sequence *GGSSSGG*. We then apply a binary mask boundary-tracing algorithm, which often produces an

overly detailed delineation of the desired hair region. This initial delineation is subsequently replaced by its convex hull. We observed that the largest circle that can be embedded inside the unsmoothed boundary is often a better choice as a region for extracting hair features because it reduces contamination by the background. Figure 2 shows examples of our hair delineation results.

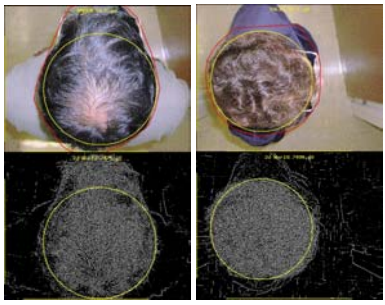


Fig. 2. Hair delineation results

3.2 Hair Feature Extraction

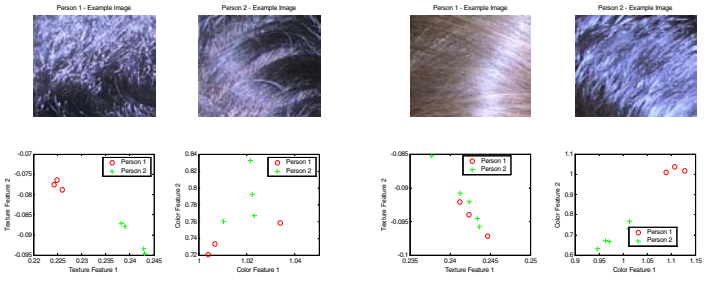
In the work reported here, we have investigated the use of two broad classes of features to characterize hair in general and hair texture in particular. The first class of features is based on sliding windows operating directly on a patch of pixels within delineated hair regions. These pixel-based features treat hair as a textured surface to be characterized via statistical texture segmentation operations. The second class of features exploits the fact that hair is a collection of individual linelike structures.

Pixel-Based Features. We direct the reader to the cited publication for details about the MR-RISAR model. We begin with a square patch of pixels located at the center of the delineated hair boundary. Rotationally invariant simultaneous auto-regressive coefficients, as defined by Mao and Jain, are then calculated to describe sliding square windows of width W over L levels of a Guassian pyramid constructed from the input image. In addition to these texture features, we compute two color features for each window, defined as $\bar{I_G}/\bar{I_R}$ and $\bar{I_B}/\bar{I_R}$ [9], where $\bar{I_R}$, $\bar{I_G}$, $\bar{I_B}$ are average intensities in the R, G, B color channels of the window, respectively. The sliding windows were made to overlap by half the window size along both X and Y directions.

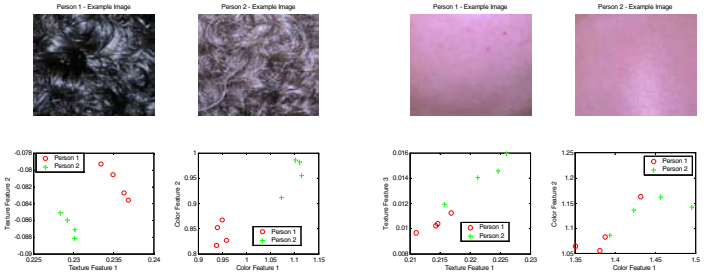
The size of the square patch, N, depends on the resolution of the input imagery, chosen such that the patch covers roughly 15% of the head surface of an average size head. Since the still images in our set were of higher resolution, a patch size of $N = 256$ and window size of $W = 32$ sufficed with $L = 3$. We chose $N = 128$ and $W = 16$ with $L = 2$ for the video experiments, due to the lower resolution. These choices led to 14 and 10 features for each sliding window in the still and video imagery, respectively. We thus have, for a given square patch of hair, a feature matrix with 14 or 10 columns.

The relative importance of color vs. texture features changes from person to person. Figures 3(a) and (b) show examples illustrating the effectiveness of our pixel-based

texture and color features individually. The top row of Figure 3(a) shows sample hair patches extracted from still images of two different subjects, both with dark hair. The bottom-left subgraph of that figure shows a plot of the first vs. the second *texture* feature extracted from the complete set of images available for these two subjects from our data set, including the patches shown in the top row of Figure 3(a). On the other hand, the bottom-right subgraph shows a plot of the first vs. the second *color* features extracted from the same set of images. Note that the X and Y coordinates of each point in these illustrations are the mean values of the corresponding features over a hair patch (i.e., the mean across the corresponding columns of the feature matrix for the patch). As intuitively expected, the two subjects are more separable in the texture space (left plot) than in the color space (right plot). Similarly, Figure 3(b) shows hair patches extracted from two different persons with straight hair of different colors. The separability in the color space (right plot) is clearly larger than the texture space (left plot), as intuitively expected.



(a) Effect of texture features (b) Effect of color features



(c) Interesting case: curly hair (d) Interesting case: bald head

Fig. 3. Pixel-based feature analysis

Figures 3(c) and (d) show two interesting cases that indicate the relative efficacy of the texture and color features and their intuitive interpretation. Two samples of subjects with curly hair are shown in Figure 3(c). Since the color difference is clearly significant, the color space (right plot) shows clear separability, as expected. Interestingly, the texture space (left plot) shows clear separability as well, evidently due to the differences in the density and/or the radius of curvature of the curls. Figure 3(d) shows

two sample patches extracted from two different completely bald subjects in our data set. Even though the skin color is quite similar, as evident from the lack of clear linear separability in the color space (right plot), the two subjects are linearly separable in the texture space (left plot). On close inspection, it was evident that one of the subjects was naturally bald (top right) but the other had shaved his head (top left), giving the shaved head a distinctive texture due to tiny wounds and hence causing the separability in the texture space.

Line-Based Features. To exploit the fact that hair is a collection of individual linelike structures, we have designed the following features in our current system:

1. *Macrotexture*: The macrotexture attributes are based on first extracting straight or curved line segments from the binary mask image described in Section 3.1, as shown in Figure 4(a). Each such segment is called a *hair-line segment*. Only the longer and smoother segments are selected (Figure 4(b)), and are later clipped to the region within the delineated hair boundary. We define two types of macrotexture attributes based, respectively, on the orientation and length of the detected hair-line segments.



(a) Line segment extraction (b) Delineated hair segments

Fig. 4. Hair segment delineation

(a) *Orientation*: We represent each hair-line-segment as a sequence of straight line subsegments and compute the orientation and direction for each subsegment. We then compute a histogram over the orientations of all such subsegments over the entire hair area, using 18 bins (each spanning 10 degrees). We then find the partition that maximizes the total number of entries in any nine adjacent bins (with wrap-around). The value we compute as the *orientation* metric is the ratio of the number of subsegments in such a maximum partition to the total number of subsegments. Intuitively, the *orientation* metric is a measure of the degree of order in the arrangement of hair. Curly or uncombed hair would have a value close to 0.5, whereas relatively straight combed hair would have a value close to 1.
(b) *Length*: We calculate a cumulative distribution function (cdf) of the lengths of the hair-line segments. The lengths corresponding to the cdf values of 0.1, 0.3, 0.6, and 0.9 are chosen as the four *length* metrics.

2. *Shape*: Our *shape* metrics are the total length of the hair boundary and the width-to-length ratio for its bounding minimum-width rectangle.
3. *Color*: We employ a color-labeling technique named *Relative Ordering*. It assigns one of 12 categorical labels to each pixel by applying the conditions listed in the rows of Table 1, going in the order 1 to 12 and stopping as soon as the first matching condition is found. We then assign categorical values to the most and second most frequently occurring label within the hair boundary.

Table 1. Relative ordering conditions

Order	RGB Condition	Label	Order	RGB Condition	Label
1	$I_R < T$ AND $I_G < T$ AND $I_B < T$, where $T = 30$	*dark*	7	$I_R > I_G > I_B$	*color4*
2	$I_R = I_G \pm 20$ AND $I_G = I_B \pm 20$ AND $I_B = I_R \pm 20$	*white*	8	$I_R > I_B > I_G$	*color5*
3	Skin determination	*skin*	9	$I_G > I_R > I_B$	*color6*
4	I_R*1.8 $< I_G$ OR I_B	*color1*	10	$I_G > I_B > I_R$	*color7*
5	I_G*1.8 $< I_R$ OR I_B	*color2*	11	$I_B > I_R > I_G$	*color8*
6	I_B*1.8 $< I_R$ OR I_G	*color3*	12	$I_B > I_G > I_R$	*color9*

In the discussion above, we have defined a number of pixel-based and line-based texture and color features. Given the small number of training samples for each subject, it is desirable to select only a subset of available features to avoid the curse of dimensionality. For the single-image enrollment scenario only the pixel-based features were used. When multiple images were available for training, line-based features were used whenever possible.

3.3 Recognition Procedure

Single Image Enrollment. For a given patch of hair, we compute a feature matrix, as discussed in Section 3.2. Assuming a multivariate Gaussian distribution, we then compute the mean feature vector μ and covariance matrix Σ. Given two patches of hair extracted from two different images, indexed as i and j, we estimate the likelihood that they belong to the same person as the following quantity:

$$p_{ij} = N_{\mu_j,\Sigma_j}(\mu_i) \times N_{\mu_i,\Sigma_i}(\mu_j), \tag{1}$$

where $N_{\mu,\Sigma}(x)$ is the multivariate Gaussian probability density evaluated at x, with mean μ and covariance matrix Σ. Given a patch of hair, i, the best match within a library of enrolled patches of hair with known identities is obtained as the patch j with the maximum likelihood p_{ij}, defined in Equation 1. Since the texture color features used are rotation invariant, this recognition strategy is inherently rotation invariant.

Multiple Image Enrollment. We have devised a recognition algorithm that operates in two phases: *ambiguity determination* and *ambiguity reduction*.

Phase 1: Ambiguity Determination: When multiple training images are available for each person, if the subjects are not completely or nearly bald, we characterize the hair first by the line-based features described in Section 3.2. For each enrolled subject, we first determine the feature interval between the minimum and maximum observed feature values. This interval is expanded by a *feature tolerance* factor proportional to its width (typically 25% in our experiments) along both ends. These expanded intervals for an enrolled subject now act as a filter to decide if a test image with unknown identity is sufficiently similar to the enrolled subject. If a feature value of the test subject lies outside these intervals for an enrolled subject, the corresponding feature filter is considered to fail. For multiple features, each feature interval acts independently to filter a given set of possible identities for the test subject (input ambiguity set) to a smaller or equal-size set (output ambiguity set). When the unknown subject is determined to be completely or nearly bald, we compose the ambiguity set as a collection of all the subjects in the enrolled images who are also completely or nearly bald.

Phase 2: Ambiguity Reduction: When the unknown subject has not been determined to be completely or nearly bald, the feature tolerances for the members of the ambiguity set are reduced to zero, thereby making the filter matching criteria more strict. We also introduce one additional filter for hair boundary shape similarity. We compare the first two members of the ambiguity set, using the extent of similarity with the test image defined as the total number of interval-based filters that succeed in matching. If any feature value for the test subject falls outside the feature intervals of both members, and does not lie in the region between the two intervals, the filter with an interval-endpoint closest to the observed feature value is said to succeed. The member with the lower extent of similarity is removed from contention. The surviving member is then compared with the next remaining member (if any) of the ambiguity set, and the procedure is repeated until the ambiguity list is exhausted.

Because of the possibility of a tie, the result produced in this phase could still be an ambiguity set of length greater than 1. In such a case, or when the test subject and therefore members of the ambiguity set are completely bald, the means of pixel-based features are used in the above interval-based recognition scheme (with feature tolerance zero) to select a single best hypothesis. There is also the possibility, in either of the two phases, that no entries in the ambiguity set are similar enough to be acceptable. In this case the test image is *rejected*.

4 Results

4.1 Data Collection

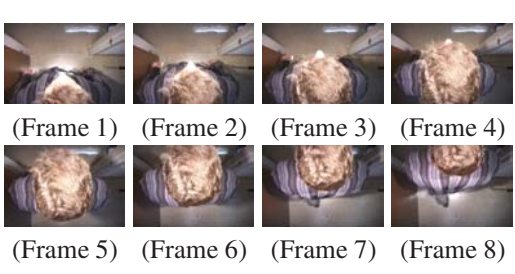
Still Image Data. We have acquired 116 top-view color (1600×1200) JPEG images from 30 subjects (Figure 5), which translated to approximately 250 μm per pixel on the head surface. Two to five images of each subject were collected, each image with a different, unconstrained head orientation in the horizontal plane. Head rotation in the vertical plane was restricted to roughly ± 10 degrees, based on the assumption that the subjects are more likely to rotate their heads sideways than straight up or down when walking along a hallway, turning a corner, or entering a room. Depending on the position and head orientation of the subject, different images for the same subject



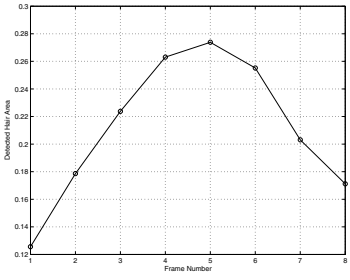
Fig. 5. Hair samples of stationary subjects (30 subjects; 116 images)

had differing extents and positions of imaged hair specularity and highlights. The time interval between different captured images of the same subject ranged roughly between one and five minutes.

Video Data. In roughly the same setup as the still image capture, we captured a few seconds of video data for each subject entering and leaving a room at 15 frames per second progressive scan. Given the close proximity of the camera to an average person's head, the field of view of the camera is narrow at head level. As a result, at average walking speeds, the entire head is visible only for a split second (i.e., only 1 or 2 frames for most people) (Figure 6(a)). The frame resolution was 720×480, which translated to approximately 250 μm per pixel on the average head surface. As before, we allowed unconstrained head orientation in the horizontal plane and restricted the head rotation in the vertical plane to roughly ± 10 degrees. Our pool of subjects included 27 individuals.



(a) Moving subject with 15 fps progressive scan video capture



(b) Frame selection from video

Fig. 6. Video data capture and frame selection

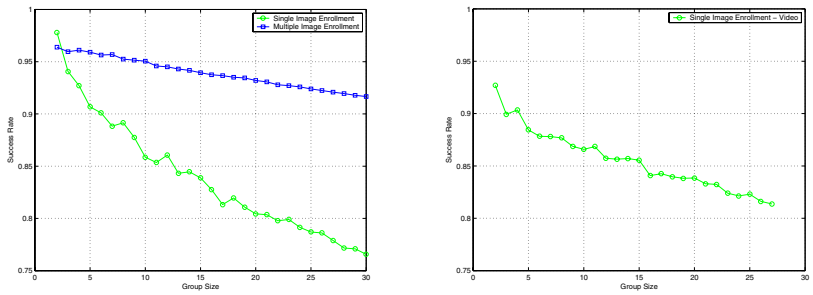
4.2 Experimentation

Authentication with High-Resolution Still Imagery. The present technique is envisaged to be used in situations involving small groups of people. It is intuitive that the recognition accuracy would reduce as the number of persons in the group increases.

Our experiments gradually increased the size of the group from two to a maximum of 30 persons per group. For each group size, individuals were randomly selected from our still image corpus of 30 persons. For the single image enrollment scenario, for each subject, a training image and a different test image was randomly chosen from the set of images available for that person. For the multiple-image enrollment scenario, each image of each selected subject was classified using the rest of the images of the selected subjects for training (i.e., according to the *leave-one-out* paradigm). The success rate (defined as the ratio of the number of correct identifications to the total number of test images) was recorded. For each group size, the random selection of members and the evaluation of the success rate was repeated 1000 times, so that the average success rate over the 1000 experiments is representative for that group size in a Monte Carlo sense.

For single image enrollment, our recognition strategy does not have a *reject* or *don't know* category. Therefore, the reject rate is constant at 0. As shown in Figure 7(a) (green curve), the success rate is fairly high for group sizes ≤ 5, given that only one image was used for training for each person. The average success rate drops as the size of the group increases, as expected: 91% for a group size of 5, 86% for a group size of 12, and 77% for a group size of 30.

For multiple-image enrollment, our reject rate was more or less independent of the group size, at approximately 3%. The error rate increased with the group size, as before. Overall, the average success rates were 96% for a group size of 5, 95% for a group size of 12, and 92% for a group size of 30 (Figure 7(a), blue curve). As can be observed, the success rates were higher when multiple images were available for training.



(a) Stationary Subjects and Still Imagery (b) Moving Subjects and Video Imagery

Fig. 7. Person authentication results using hair

Authentication with Low Resolution Video. The person authentication problem using video is in many ways more difficult than authentication using still imagery due to lower resolution, lower color sensitivity, and motion blur. Our experimental setup involving video capture sometimes resulted in only a couple of full head views at average walking speeds. For the purposes of the experimental results presented here, we restricted the video analysis to single image enrollment-based person authentication. As the subject entered the data capture room, of the several frames in which the person was visible in the captured video, our video analysis algorithm chose a single frame in which the area of the detected hair region was maximum. For example, Figure 6(b) shows the extent

of detected hair areas in the eight frames shown in Figure 6(a). The X axis represents the frame index, and the Y axis represents the ratio of the detected hair area to the area of the frame. The plot shows a clear peak at frame 5, where the detected hair area is maximum. This frame is then chosen to enroll the subject using the single image enrollment scheme. Analogously, as the subject left the room, another single frame was similarly chosen for testing. Using the framework for simulating smaller group sizes described above, we estimated the recognition accuracy for several subgroups of subjects.

The error rate increased with the group size, as before. Overall, the average success rates were 89% for a group size of 5, 86% for a group size of 12, and 81% (extrapolated) for a group size of 30 (Figure 7(b)). The performance is therefore comparable to one reported with high resolution still imagery, thereby attesting to the robustness of the feature extraction and analysis steps.

5 Conclusion

One of the main objectives of this work was to select a relevant scenario and develop an experimental system to demonstrate and quantify the utility of using images and videos of hair for automated person recognition. We are not aware of any published work in which characteristics of human hair were deliberately selected to play a central or critical role in automatically establishing human ID from color images. The selected scenario was the reacquisition of the identities of people being tracked in an indoor environment after a short interruption in tracking.

Subsequently, this work successfully defined a set of features for pixel-based and line-based characterizations of the imaged appearance of human hair. These features can be used for person recognition in the above (or similar) setting. For pixel-based characterization, we used rotation invariant statistical features to determine texture and color. For line-based characterization, we developed many different types of attributes including shape, line texture and orientation, and color. Taking into account the small number of images that can reasonably be assumed to be available for training purposes for the given scenario, we described two reliable decision procedures for single and multiple image enrollments.

Overall, for still images, our performance is in the same range as that reported by Cohen et al. [3]: 96% success rate, 4% error rate, and 0% reject rate for 12 subjects, compared with our result of 95% success rate, 2% error rate, and 3% reject rate. However, the results reported here were obtained with only a single or up to 4 training images per test subject, whereas Cohen et al. used 60 training samples per subject. In contrast to the proposed approach, an explicit delineation and characterization of hair did not play a central role in the cited previous work [3], which included the color and texture of clothing for identification. The hair characterization procedure described here was demonstrated to be capable of handling completely or nearly bald subjects, whereas Cohen et al. make no such claim. Our method has also been illustrated on subjects in motion as captured in video clips, whereas Cohen et al. focused only on still images.

In summary, one of the primary contributions of this work is our successful demonstration of how hair, far from being a recognition impediment, can be an important asset

in person identification. In this work, we have introduced a collection of pixel-based and line-based attributes, and methods for their measurements, that may have general utility beyond hair characterization. Our decision procedures can be parameterized with little image information, and are effective in exploiting observed interactions between individual objects and the feature extraction algorithms operating on these objects. Our ongoing work in this research direction includes the development of a robust hair detection technique that loosens the constraints imposed by context, and the use of hair for human ID from a more general perspective than the overhead viewpoint we described here.

References

1. J. Gray, *The World of Hair: A Scientific Companion*, Delmar Learning, 1997.
2. X. Jia, *Extending the Feature Set for Automatic Face Recognition*, Ph.D. thesis, University of Southhampton, 1993.
3. I. Cohen, A. Garg, and T. Huang, "Vision-based overhead view person recognition," in *Proceedings of the ICPR*, 2000, pp. 1119–1124.
4. J. Mao and A.K. Jain, "Texture classification and segmentation using multiresolution simultaneous autoregressive models," *Pattern Recognition*, vol. 25, no. 2, pp. 173–188, 1992.
5. M.A. Fischler and H.C. Wolf, "Linear delineation," in *Proceedings of IEEE CVPR*, 1983, pp. 351–356.
6. M. A. Fischler and H. C. Wolf, "Locating perceptually salient points on planar curves," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 113–129, 1994.
7. M.A. Fischler and A.J. Heller, "Automated techniques for road network modeling," in *DARPA Image Understanding Workshop*, 1998, pp. 501–516.
8. T. Rzeszewski, "A novel automatic hue control system," *IEEE Transactions on Consumer Electronics*, vol. CE-21, pp. 155–162, 1975.
9. A. Khotanzad and O.J. Hernandez, "Color image retrieval using multispectral random field texture model and color content features," *Pattern Recognition*, vol. 36, no. 8, pp. 1679–1694, 2003.

A Survey of 3D Face Recognition Methods

Alize Scheenstra[1], Arnout Ruifrok[2], and Remco C. Veltkamp[1]

[1] Utrecht University, Institute of Information and Computing Sciences,
Padualaan 14, 3584 CH Utrecht, The Netherlands
alize.scheenstra@tiscali.nl, remco.veltkamp@cs.uu.nl
[2] Netherlands Forensic Institute,
Laan van Ypenburg 6, 2497 GB Den Haag, The Netherlands,
arnout@holmes.nl

Abstract. Many researches in face recognition have been dealing with the challenge of the great variability in head pose, lighting intensity and direction,facial expression, and aging. The main purpose of this overview is to describe the recent 3D face recognition algorithms. The last few years more and more 2D face recognition algorithms are improved and tested on less than perfect images. However, 3D models hold more information of the face, like surface information, that can be used for face recognition or subject discrimination. Another major advantage is that 3D face recognition is pose invariant. A disadvantage of most presented 3D face recognition methods is that they still treat the human face as a rigid object. This means that the methods aren't capable of handling facial expressions. Although 2D face recognition still seems to outperform the 3D face recognition methods, it is expected that this will change in the near future.

1 Introduction

One of the earliest face recognition methods was presented in 1966 by Bledsoe [1]. In one of his papers [2], Bledsoe described the difficulties of the face recognition problem:

> "This recognition problem is made difficult by the great variability in head rotation and tilt, lighting intensity and angle, facial expression, aging, etc. Some other attempts at facial recognition by machine have allowed for little or no variability in these quantities. Yet the method of correlation (or pattern matching) of unprocessed optical data, which is often used by some researchers, is certain to fail in cases where the variability is great. In particular, the correlation is very low between two pictures of the same person with two different head rotations."

Since that time many researches have been dealing with this subject and have been trying to find an optimal face recognition method. The main purpose of this overview is to describe the recent face recognition algorithms on still images. Previous face recognition surveys were presented by Samal and Iyengar [3],

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 891–899, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Chellappa et al. [4] and Zhao et al. [5]. However, they all are primarily focussed on 2D face recognition. In the Vendor Test 2002 the performance of different commercial face recognition methods were compared [6]. Most commercial face recognition systems use one or more algorithms as presented in the literature. However, all systems conceal which algorithms are used in their application. Therefore, commercial systems are excluded in this survey. The last few years 3D facial models can be more easily acquired since the acquisition techniques have improved. Therefore, some face recognition methods originally developed for 2D face recognition have been extended for 3-dimensional purposes. Using 3D models one can deal with one main problem in 2D face recognition: the influence of the pose of the head. Also the surface curvature of the head can now be used to describe a face. A recent survey of 3D face recognition was recently presented by Bowyer [10]. Since that time new results with respect to 3D face recognition have been published. We describe the most recent approaches to the facial recognition challenge.

2 3D Supported 2D Models

Zhao and Chellappa proposed in [7] a shape-from-shading (SFS) method for preprocessing of 2D images. This SFS-based method used a depth map for generating synthetic frontal images. The The Linear Disrciminant Analysis (LDA) was applied to the synthetic images instead of the original images. The recognition rate increased with 4% when the synthetic images were used for LDA coding instead of the original images. Hu et al. proposed to use one neutral frontal image to first create a 3D model and from that create synthetic images under different poses, illuminations and expressions [8]. By applying LDA or Principal Component Analysis (PCA) to this 3D model instead of the 2D face images, the recognition rate increased with an average of 10% for the half-profile images. A similar idea was proposed earlier by Lee and Ranganath where they presented a combination of an edge model and color region model for face recognition after the synthetic images were created by a deformable 3D model [9]. Their method was tested on a dataset with 15 subjects and reached a recognition rate of 92.3% when 10 synthetic images per subject were used and 26.2% if one image for each subject was used.

3 Surface Based Approaches

3.1 Local Methods

Suikerbuik [11] proposed to use Gaussian curvatures to find 5 landmarks in a 3D model. He could find the correct landmark point with a maximal error of 4 mm . Gordon proposed to use the Gaussian and cean curvature combined with depth maps to extract the regions of the eyes and the nose. He matched these regions to each other and reached a recognition rate of 97% on a dataset of 24 subjects [12]. Moreno et al. used both median and Gaussian curvature for the

selection of 35 features in the face describing the nose and eye region [13]. The best recognition rate was reached on neutral faces with a recognition rate of 78%.

Xu et al. proposed to use Gaussian-Hermite moments as local descriptors combined with a global mesh [14]. Their approach reached a recognition rate of 92% when tested on a dataset of 30 subjects. When the dataset was increased to 120 subjects, the recognition rate decreased to 69%.

Chua et al. [15, 16] introduced point signatures to describe the 3D landmark. They used point signatures to describe the forehead, nose and eyes. Their method reached a recognition rate of 100% when tested on a dataset with 6 subjects. Wang et al. used the point signatures to describe local points on a face (landmarks). They tested their method on a dataset of 50 subjects and compared their results with the Gabor wavelet approach [17]. Their results showed that point signatures alone reached a recognition rate of 85% where the Gabor wavelets reached a recognition rate of 87%. If both 2D and 3D landmarks were combined, they reached a recognition rate of 89%. The authors remarked that these results could also be influenced by the number of landmarks used for face recognition, since for the point signatures 4 landmarks were used, for the Gabor wavelets 6 landmarks and for the combination of both 12 landmarks were used.

Douros and Buxton proposed the Gaussian Curvature to define quadratic patches to extract significant areas of the body. They claim that their method can be used for recognition of all kinds of 3D models [18]. Another local shape descriptor that was found to perform good on human bodies was the Paquet Shape Descriptor [19].

3.2 Global Methods

One global method on curvature was lately presented by Wong et al. [20]. The surface of a facial model was represented by an Extended Gaussian Image (EGI) to reduce the 3D face recognition problem to a 2D histogram comparison. The proposed measure was the multiple conditional probability mass function classifier (MCPMFC). Tested on a dataset of 5 subjects the MCPMFC has a recognition rate of 80.08% where a minimum distance classifier (MDC) reached a recognition rate of 67.40%. However a test on synthetic data showed that for both methods the recognition rate decreased with 10% when the dataset was increased from 6 subjects to 21 subjects.

Papatheodorou and Rueckert proposed to use a combination of a 3D model and the texture of a face [21]. They also proposed some similarity measures for rigid alignment of two faces for 3D models and for 3D models combined with the texture. Their results showed an increase for frontal images when adding a texture to the model.

Beumier and Acheroy proposed to use vertical profiles of 3D models for face recognition. Their first attempt was based on three profiles of one face and had an error rate of 9.0% when it was tested on a dataset of 30 subjects [22]. In their second attempt they added grey value information to the matching process [23]. This attempt reduced the error rate to 2.5% when it was tested on the

same database. Wu et al. proposed to perform 3D face recognition by extracting multiple horizontal profiles from the 3D model [24]. By matching these profiles to each other they reached an error rate between 1% and 5.5% tested on a database with 30 subjects.

4 Template Matching Approaches

Blanz, Vetter and Romdhani proposed to use a 3D morphable model for face recognition on 2D images [25–27]. With this method tested on a dataset of 68 subjects they reached a recognition rate of 99.8% for neutral frontal images and a recognition rate of 89% for profile images. Huang et al. added a component based approach to the morphable model [29] based on the approach of Heisele [28]. However, the recognition rate was for all approaches of the morphable model between the 75% and the 99%.

Naftel et al. presented a method for automatically detecting landmarks in 3D models by using a stereo camera [30]. The landmarks were found on the 2D images by an ASM model. These landmark points were transformed to the 3D model by the stereo camera algorithm. This algorithm was correct in 80% of all cases when tested on a dataset of 25 subjects.

A similar idea was proposed by Ansari and Abdel-Mottaleb [31]. They used the CANDIDE-3 model [32] for face recognition. Based on a stereo images landmark points around the eyes, nose and mouth were extracted from the 2D images and converted to 3D landmark points. A 3D model was created by transforming the CANDIDE-3 generic face to match the landmark points. The eyes, nose and mouth of the 3D model were separately matched during the face recognition. Their method achieved a recognition rate of 96.2% using a database of 26 subjects. Lu et al. had used the generic head from Terzopoulos and Waters [33] which they adapted for each subject based on manually placed feature points in the facial image [34]. Afterwards the models were matched based on PCA. This method was tested on frontal images and returns in 97% of all cases the correct face within the best 5 matches.

5 Other Approaches

The original principal component method for 3D facial models was implemented by Mavridis et al. for the European project HiScore [35]. Chang et al. had compared the performance of 3D eigenfaces and 2D eigenfaces of neutral frontal faces on a dataset of 166 subjects [36]. They found no real difference in performance for the 2D eigenfaces and 3D eigenfaces. However, a combination of both dimensionalities scored best of all with a recognition rate of 82.8%. Xu et al. proposed to use 3D eigenfaces with nearest neighbor and k-nearest neighbors as classifiers [37]. Their approach reached a recognition rate around the 70% when tested on a dataset of 120 subjects.

Bronstein et al. had proposed to transform the 3D models to a canonical form before applying the eigenface method to it [38]. They claimed that their

method could discriminate between identical twins and was insensitive for facial expressions, although no recognition rates were given.

Tsalakanidou et al. proposed to combine depth maps with intensity images. In their first attempt they used eigenfaces for the face recognition and his results showed a recognition rate of 99% for a combination of both on a database of 40 subjects [39]. In a second attempt embedded hidden markov models were used instead of eigenfaces to combine the depth images and intensity images [40]. This approach had an error rate between the 7 % and 9%.

6 Discussion and Conclusion

It is hard to compare the results of different methods to each other since the experiments presented in literature are mostly performed under different conditions on different sized datasets. For example one method was tested on neutral frontal images and had a high recognition rate, while another method was tested on noisy images with different facial expressions or head poses and had a low error rate.

Some authors presented combinations of different approaches for a face recognition method and these performed all a little better than the separate methods. But besides recognition rate, the error rate and computational costs are important, too. If the error rate decreases significantly, while the recognition rate increases only a little bit, the combined method is still preferred. But, if the computational costs increase a lot, calculation times could become prohibitive for practical applications.

Most interesting for this survey were the studies that presented method comparisons, like [41–43]. Phillips et al. [6] present comparison studies performed on the FERET database. The latest FERET test performed on different algorithms was presented in 2000 [44]. An important conclusion from this survey was that the recognition rates of all methods improved over the years. The dynamic graph matching approach of Wistkott et al. [17] had the best overall performance on identification. For face verification the combination of PCA and LDA presented by Zhao et al. [46] performed best.

In table 1 a summary is given for the most important and successful 2D and 3D face recognition methods. One can see that the 3D face recognition approaches are still tested on very small datasets. However, the datasets are increasing during the years since better acquisition materials become available. By increasing a dataset, however, the recognition rate will decrease. So the algorithms must be adjusted and improved before they will be able to handle large datasets with the same recognition performance. Another disadvantage of most presented 3D face recognition methods is that most algorithms still treat the human face as a rigid object. This means that the methods aren't capable of handling facial expressions. In contrast to 3D face recognition algorithms, most 2D face recognition algorithms are already tested on large datasets and are able to handle the size of the data tolerable well. The last few years more and more 2D face recognition algorithms are improved and tested on less perfect images,

Table 1. A summary on most important presented 2D and 3D face recognition methods. The variation in images column shows if images in de dataset were taken under different conditions, like facial expression, illumination, head pose et cetera

method	modality	reference	number of subjects in dataset	images per subject per subject	rank one recognition performance in %	error rate in %	variation in images
baseline PCA	2D	[44]	1196	1	79	?	no
baseline LDA	2D	[42]	200	at least 4	97.5	1.48	yes
baseline Correlation	2D	[44]	1196	1	82	?	no
PCA-LDA	2D	[5]	298	1	95	?	yes
Bayesian PCA	2D	[47]	1196	1	95	?	no
ASM-AAM	2D	[48]	30	10	92	?	no
ASM-AAM	2D	[48]	30	3	48	?	yes
Face Bunch Graph	2D	[17]	300	1	97	?	no
Face Bunch Graph	2D	[17]	250	1	84	?	yes
Infra-Red images	2D	[49]	200	7	89	?	no
Gaussian images	3D	[12]	8	3	100	?	no
Point Signatures	3D	[16]	6	1	100	?	no
Extended Gaussian Images	3D	[20]	5	1	80	?	no
Profiles	3D	[23]	26	1	?	2.5	no
Morphable model	3D	[25]	68	21	99.8	?	no
Morphable model	3D	[25]	68	21	89	?	yes
3D eigenfaces	3D	[36]	166	2	83.7	?	no
3D eigenfaces	3D	[37]	120	1	71.1	5	yes
Canonical forms	3D	[38]	157	1	?	?	yes

like noisy images, half profile images, occlusion images, images with different facial expressions, et cetera. Although not single algorithm can be assumed to handle the difficult images good enough, an increasing line in performance can be found.

Although 2D face recognition still seems to outperform the 3D face recognition methods, it is expected that in the near future 3D face recognition methods outperform 2D methods. 3D models hold more information of the face, like surface information, that can be used for face recognition or subject discrimination. Another major advantage is that 3D face recognition is pose invariant. Therefore, 3D face recognition is still a challenging but very promising research area.

References

1. Bledsoe, W.W.: The Model Method in Facial Recognition. Technical Report PRI-15. Panoramic Recsearch Inc. California (1966)
2. Bledsoe,W.W.: Man-Machine Facial Recognition: Report on a Large-Scale Experiment. Technical Report PRI-22. Panoramic Recsearch Inc. California (1966)
3. Samal, A., Iyengar,P.A.: Automatic Recognition and Analysis of Human Faces and Facial Expressions. Pattern Recognition. 25 (1992) 65–77
4. Chellappa, R., Wilson, C.L., Sirohey. S.: Human and Machine Recognition of Faces: A survey. Proceedings of the IEEE, 83 (1995) 705–740
5. Zhao, W.Y., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Surveys. 35 (2003) 399–458
6. Phillips, P.J., Grother,R., Micheals, R.J., Blackburn, D.M., Tabassi, E., Bone, M.: Face Recognition Vendor test 2002. "http://www.frvt.org/" (December 2004)
7. Zhao,W.Y., Chellappa,R.: SFS Based View Synthesis for Robust Face Recognition. Proceedings of the IEEE International Automatic Face and Gesture Recognition (2000) 285–292
8. Hu,Y., Jiang, D., Yan, S.,Zhang, L., Zhang, H.: Automatic 3D reconstruction for Face Recognition. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. (2000) 843–850
9. Lee, C.H., Park, S.W.,Chang, W., Park, J.W.: Improving the performance of Multi-Class SVMs in face recognition with nearest neighbour rule. Proceedings of the IEEE International Conference on Tools with Artificial Intelligence. (2003) 411–415
10. Bowyer, K.W., Chang, K., Flynn, P.J.: A survey of approaches to three-dimensional face recognition. International Conference of Pattern Recognition, Vol I. (2004) 358–361
11. Suikerbuik, C.A.M., Tangelder, J.W.H., Daanen, H.A.M., Oudenhuijzen, A.J.K.: Automatic feature detection in 3D human body scans. Proceedings of the conference "SAE Digital Human Modelling for Design and Engineering. (2004)
12. Gordon, G.G.: Face Recognition Based on Depth Maps and Surface Curvature. Proceedings of the SPIE, Geometric Methods in Computer Vision, Vol. 1570. (1991) 108–110
13. Moreno, A.B., Sánchez, A., Vélez, J.F., Díaz, F.J.: Face Recognition using 3D Surface-Extracted Descriptors. Proceedings of the Irish Machine Vision and Image Processing Conference. (2003)

14. Xu, C., Wang, Y., Tan, T., Quan, L.: Automatic 3D Face recognition combining global geometric features with local shape variation information. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. (2004) 302–307
15. Chua, C.S., Jarvis, R.: Point Signatures: A New Representation for 3D Object Recognition. International Journal on Cumputer Vision. 25 (1997) 63–85
16. Chua, C.S., Han, F., Ho, Y.K.: 3D human face recognition using point signature. Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. (2000) 233–238
17. Wiskott, L., Fellous, J.M., kruger, N., van der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence. 19 (1997) 775–779
18. Douros, I., Buxton, B.F.: Three-Dimensional Surface Curvature Estimation using Quadric Surface Patches. Proceedings of the Scanning 2002 Conference. (2002)
19. Robinette, K.M., An Alternative 3D descriptor for database mining. Proceedings of the Digital Human Modelling Conference. (2004)
20. Wong, H.S., Chueng, K.K.T., Ip, H.H.S.: 3D head model classification by evolutionary optimization of the extended Gaussian image representation. Pattern Recognition. 37 (2004) 2307–2322
21. Papatheodorou, T., Rueckert, D.: Evaluation of automatic 4D Face recognition using surface and texture registration. Proceedings of the IEEE International Conference in Automatic Face and Gesture Recognition. (2004) 321–326
22. Beumier, C., Acheroy, M.: Automatic 3D face authentication. Image and Vision Computing. 18 (2000) 315–321
23. Beumier, C., Acheroy, M.: Face verification from 3D and grey level clues. Pattern Recognition Letters. 22 (2001) 1321–1329
24. Wu, Y., Pan, G., Wu, Z.: Face Authentication Based on Multiple Profiles Extracted from range data. Proceedings of the Audio- and Video-Based Biometric Person Authentication, Vol. 2688. (2003) 515–522
25. Blanz, V., Romdhani, S., Vetter, T.: Face Identification across different poses and illuminations with a 3D morphable model. Proceedings of the IEEE International Automatic Face and Gesture Recognition. (2002)
26. Blanz, V., Vetter, T.: Face Recognition Based on Fitting a 3D Morphable Model. IEEE Transactions on Pattern Analysis and Machine Intelligence. 25 (2003)
27. Romdhani, S., Vetter, T.: Efficient, Robust and Accurate Fitting of a 3D Morphable Model. Proceedings of the European Conference on Computer Vision. (2003)
28. Heisele, B., Koshizen, T.: Components for Face Recognition. Proceedings of the Audio- and Video-Based Biometric Person Authentication, Vol 2688. (2003) 153–159
29. Huang, J., Heisele, B., Blanz, V.: Component-Based Face Recognition with 3D Morphable Models. Proceedings of the Audio- and Video-Based Biometric Person Authentication, Vol 2688. (2003) 27–34
30. Naftal, A.J., Mao, Z., Trenouth, M.J.: Stereo-assisted landmark detection for the analysis of 3D facial shape changes. Technical Report TRS-2002-007. Deprtment of Computation UMIST, Manchester. (2002)
31. Ansari, A., Abdel-Mottaleb, M.: 3D Face modeling using two views and a generic face model with application to 3D face recognition. Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance. (2003) 37–44
32. Ahlberg, J.: CANDIDE-3 - an updated parameterized face. Technical Report LiTH-ISY-R-2326. Dept. of Electrical Engineering, Linköping University. (2001)

33. Terzopoulos, D., Waters, K.: Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 15 (1993) 569–579
34. Lu, X., Hsu, R., Jain, A., Kamgar-Parsi, B.: Face Recognition with 3D Model-Based Synthesis. Proceedings of the International Conference on Biometric Authentication. (2004) 139–146
35. Mavridis, N., Tsalakanidou, F., Pantazis, D., Malasiotis, S., Strintzis, M.: The HISCORE face recognition application: Affordable desktop face recognition based on a novel 3D camera. Proceedings of the International Conference on Augmented Virtual Environments and 3D Images. (2001) URL: http://uranus.ee.auth.gr/hiscore
36. Chang, K.I., Bowyer, K.W., Flynn, P.J.: Multi-Modal 2D and 3D Biometrics for Face Recognition. Proceedings of the IEEE International Workshop on Analysis and Modeling Journal on Computer Vision. (2003)
37. Xu, C., Wang, Y., Tan, T., Quan, L.: A New Attempt to Face Recognition Using Eigenfaces. Proceedings of the Asian Conference on Computer Vision. 2 (2004) 884–889
38. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Expression-Invariant 3D Face Recognition. Audio- and Video-Based Biometric Person Authentification, Vol 2688. (2003) 62–70
39. Tsalakanidou, F., Tzovaras, D., Strintzis, M.G.: Use of depth and colour eigenfaces for face recognition. Pattern Recognition Letters. 24 (2003) 1427–1435
40. Tsalakanidou, F., Malassiotis, S., Strintzis, M.G.: Integration of 2D and 3D images for Enhanced Face Authentication. Proceedings of the IEEE International Conference in Automatic Face and Gesture Recognition. (2004) 266–271
41. Navarrete, P., Ruiz-del-Solar, J.: Comparative Study between different eigenspace-based approaches for face recognition. AFSS International Conference on Fuzzy Systems. (2002) 178–187
42. Jonsson, K., Kittler, J., Matas, J.: Support Vector Machines for face authentication. Image and Vision Computing. 20 (2002) 369–375
43. Sadeghi, M., Kittler, J., Kostin, A., Messer, K.: A comparative study of automatic face verification algorithms on the BANCA database. Proceedings of the Audio- and Video-Based Biometric Person Authentication, Vol 2688. (2003) 35–43
44. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence. 22 (2000) 1090–1104
45. Wiskott, L.: Phantom Faces for Face Analysis. Proceedings of the Joint Symposium on Neural Computation. (1996) 46–52
46. Zhao, W.Y., Krishnaswamy, A., Chellappa, R.: Discriminant Analysis of Principal Components for Face Recognition. Proceedings of the International Conference on Automatic Face and Gesture Recognition. (1998) 336–341
47. Moghaddam, B., Jebera, T., Pentland, A.P.: Bayesian face recognition. Pattern Recognition. 33 (2000) 1771–1782
48. Lanitis, A., Taylor, C.J., Cootes, T.F.: An automatic face identification system using flexible appearance models. Image and Vision Computing. 13 (1995) 393–401
49. Pan, Z., Healey, G., Prasad, M., Tromberg, B.: Face Recognition in Hyperspectral Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 25 (2003) 1552–1560

Influences of Image Disturbances on 2D Face Recognition

Henning Daum

Fraunhofer-Institute for Computer Graphics
daum@igd.fhg.de

Abstract. In the current development of deploying biometric methods, face recognition takes a leading role. Therefore it is of interest what barriers might arise in using face recognition for certain applications. Because of the wide range of possible uses of biometrics and face recognition a technology evaluation on which image disturbances influence face recognition algorithms was done in the project BioFace III. This includes tests regarding head rotation, contrast, compression and more.

1 Introduction

Face recognition takes a special role among the biometric methods, as the face is one of the few biometric features that can be acquired relatively unnoticed. Therefore it can be used for covered surveillance. Furthermore a face picture is very common in passport documents. Connected to this the face picture features another advantage: "Manual" comparison. In case the system fails a human guard can at least compare a face picture to a person while this is not feasible for e.g. fingerprints or iris.

The wide range of possible applications brings a wide range of possible capturing scenarios: Pictures could be unsharp, the scene could be illuminated in a bad way, the angle of view is not frontal[1], etc. All of these disturbances on the image quality and the face pictured can influence the comparison result of a face recognition algorithm. In the worst case the degradation is so severe that the captured images cannot be used for biometric face recognition and therefore the targeted application scenario doesn't work.

Several disturbances (e.g. head rotations) are commonly known to degrade face recognition performance but a number of disturbances – especially in the acquisition area – remained unexamined.

2 Disturbances

The sources of disturbances on image material can be roughly grouped into two classes. Firstly the person being the motif can introduce disturbances into the

[1] Probably this should read "optimal", but for common face recognition algorithms the frontal view is the optimal view.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 900–908, 2005.
© Springer-Verlag Berlin Heidelberg 2005

picture, e.g. grinning or wearing a hat. The second step is the photographic process itself, e.g. focal settings, shutter. As the face recognition software needs a digital image the conversion process from a printed copy to a data file can also introduce degradations into the image, like noise, dust or a tint.

2.1 Motif Originated Disturbances

Motif based Disturbances are disturbances that arise from the projection of the three-dimensional head to the two-dimensional image. In addition to that the person being pictured can have a non-standard mimic or wear disturbing clothes.

Rotations. The head is a three-dimensional object, that can be rotated by three axis. The first axis passes through the head from neck upward, a rotation around this axis would roughly represent a non-verbal "No". The second axis can be imagined from ear to ear, the corresponding rotation being a non-verbal "Yes". The third and last axis will enter the head at the nose and exit at the back of the head, a rotation will tilt the head to the side.

Rotations around the first two axis (neck and ears) will after a certain degree result in a loss of face area to start getting invisible and therefore a loss in information that can not be recovered by the recognition algorithm. It is obvious that the rotation around the neck axis is far more common and wide angles will arise more often.

A rotation around the nose axis by comparison can be corrected nearly without any loss. Here a rotation greater that twenty degree will be encountered very rarely.

Mimic. The mimic of the pictured person during the acquisition can decrease the recognition performance considerably. Especially the mouth area is affected by mimic changes. But also the eyes can cause a comparison to fail, as these are often used for face localization, if the person closes the eyes this can fail, causing the whole comparison to fail.

Classification of mimic originated disturbances are difficult. Measuring the strength of a smile is hard for one person and even more complicated for different persons. This is why for disturbances caused by mimic only the existence or absence can be determined.

2.2 Acquisition Originated Disturbances

This group of disturbances arise from errors in lighting, camera settings, wrong handling of the camera, bad or wrong film material etc. These disturbances cause a bad reproduction in a technically (and often also optically) view.

Over-/Underexposure. Main reason for an over- or underexposure is a too long or too short exposure time of the film material or the digital camera CCD. An exposure error is often correlated to a contrast error.

Contrast. The definition of contrast in photography is typically the difference in brightness between bright and dark areas of the image. As these are important for segmentation and edge detection the contrast of an image can affect the performance of a recognition algorithm. In the digital world this is even worse as there is a upper and lower border for the brightness of pixels. If the contrast is raised or lowered extremely, all brightness values exceeding the border will be mapped to the border, resulting in loss of information.

Sharpness. In the optimal case a punctiform item in the motif will appear punctiform in the image. In a pinhole camera without lens the picture all motif points are nearly sharp. As soon as a lens comes into play, the camera apparatus has to be focused, so that the motif points are on the targeted sharpness layer, i.e. the focal point lies on film or CCD layer. A bad focus results in a unsharp image, i.e. one motif point is assigned to multiple image points.

Resolution. The term resolution is used in two contexts: 1. For the size of digital raster images and 2. for the image proportion between an analogue original and the digital copy. These are somewhat correlated, if one assumes that a digital image is captured from an indefinitely exact analogue original, the size of the digital copy will rise with the resolution.

In the face recognition area the minimum resolution is commonly defined over the minimum pixel distance from eye to eye. Is the resolution to low, differences between persons will start to be averaged. High resolutions are expected not to cause problems other than a delay in processing the large amount of pixels.

Compression. For storing biometric images in small data storages (e.g. RFID, smart card), compression of digital face images will probably become a very common disturbance. As JPEG is a very common image compression method with a common compression artifact it was chosen.

Some image compression algorithms allow to control the rate of compression over a parameter. For JPEG this parameter controls the number of frequency coefficients that are used for each JPEG image block (8x8 pixels). For low to medium values the changes are nearly not noticeable, while with high compression rates blocky disturbances arise, which are caused by the effect that two adjacent JPEG image blocks differ in color noticeably because the information amount available for each block is very low.

Grey Scale. The color information of an image is an additional dimension that can be used by face recognition algorithms, e.g. examining only the red channel for locating faces.

The one of the main dangers of using color as information source is the its dependency on illumination and acquisition equipment. Different lighting conditions and/or cameras will most probably result in a difference in the color distribution. This is the reason why most recognition algorithms work with grey scale images internally.

The influence of color is of special interest as this is an excellent possibility of reducing the image size whilst maintaining the image size and/or constrict compression degradation.

Occlusion by Hats and Sunglasses. Wearing sunglasses can disturb the image localization process noticeable as one of the main invariant face features – the eyes – are masked. In standard illumination environments hats or caps can cast a shadow over the face and darkening the upper face area drastically.

3 Implementation and Methodology

3.1 Reusing vs. Generating

As the predecessor BioFace I/II has shown [1], image quality is one of the main parameters influencing the recognition performance. In fact this was the motivation to implement BioFace III. In BioFace I/II over 200,000 images have been categorized regarding image disturbances and could have been re-used for BioFace III, but two main reasons speak against this.

Firstly the images often not only contain one disturbance but a combination of these. This makes the task of measuring the influences of the single disturbances hard. If we define the result $s(c_{u_1 \to u_2})$ of a biometric face recognition comparison[2] $c(u_1 \to u_2)$ of two undisturbed images u_1 and u_2, we can measure the influence of a disturbance d_m on the similarity score s as Δs_{d_m}. If we now combine two disturbances d_i and d_j in one image, the resulting degradation $\Delta s_{d_i \cup d_j}$ does not necessarily be the same as $\Delta s_{d_i} + \Delta s_{d_j}$ although this can be. As the BioFace I/II images were mostly images with combined disturbances the influence of the synergy factors would possibly influence the results noticeably.

The second problem by reusing already disturbed material is the measurement of the strength of a given disturbance. While contrast or brightness disturbances can be easily detected and measured this is not the case e.g. for head rotations.

Connected to this a third fact against reusing the BioFace I/II images arises: The number of disturbances is not equally distributed, either over the disturbance types or over the disturbances strengths.

All this led to the conclusion that new images need to be taken of a given test group. Some of these had to be taken during the photographic session (see table 2a), while the rest have been simulated by computer graphics measures (see table 2b) by using the open source software ImageMagick™.

As Δs_{d_i} is calculated as a distance of two score results from comparisons, a reference of how good face recognition algorithms perform on undisturbed data had to be established first. Therefore two undisturbed images were taken of all test persons.

[2] The operation $c(u_1 \to u_2)$ denotes that u_1 was enroled in the gallery and u_2 being the probe image for the verification $c(u_1 \to u_2)$.

Table 1. Disturbances

Disturbance	Variation	Number of Images
Undisturbed	Frontal, neutral mimic, no headdress, spectacle wearers without glasses	2
Undisturbed alternative[a]	As above, with glasses	2
Rotation neck axis	±10°, ±15°, ±20°	6
Rotation ear axis	±20°, ±45°	4
Rotation nose axis	±15°, ±30°	4
Mimic	Smile, Grin, Closed Eyes	3
Other disturbances	Sun Glasses, Cap	2

[a] Spectacle wearers only

(a) Photographed Disturbances

Disturbance	Variation	Number of Images
Chrominance / Gray scale	None	1
Overexposure	+10%, +20%, +30%, +40%, +50%	5
Underexposure	-10%, -20%, -30%, -40%, -50%	5
Contrast	±1, ±2, ±3, ±4, ±5 steps	10
Sharpness	±1, ±2, ±3, ±4, ±5 steps	10
Resolution	10, 20, 50, 72, 96, 100, 150, 200, 300, 600, 1200 dpi	11
Compression	steps 10, 20, 40, 60, 80, 100	6
Digitizing Disturbances	Reflections, Scratches, Staining	4

(b) Generated Disturbances

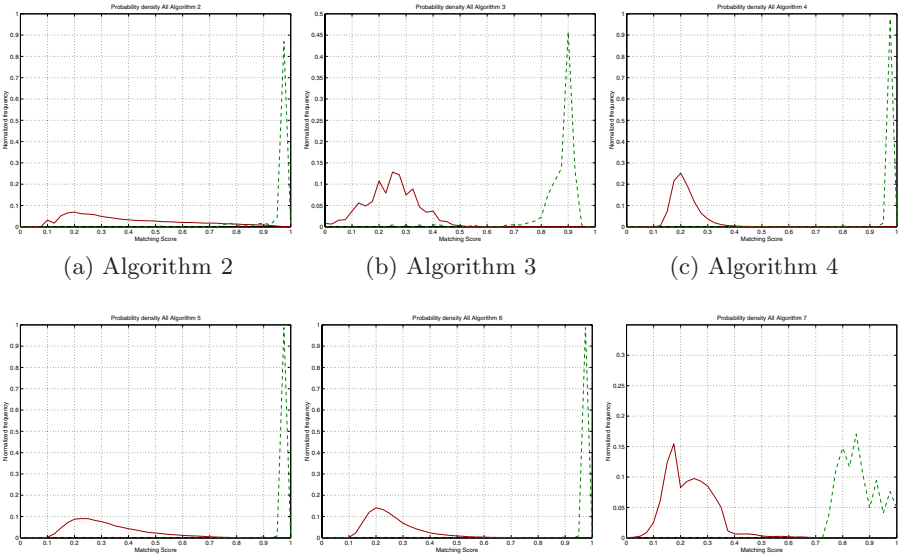
3.2 Algorithms

In the test seven algorithms were used that were developed by five different vendors. Most of the algorithms were the commercially available products, but some were lab prototypes with new developments or different parametrization. The test was done in a "black box" setup where the testers had no knowledge about the inner workings of the algorithms.

During the test analysis it became obvious that algorithm 1 was severely flawed, the error could be tracked to a bug in the software that was developed to comply to the test interface. Therefore algorithm 1 was excluded from further analysis to not influence biometric performance considerations by buggy results.

3.3 Test Procedure

The goal of BioFace III was to measure the influence of image disturbances on biometric (two-dimensional) face recognition. The test didn't have a competitive character unlike other comparable tests like the FRVT [2][3], the FRGC [4] or the FVC [5][6][7]. Therefore the vendor and algorithm names are only published as pseudonyms.



(a) Algorithm 2 (b) Algorithm 3 (c) Algorithm 4

(d) Algorithm 5 (e) Algorithm 6 (f) Algorithm 7

Fig. 1. Probability Densities of the Undisturbed Verification

The test type resembles roughly a technology evaluation [8] without targeting a certain application, which also reflects the rather wide selection of disturbances, even though in the back of the mind a border crossing scenario and the ICAO choice for face recognition as primary biometric [9] where inspirations.

The test itself was run as a crossover verification test, which means that for each algorithm a, firstly for the set of all images I, each image $i \in I$ was enroled, let E be the set of successfully enroled images. Then sequentially each image was used as probe $p \in I$ for and a verification $v(e_k, p_l)$ $k = 1 \ldots |E|, l = 1 \ldots |I|$.

Even though each image $e \in E$ was used as gallery image for the verification the analysis in BioFace III was limited to the verifications with the second of both undisturbed image being the gallery image.

4 Results

4.1 Undisturbed Verification

Firstly the verifications of the both undisturbed images were filtered from the verification results to build a reference for measuring the score difference introduced by the disturbances. Of course these results are interesting on their own, as they show the capabilities of the algorithms when working on "optimal" data. If we take a close look at the probability densities of the undisturbed verification in 1, all algorithms perform very well, with algorithms 4-7 even a zero FR/FA. This obviously motivates a strict quality control as defined in [9] or [10].

4.2 Selected Findings from the Disturbed Verification

While all results would go beyond the scope of this article, we will analyze some selected findings of special interest. These were chosen, because they show the different impacts of the influences on therecognition approaches of the algorithms.

Resolution. The resolution test group consisted of images ranging from 10 dpi to 1200 dpi (see table 2b on page 904). Nearly all algorithms didn't enrol any images below 72 dpi, one exception was algorithm 3. As figure 2 shows this wasn't successful either, as the low resolution images reach very low performance. This is no wonder, as the low resolutions also bring a small eye distance with them, which is usually a good measure for estimating the usability of an image for face recognition, as the image size itself does not tell anything about the size of the face pictured.

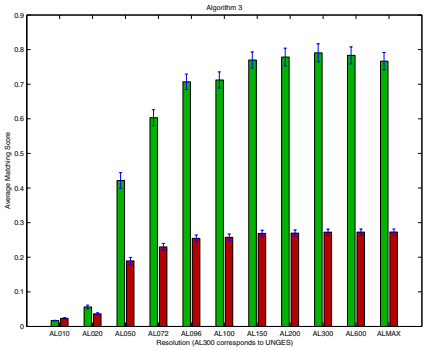
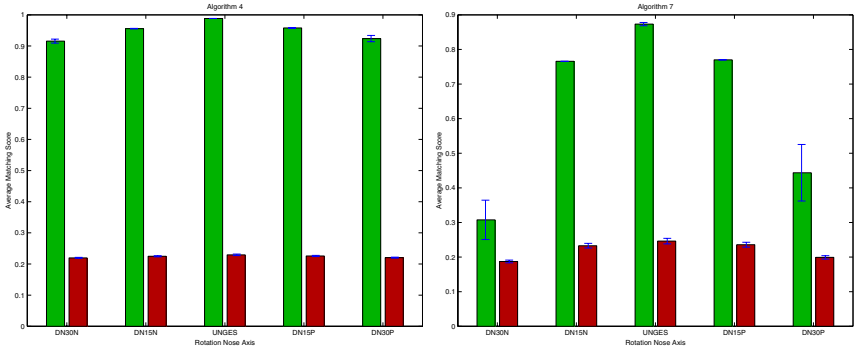


Fig. 2. Average Matching Scores (Genuine/Impostor) regarding Resolution (key AL) Algorithm 3

Head Rotation Around Nose Axis. Regarding the information loss for the face image the rotation around the neck axis can be seen as the rotation with the most impact followed by the rotation around the ear axis[3]. In contrast to this the impact of the rotation around the nose axis is negligible. Nevertheless with the rotated face image certain face localizations or approaches for recognition (e.g. template matching) might fail. This can be seen in figure 3, comparing the degradations of the matching scores of algorithm 4 and 7 regarding the rotation around the nose axis. Algorithm 4 seems to be able to compensate the rotation quite well, while algorithm 7 has to cope with the $\pm 15°$ rotations already and showing severe problems for the $\pm 30°$ rotations. Another indicator for the problems handling the $\pm 30°$ rotations is the significant rise of the standard deviation, depicted as "wick" of the bar plot.

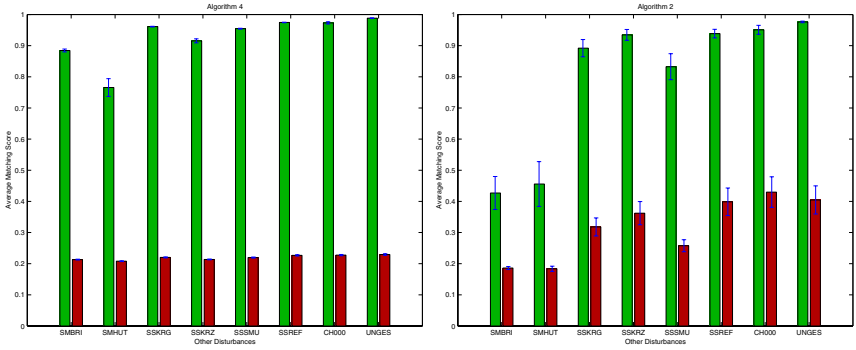
Sunglasses, Headdress. One disturbance aimed on hiding the eyes by sunglasses and another on dropping a shadow over the eyes by a baseball cap. As shown in figure 4 algorithm 7 shows a severe drop of the average matching scores

[3] See 2.1 on page 901



(a) Algorithm 4 (b) Algorithm 7

Fig. 3. Average Matching Scores (Genuine/Impostor) regarding Head Rotation around Nose Axis (key DN)



(a) Algorithm 4 (b) Algorithm 7

Fig. 4. Average Matching Scores (Genuine/Impostor) regarding Several Disturbances (SMBRI = glasses, SMHUT = hat/cap, SSKRG, SSKRZ = scratches, SSSMU = dirt, SSREF = reflection, CH000 = grey scale, UNGES = undisturbed)

regarding sunglasses (SMBRI) and the cap (SMHUT), while other disturbances do not show this behavior. This leads to the assumption that the eye detection is a key element of algorithm 7.

As in section 4.2 we take algorithm 4 as comparison, where the degradation is still noticeable, but much less severe.

References

1. C. Busch, H. Daum, F. Graf, M. Martin, A. Munde, A. Pretzel, T. Reinefeld, R. Salamon, C. Vogel, M. Wiegand. BioFace – Comparative Study of Face Recognition Systems, Public Final Report BioFace I and II, Version 2.1. Technical report, Bundesamt für Sicherheit in der Informationstechnik, June 2003. `http://www.bsi.de/english/fachthem/BioFace/BioFaceIIReport.pdf`.

2. Duane M. Blackburn, Mike Bone and P. Jonathon Phillips. Facial Recognition Vendor Test (FRVT) 2000 Evaluation Report, February 2001. http://www.frvt.org/DLs/FRVT_2000.pdf.
3. P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duane M. Blackburn, Elham Tabassi, Mike Bone. Face Recognition Vendor Test (FRVT) 2002 Evaluation Report, March 2003. http://www.frvt.org/DLs/FRVT_2002_Evaluation_Report.pdf.
4. P. Jonathon Phillips. Face Recognition Grand Challenge. Presentation at BC2004, September 2004. http://www.bee-biometrics.org/files/presentations/BC2004 FRGC Phillips.pdf.
5. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, A. K. Jain. Fingerprint verification competition 2000, September 2000. http://bias.csr.unibo.it/fvc2000/.
6. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, A. K. Jain. Fingerprint verification competition 2002, April 2002. http://bias.csr.unibo.it/fvc2002/.
7. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, A. K. Jain. Fingerprint verification competition 2004, March 2003. http://bias.csr.unibo.it/fvc2004/.
8. ISO/IEC TC JTC1 SC37 Biometrics. Text of CD 19795-1, Biometric Performance Testing and Reporting – Part 1: Principles and Framework, Jul 2004. http://www.jtc1.org/FTP/Public/SC37/DOCREG/37N0684.pdf.
9. ICAO TAG 14 MRTD/NTWG. Biometrics Deployment of Machine Readable Travel Documents, Version 2.0. Technical report, ICAO, May 2004. http://www.icao.int/mrtd/download/documents/Biometrics deployment of Machine Readable Travel Documents.pdf.
10. ISO/IEC TC JTC1 SC37 Biometrics. Text of FCD 19794-5, Biometric Data Interchange Formats – Part 5: Face Image Data, March 2004. http://www.jtc1.org/FTP/Public/SC37/DOCREG/37N0506.pdf.

Local Feature Based 3D Face Recognition

Yonguk Lee, Hwanjong Song, Ukil Yang, Hyungchul Shin, and Kwanghoon Sohn*

Biometrics Engineering Research Center
Department of Electrical & Electronics Engineering,
Yonsei University, Seoul, 120-749, Korea
{quience,ultrarex,starb612,k9doli}@diml.yonsei.ac.kr, khsohn@yonsei.ac.kr

Abstract. This paper presents a 3D face recognition system based on geometrically localized facial features. We propose the feature extraction procedure using the geometrical characteristics of a face. We extract three curvatures, eight invariant facial feature points and their relative features. These features are directly applied to face recognition algorithms which are a depth-based DP (Dynamic Programming) and a feature-based SVM (Support Vector Machine). Experimental results show that face recognition rates based on the depth-based DP and the feature-based SVM are 95% for 20 people and 96% for 100 people, respectively.

1 Introduction

Face recognition technologies have made great progress using 2D data for the past few decades. Although they played an important role in many applications such as identification, crowd surveillance and access control under the controlled inner and outer environments [1], there are still many unsolved problems in varying environments such as pose, illumination and expression. With the development of a 3D acquisition system, face recognition based on 3D information is attracting greatly in order to solve problems of using 2D data. Early work on 3D face recognition was launched decades ago, and a few approaches have been reported about face recognition using 3D data which were acquired by a 3D sensor [2] and a stereo-based system [3]. Having a 3D system removes many problems associated with lighting and pose that can affect 2D systems. The solution of 3D face recognition should be more accurate as we can concentrate on invariant features of the face.

In this paper, we propose a 3D face recognition system based on geometrically localized facial features with two different 3D sensors. We use Genex 3D FaceCam system for probe data and a 3D full laser scanner of Cyberware which provides good quality of face image for gallery data in our system. Fig. 1 shows the block diagram of the proposed system. As shown in Fig. 1, the system consists of three stages, which are data acquisition stage, feature extraction stage and recognition stage. Firstly, in the data acquisition stage, we explore the data acquisition

* Corresponding author

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 909–918, 2005.
© Springer-Verlag Berlin Heidelberg 2005

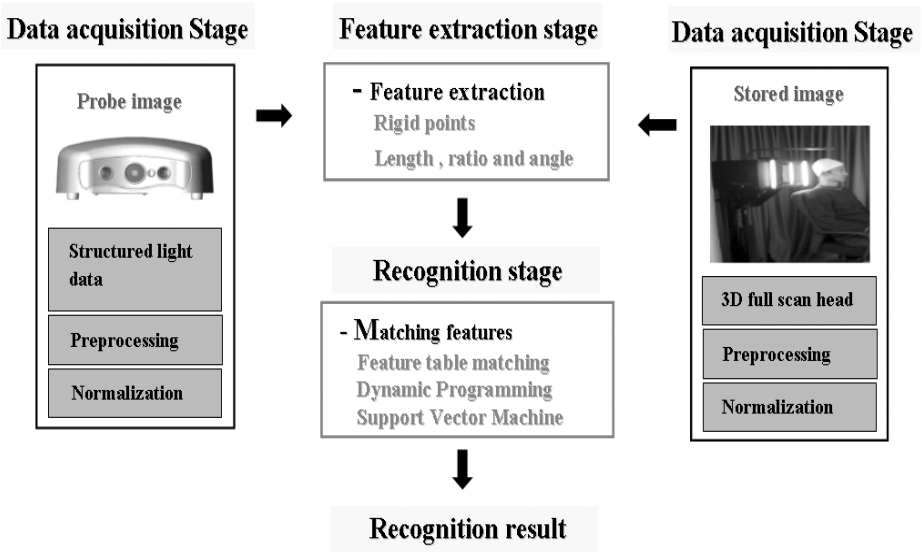


Fig. 1. The block diagram of the proposed system

process of probe and gallery data. The obtained probe and gallery data are in the established 3D normalized space through the preprocessing step. Secondly, in the feature extraction stage, the proposed 3D facial feature extraction method using the geometrical characteristics is described. We also propose the relative features that are obtained by using the relations among the previously extracted feature points. Finally, in the recognition stage for the matching a probe data with the gallery data, we induce the recognition method using a depth-based DP (Dynamic Programming) and a feature-based SVM (Support Vector Machine).

The remainder of this paper is organized as follows. Chapter 2 explains the extraction of 3D facial feature points and relative features using geometrical characteristics of 3D face. Chapter 3 describes the proposed 3D face recognition methods using the depth-based DP and the feature-based SVM in detail. Test performance is analyzed to explain the efficiency of the proposed algorithm and discussion is presented in Chapter 4. Finally, Chapter 5 concludes a summary of the contributions of this paper and the future works.

2 Feature Extraction Algorithms

2.1 Facial Feature Extraction

As for acquiring 3D probe data, we utilize a 3D device based on the structured light method, and for 3D gallery data, we utilize 3D Laser Scanner. Then, normalization of probe and gallery data is required to perform a face recognition algorithm. Therefore, we normalize the face data to make all the equivalent face space [4].

In this section, we propose a facial feature extraction method for discriminating individual faces. We define generic facial feature points related to nose and eyes. Typical locations of some of these features are indicated for reference in Fig. 2. Each of these face descriptors is defined in terms of a high level set of relationships among depth and curvature features [5]. The procedure of the facial feature extraction is as follows:

1. We can vertically and almost symmetrically divide the face using the Y-Z plane which includes the NPP (Nose Peak Point) and the Y axis, and obtains the face dividing curvature. Therefore, we can extract the face center curve.

2. On the face center curve, we use curvature characteristics to extract facial feature points, which are convex and concave points by differentiating the depth variable z with respect to the length variable y. These points are named as the NBP (Nose Base Point), the NBRP (Nose BRidge Point)and the CPE (Center Point between Eyebrows).

3. We can extract a horizontal curvature which passes the NPP. We can obtain two feature points named as the NEP (Nose End Points) by using the derivative of the depth variable z with respect to the width variable x.

4. As the procedure as mentioned above, we can extract a horizontal curvature through NBRP. We also obtain two points on this curvature named as the EIP (Eye Inner corner Points) on this curvature.

Finally, we extract three distinctive curvatures of a face using geometrical facial characteristics. These curvatures consist of geometrical concavities of a frontal face; therefore, we obtain eight important feature points on these curvatures by using the derivative of the depth value z with respect to the width variable x and the length variable y.

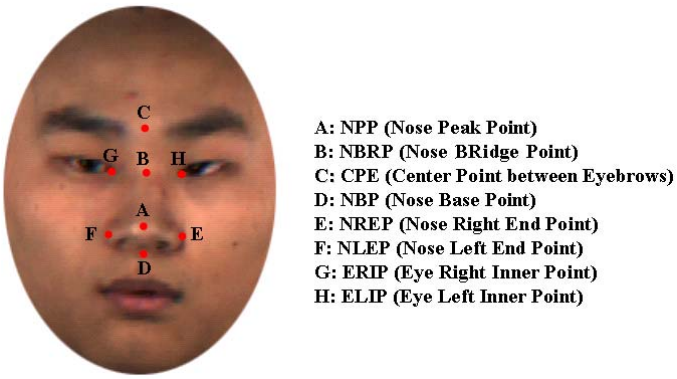


Fig. 2. Generic-facial feature points

2.2 Relative Features

In this section, we propose the relative features which are composed of previously extracted feature points. They are the distances and the ratios between feature points and the angles among feature points. These relative features can

be utilized for distinguishing the individuals explicitly, because relations of 3D data are invariant. Each alphabet from A to H referred to Fig. 2 stands for eight extracted feature points. Each feature point has x, y and z coordinate values. For an example, we denote A feature point as $A(a_x, a_y, a_z)$. Equations of the relative features are as follows:

$$L_1 = |c_y - b_y|,\ L_2 = |f_x - e_x|,\ and\ L_3 = |g_x - h_x|\ , \tag{1}$$

where L_1 ,L_2 and L_3 are the relative lengths of facial feature points.

$$V_1 = |A(a_x, a_y, a_z) - B(b_x, b_y, b_z)|\ and\ V_2 = |B(b_x, b_y, b_z) - C(c_x, c_y, c_z)|\ , \tag{2}$$

where V_1 and V_2 are the relative distances of facial feature points.

$$R_1 = \frac{|c_y - b_y|}{|g_x - h_x|},\ R_2 = \frac{|a_y - b_y|}{|f_x - e_x|}\ and\ R_3 = \frac{|f_x - e_x|}{|a_z - d_z|}\ , \tag{3}$$

where R_1 , R_2 and R_3 are the relative ratios of distances.

$$\Theta_1 = \angle EBF,\ \Theta_2 = \angle GAH\ \ \Theta_3 = \angle EAF\ and\ \Theta_4 = \angle GCH\ , \tag{4}$$

where Θ_1,Θ_2,Θ_3 and Θ_4 are the relative angles among facial feature points.

3 Face Recognition Based on Facial Features

3.1 Face Recognition by a Depth-Based DP

DP was introduced to pattern recognition field related to a time-dependent process. The objective of pattern recognition using DP is to find an optimal time-alignment between two sequential patterns [6]. DP is usually used in optimization problems in which a set of decisions must be made to arrive at an optimal solution [7].

The main advantage of using a depth-based DP for our face recognition system is that we do not care about the different number of data points and correspondence problems between face data having two different 3D sensors, and we would also expect good recognition results more than simple correlation algorithm. However, one of disadvantages of the depth-based DP is to take a long time to find an optimal path, if there are many lots of sequential vectors which consist of facial curvatures for matching. In order to solve this problem, we select the several candidate faces of gallery data by using the Euclidean distance.

We now extract facial curvatures of the face candidates in order to be fed into the the depth-based DP system for graph matching. There are four vertical curvatures which are parallel to the center curvature through the NPP, and two horizontal curvatures related to eyes and nose. These facial curvatures are distinguishable features among the individuals; because, everyone has different depth value of the facial concave and convex points. As the same as sequential patterns in signature or speech recognition, extracted facial curvatures of a probe data are very similar to those of the gallery data of the same face. On the other hand, the curvature patterns of the other gallery data are more different from the probe data. The extracted seven facial curvatures, as described in Fig. 3, are directly used as sequential time-aligned vectors having only depth values.

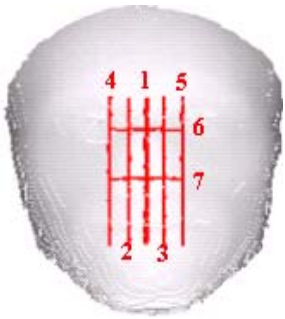


Fig. 3. Extracting seven curvatures of a face

3.2 Face Recognition by a Feature-Based SVM

In this section, we propose a face recognition algorithm by using a feature-based SVM. A SVM has been recently proposed as a new technique for pattern recognition [8]. The main objective of pattern recognition using a SVM is to solve two-class classification problem. Given a set of points which belong to either of two classes, a SVM finds the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyperplane. According to [9], given fixed but unknown probability distributions, this hyperplane called OSH (Optimal Separating Hyperplane) which minimizes the risk of misclassifying not only the examples in the training set but also the yet-to-be-seen examples of the test set.

A SVM used in 2D face recognition requires more feature vectors or intensity values of the gallery data per a class for training support vectors in order to find the optimal hyperplane, because the more feature vectors per a class are used, the better recognition rate is performed. Therefore, elapsed time of classification is not in real time. However, in our system, advantages of using a feature-based SVM are that training feature vectors which have a gallery data per a class is accomplished in real time, and SVMs can find the optimal hyperplanes among one and the others by using well defined facial features of 3D face data as feed of the SVM. We use the extracted eight facial feature points and their relative features such as lengths, ratios and angles as mentioned in Chap 2. We apply 100 classes out of BERC (Biometrics Engineering Research Center) face database for the experiments [10]. In order to solve multi-class problem using the SVM, we induce the one-vs-all strategies. It separates one class from the other classes by positive value +1 and negative value -1. When a probe data is used for the test, facial feature vectors of a probe data are directly fed into 100 SVMs during the learning, and a SVM having the highest value of output values from each SVM is ether matched or not matched with target class.

4 Simulation Results and Analysis

For experimental results, we utilize Visual studio 6.0 C++ programming tool to implement our face recognition system and OpenGL programming for rendering

Table 1. Facial feature points for probe and gallery data

Features	Gallery data(13)			Probe data(13)			Probe data(12)		
	x	y	z	x	y	z	x	y	z
CPE	0.738	60.11	86.41	0.490	60.34	86.21	0.6009	55.7911	78.6374
NBRP	0.390	39.14	81.64	0.434	39.264	79.93	-0.6437	37.23331	76.4823
NPP	0	0	100	0	0	100	0	0	100
NBP	0.734	-10.48	86.35	0.7052	-10.25	87.77	0.83728	-13.681	80.1466
NRIP	21.21	0	78.29	21.222	-0.028	78.09	17.4555	0.19370	74.8431
NLIP	-20.24	0	76.81	-20.63	-0.497	75.72	-17.853	-0.4307	73.6493
ERIP	14.34	39.14	74.68	14.096	39.116	74.91	11.4168	37.6236	70.2066
ELIP	-14.29	39.14	73.94	-14.57	38.970	72.12	-11.648	37.046	70.0277

3D face data. We acquire 3D face database by using the laser scanner, Cyberware 3030/ RGB. We also obtain 3D probe data by the structured light device,3D FaceCam. We adopt the frontal faces of 100 people of BERC face database.

4.1 Simulation Results of Extracted Facial Features

The simulation results of the proposed feature extraction algorithm are illustrated in this section. Table 1 shows the comparisons of x, y and z values of eight features about gallery data and probe data. x, y and z values of gallery data (13) as shown in Table 1, are very similar to probe data (13), but x, y and z values of probe data (12) are different and distinguishable from gallery data (13). Table 2 also shows the similarity of eight feature points and their relative features about probe and gallery data respectively. We normalize all face data for NPP to be located on 0, 0 and 100 for x, y and z coordinates respectively. This means that all the face data are in the same space of the view port.

4.2 Simulation Results of Face Recognition

Depth-Based DP: In this section we show the simulation results related to a face recognition algorithm by a depth-based DP. We use 20 people of BERC face database as gallery and probe data respectively. Before using the depth-based DP for 3D face recognition, we use the Euclidean distance between extracted feature values of gallery data and probe data in order to select the similar face candidates among gallery data. Fig. 4 shows an example of selected candidate faces, when face number 19 is used as probe data. There are three face candidates which are face number 11, 17 and 19 selected under the acceptable error score 30. The curvatures of these face candidates are directly connected to sequential patterns of DP. Fig. 5 shows the simulation results by the depth-based DP when gallery data 17, 19 and 20 are chosen. The matching errors of seven curvatures per each probe data are computed as a total error. When we compare probe data number 19 with gallery data number 19, the total matching error is 98.82. On the other hand, in case of probe data number 17 ,the total matching error are 121.34. However, when we also compare the gallery data 19 with a probe data

Table 2. Facial feature points and their relative features for probe and gallery data

Feature types	Gallery data values(001)	Probe data values(001)	Probe data values(002)
CPE(z value)	83.639220	84.263680	87.731530
NBRP(z value)	77.616940	77.541620	81.993890
NBP(z value)	86.118020	86.911620	82.352200
NRIP(z value)	76.999370	76.224990	71.538910
NLIP(z value)	77.217860	79.061580	73.571020
ERIP(z value)	73.406790	72.759880	76.142100
ELIP(z value)	73.643800	75.527330	76.527230
L_1	19.570000	19.621490	20.846400
L_2	42.786420	42.303920	39.114720
L_3	31.678770	31.573760	35.448090
Θ_1	55.620597	54.895556	51.462474
Θ_2	36.450379	36.274882	39.249164
Θ_3	86.111687	86.937483	71.002492
Θ_4	68.435927	70.513071	87.305494
R_1	0.617764	0.621449	0.58808
R_2	0.947590	0.958799	0.964348

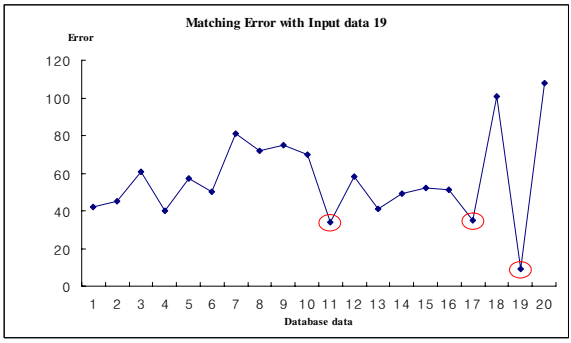


Fig. 4. Selecting face candidate based on Euclidean distance

number 20 which is not the face candidates, the total error score is very higher than the face candidates.

As we match the distance between feature points among face candidates, we obtain similar score values, but the extracted seven curvatures include all the feature points as mentioned in Chap. 3. In other words, curvature matching is much robust than matching based on MSE, because the curvatures have much higher dimensions than facial feature points. Therefore, we achieved 95% face recognition rate for 20 people according to our proposed method.

Feature-Based SVM: In this section we describe the simulation results related to face recognition algorithm by a feature-based SVM. We use 100 people of BERC face database as gallery and probe data respectively. Eight feature points and their relative features illustrated in Chap. 2 are fed into directly SVM. For

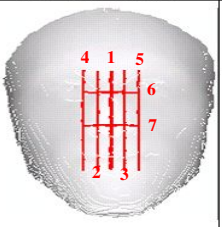
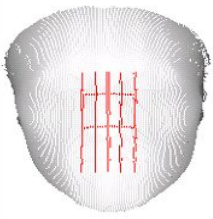
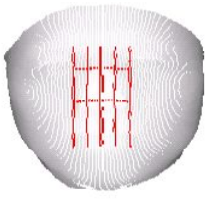
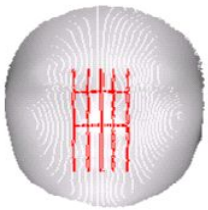
DB_019_Normal	Input_019_Normal	Input_017_Normal	Input_020_Normal
			
Curvature_1	9.337120	15.01710	37.76449
Curvature_2	12.57815	16.34530	35.67503
Curvature_3	11.19463	14.38401	34.77351
Curvature_4	14.11903	18.72886	51.15996
Curvature_5	13.18248	16.15266	43.75500
Curvature_6	19.86787	17.59605	40.56902
Curvature_7	18.54281	23.12011	31.56345
Total	**98.82209**	**121.34409**	**275.26046**

Fig. 5. Simulation result based on Dynamic Programming

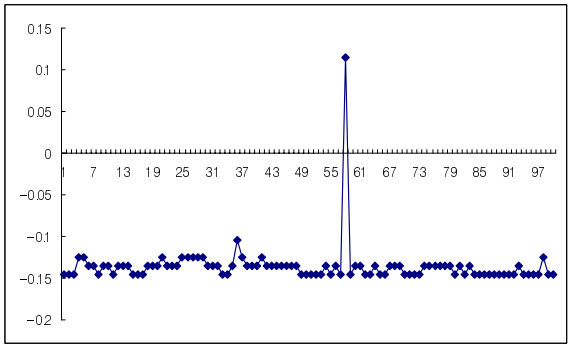


Fig. 6. Face recognition result for the probe data number 58

testing our recognition system, when we put the probe data number 58 into the system, the same data among gallery data has only a positive value and the other classes have negatives value as shown in Fig. 6. We compare the simulation results of the feature-based SVM with those of a simple correlation method as described in Fig. 7. The cumulative matching score of the feature-based SVM shows 96% at the first rank and 97% at the third rank.

It means that face recognition rates are 96% at the first rank and 97% at the third rank respectively. However, the cumulative matching score of MSE shows 89% at the first rank and 90% at the third rank. Therefore, we achieved about 7% higher than the simple correlation method.

5 Conclusion and Future Work

In this paper, we proposed the 3D face recognition system based on two different devices for probe and gallery data respectively. We utilize the 3D laser scanner

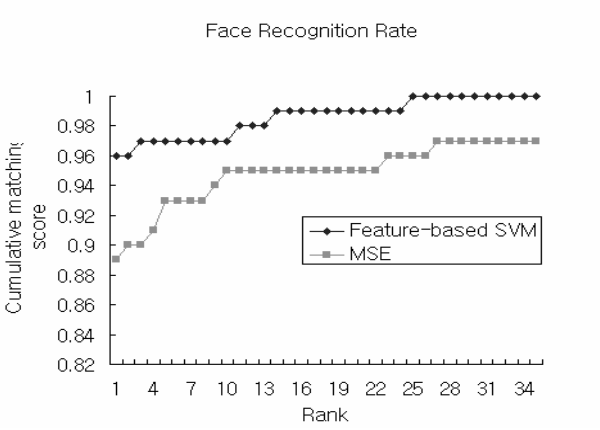


Fig. 7. Face recognition rate based on SVM and MSE

for obtaining gallery data and the structured-light based 3D device for acquiring probe data. According to the proposed feature extraction method, we extract eight feature points that are geometrically invariant, and we obtain relative features such as the distance and the ratio between points and the angle among feature points. These relative features can distinguish the individuals better than only using the facial feature points. In the recognition stage, we proposed two different recognition algorithms, a depth-based DP and a feature-based SVM. In the experimental results, we show that the resulting recognition rate is 95% for 20 people by DP and 96% for 100 people respectively. When we compare the results of the feature-based SVM with these of MSE, face recognition rate by the feature-based SVM is 7% higher than that of the simple correlation method. For further works, we are researching for the pose invariant face recognition system and more robust 3D face recognition algorithms.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center at Yonsei University.

References

1. R. Chellappa, C. L Wilson, and S. Sirohey: Human and machine recognition of faces: A survey. Proc. the IEEE, Vol. 83, pp.705-740, 1995.
2. W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phllips: Face recognition: A survey. CVL Tech. Report, Center for Automation Research, University of Maryland at College Park, 2000.
3. G. Medioni and R. Waupotitsch: Face recognition and modeling in 3D. Proc. the IEEE Int'l Workshop on Analysis and Modeling of Faces and Gestures, pp.232-233, 2003.

4. H. Song, U. Yang, K. Sohn: 3D face recognition under pose varying environments. Lecture Notes in Computer Science, Vol. 2908, pp. 333-347, 2004.
5. Peter L. Hallian: Two-and Three Dimensional patterns of the Face. A K Peters LTD, pp.202-203.
6. H.Sakoe and S.Chiba: A dynamic programming approach to continuous speech recognition. Proc. the 7th ICA, pp.20, Aug. 1971.
7. Hichem sahi and Nozha Boujemaa: Robust Face Recognition Using Dynamic Space Warping. Biometric Authenticatgion, LNCS 2359, pp.121-132, 2002.
8. Guodong Guo, Stan Z.Li, and Kapuk Chan: Face recognition by Support Vector Machines. Proceedings. Proc. the Fourth IEEE International Conference, pp.196 - 201, Mar. 2000.
9. M.Pontil and A.Verri: Support Vector Machines for 3D object recognition. IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 20, pp.637-646, 1998.
10. "Biometrics Engineering Research Center at Yonsei University": http://berc.yonsei.ac.kr.

Fusion of Appearance and Depth Information for Face Recognition

Jian-Gang Wang, Kar-Ann Toh, and Ronda Venkateswarlu

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
{jgwang,katoh,vronda}@i2r.a-star.edu.sg

Abstract. In this paper, an investigation is carried out regarding combination of 3D and 2D information for face recognition. A two-stage method, PCA and a Reduced Multivariate Polynomial Model (RMPM), is developed to fuse the appearance and depth information of face images in feature level where simplicity (number of polynomial coefficients increases linearly with model-order and input-dimension, i.e. no dimension explosion as in the case of full multivariate polynomials) and ease of use (can be easily formulated into recursive learning fashion) are major concerns. To cater for fast on-line registration capability when a new user arrives, the learning is formulated into recursive form. The improvement of the face recognition rate using this combination is quantified. The recognition rate by the combination is better than either appearance alone or depth alone. The performance of the algorithm is verified on both XM2VTS database and a real-time stereo vision system, showing that it is able to detect, track and recognize a person walking towards a stereo camera within reasonable time.

1 Introduction

A good account of research efforts has been devoted to grey-level analysis of frontal face recognition [4]. However, most of the appearance-based methods are still sensitive to the pose and illumination conditions. A robust identification system may require fusion of several modalities. Ambiguities in one modality like lighting problem may be compensated by another modality like depth features. Multi-modal identification system hence usually performs better than any one of its individual components [5].

While a lot of work has been carried out in face modeling and recognition, 3D information is still not widely used for recognition. Main reason could be due to the low resolution offered by those affordable equipments. Intuitively, a 3-D representation provides an added dimension to the useful information for the description of the face. This is because 3D information is relatively insensitive to illumination, skin-color, pose and makeup, and this can be used to compensate the intrinsic weakness of 2D information. Studies have demonstrated the benefits of having this additional information using reasonably high-resolution 3D scanners [6]. On the other hand, 2D image complements well 3D information. They are localized in hair, eyebrows, eyes, nose, mouth and facial hairs, skin color precisely where 3D capture is difficult and not accurate.

Beumier et al [1] investigated the improvement of the recognition rate by fusing 3D and 2D information. Error rate was reported to be 2.5% by fusing 3D and grey

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 919–928, 2005.
© Springer-Verlag Berlin Heidelberg 2005

level on a 26 subjects database. Both Gordon [6] and Beumier [1, 2] realized that the performance of 3D facial features based face recognition depends on the 3D resolution. Thanks to the technical progress in 3D capture/computing, an affordable real-time stereo system soon becomes available by which one can get a reasonably good resolution of 3D data in real-time.

There is a rich literature on fusing multiple modalities for identity verification, e.g. combining voice and fingerprint, voice and face biometrics [5], visible and thermal imagery [9]. Evaluation on fusion face recognition algorithms on intensity image can be found in [3]. The fusion can be done in feature level or decision level with different fusion models, e.g. appearance and depth are fused for face recognition by min, sum, product in [13, 14], by weighted sum in [1, 15]. The fusion algorithm is critical part to obtain a high recognition rate [7]. There are some limitations in existing decision fusion models. Statistical models (e.g. kNN, Bayesian) rely heavily on prior statistical assumptions which can depart from reality; Linear models (e.g. weighted sum, LDA) are limited to linear decision hyper-surfaces; Nonlinear models (e.g. Neural Networks, RBF, SVM) involves nonlinear optimization where only local solutions are obtained. Moreover, the learning process could be very tedious and time consuming.

Multivariate Polynomial (MP) provides an effective way to describe complex nonlinear input-output relationship since it is tractable for optimization, sensitivity analysis, and predication of confidence intervals. With appropriate incorporation of certain decision criteria into the model output, MP can be used for pattern analysis and could be a fusion model to overcome the limitations of the existing decision fusion models. However, the full MP has dimension explosion problem for large dimension and high order system. The MP model is a special example of kernel ridge regression (KRR) [11]. Also, KRR overcomes the computational difficulty by using the kernel trick. This computational difficulty drives us to use reduced multivariate polynomial model.

In this paper, we proposed to use a Reduced Multivariate Polynomial Model (RMPM) [10] to fuse appearance and depth information for face recognition where simplicity and ease of use is our concern. We also report a stage of development on fusing the 2D and 3D information, catering for on-line new user registration. This issue of new user registration is non-trivial since current available techniques require large computing effort on static database. Based on a recent work by [10], a recursive formulation for on-line learning of new user parameters is proposed in this paper. The verification performance of the face recognition system where color and depth images are fused will be reported.

The rest of the paper is organized as follows. The Fusion of depth and appearance information is discussed in Section 2. Section 3 presents some issues related to normalization of images. Section 4 discusses the experiment on XM2VTs and the implementation of the algorithm on a stereo vision system. Section 5 concludes the work with future enhancements to the system.

2 Fusing Appearance and Depth Information

As discussed in section 1, we aim to improve the recognition rate by combining appearance and depth information. The matter of the combination is crucial to the per-

formance of the system. The criteria for this kind of combination is to fully make use of the advantages of the two sources of information to optimize the discriminant power of the whole system. The degree to which the results improve performance is dependent on the degree of correlation among individual decisions. Fusion of decisions with low correlation can dramatically improve the performance.

In the following, we briefly discuss the RMPM and the auto-update algorithm of RMPM for new user registration in face recognition.

2.1 Multivariate Polynomial Regression

The general multivariate polynomial model can be expressed as

$$g(\alpha, \mathbf{x}) = \sum_{i=1}^{K} \alpha_i x_1^{n_1} x_2^{n_2} \quad x_l^{n_l} \tag{1}$$

where the summation is taken over all nonnegative integers $n_1, n_2, \quad, n_l$, for which $n_1 + n_2 + \quad + n_l \le r$ with r being the order of approximation. $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \quad, \alpha_K]$ is the parameter vector to be estimated and $\mathbf{x}$ denotes the regressor vector $[x_1, x_2, \quad, x_l]$ containing l inputs. K is the total number of terms in g ($\boldsymbol{\alpha}$, $\mathbf{x}$).

For example, a second-order bivariate polynomial model ($r = 2$ and $l = 2$) given by

$$g(\alpha,x)=\alpha^{T}p(x) \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6]$ and $\mathbf{p}(\mathbf{x}) = [1 \; x_1 \; x_2 \; x_1^2 \; x_1x_2 \; x_2^2]^T$.

Given m data points with $m>K$ (K=6) and using the least-squares error minimization objective given by

$$s(\boldsymbol{\alpha}, \mathbf{x}) = \sum_{i=1}^{m} [y_i - g(\boldsymbol{\alpha}, x_i)]^2 = [\mathbf{y} - \mathbf{P}\boldsymbol{\alpha}]^T [\mathbf{y} - \mathbf{P}\boldsymbol{\alpha}] \tag{3}$$

the parameter vector $\boldsymbol{\alpha}$ can be estimated using

$$\boldsymbol{\alpha}=(\mathbf{P}^{\mathbf{T}}\mathbf{P})^{-1}\mathbf{P}^{\mathbf{T}}\mathbf{y} \tag{4}$$

where $\mathbf{P} \in R^{m \times K}$ denotes the Jacobian matrix of p(x):

$$\mathbf{P} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{11}^2 & x_{11}x_{21} & x_{21}^2 \\ & & & & & \\ 1 & x_{1m} & x_{2m} & x_{1m}^2 & x_{1m}x_{2m} & x_{2m}^2 \end{bmatrix} \tag{5}$$

and $\mathbf{y} = [y_1, y_2, \quad, y_m]^T$ is the known interface vector from training data. The subscripts of the matrix element, x_{jk} (j = 1,2, k = 1,2,…m), indicates the number of the inputs and the number of instances respectively.

A simple way to improve numerical stability is to perform a weighted decay regularization by

$$\boldsymbol{\alpha} = (\mathbf{P}^T\mathbf{P} + b\mathbf{I})^{-1}\mathbf{P}^T\mathbf{y} \tag{6}$$

where $\mathbf{P} \in R^{m \times K}$, $\mathbf{y} \in R^{m+1}$ and $\mathbf{I}$ is a ($K \times K$) identity matrix. b is a regularization parameter (b is chosen to be a small value for stability).

2.2 Reduced Multivariate Polynomial Model

To significantly reduce the huge number of terms in above multivariate polynomials, a reduced model was proposed [10] as:

$$g_{RM}(\boldsymbol{\alpha}, \mathbf{x}) = \alpha^T P_{RM}(x) = \alpha_0 + \sum_{k=1}^{r}\sum_{j=1}^{l}\alpha_{k,j}x_j^k + \sum_{k=1}^{r}\alpha_k(\sum_{j=1}^{l}x_j)^k + \sum_{k=1}^{r}(\sum_{i=1}^{l}\alpha_{k,i}x_i)(\sum_{j=1}^{l}x_j)^{k-1} \qquad l, r \geq 2. \tag{7}$$

where x_j $j = 1, 2, \quad l$ are the polynomial inputs, α_0, $\alpha_{k,j}$ $\alpha_{k,i}$, are the weighting coefficients to be estimated, and l, r correspond to input-dimension, order of system respectively. The number of terms in this model can be expressed in a linear relationship: $K=1+r+l(2r-1)$.

Comparing with the existing fusion approaches, RMPM has some advantages: Number of parameters (polynomial coefficients) increases linearly with model-order and input-dimension, i.e. no dimension explosion as in the case of full multivariate polynomials; Nonlinear decision hyper-surface mapping; Fast single-step least-squares optimal computation: linear in parameter space, tractable for optimization, sensitivity analysis, and prediction of confidence intervals; Good classification_accuracy: better or comparable to SVM, Neural Networks, RBF, Nearest-Neighbor, Decision Trees [10]. Furthermore, RMPM can be easily formulated into recursive learning fashion for online applications. We will discuss this in section 2.4.

2.3 Face Recognition

In this paper, a two-stage PCA+RMPM is proposed for face recognition. The reduced model PCA is used for dimension reduction and feature extraction.

Learning. PCA is applied to color and depth images, respectively. The inputs of the learning algorithm of the reduced model are

$$\mathbf{W}_{pca} = \arg\max_{\mathbf{W}} \left| \mathbf{W}^T \mathbf{S}_T \mathbf{W} \right| \tag{8}$$

where $\mathbf{S}_T$ is the total scatter matrix of the training samples. $\mathbf{W}_{pca}$ can be obtained by solving an eigen problem. In this paper, the fusion of the color and depth information is completed in feature level, it is

$$P = RM(r, \begin{bmatrix} W_{pca_color} \\ W_{pca_depth} \end{bmatrix}) \tag{9}$$

where r is the order of RMPM. Then the parameters can be learned from the training sample by (6).

Testing. A probe face, F_T, is identified as a face, F_L, of the gallery if the output of the reduced model classifier (color and depth), $\mathbf{P}^T\boldsymbol{\alpha}$, is the maximum (and ≥0.5) among the all faces in the gallery. This can be expressed as:

$$\text{Maximum}(P^T(F_T)\alpha) \tag{10}$$

Where P is the obtained from the testing sample,

$$P(F_T) = RM(r, \begin{bmatrix} W_{pca_color}(F_T) \\ W_{pca_depth}(F_T) \end{bmatrix})$$

2.4 New User Registration

A face is determined to be a new user when the output of (10) is less than 0.5. (the initial value of an element in learning matrix is set to be 1 if the sample corresponds to the indicated subject, else it is set to be 0).

Assuming we have *n* face images in the original training set. The mean of the new training set is re-computed when a new user is found. Then the eigenfaces of the new set can be computed and the RM updating algorithm can be applied to the new eigfaces. The mean of all classes including the new user will be

$$m_{\text{new}} = (S_{t-1}+f_{\text{new}})/(n+1) \tag{11}$$

where S_{t-1} is the sum of the faces of the original training samples in time t-1, we have

$$S_t = S_{t-1}+f_{\text{new}}$$

The new eigenfaces can be computed using m_{new}.

Let f_i be the vector of all polynomial terms in (7) which is applied to the *i*-th samples. $F_T = [f_1, f_2 \quad , f_t]^T$. Let $\mathbf{M}_t = (\mathbf{P}^T\mathbf{P} + b\mathbf{I})$, then (6) becomes

$$\boldsymbol{\alpha} = \mathbf{M}_t^{-1}\mathbf{P}_t^T\mathbf{y}_t \tag{12}$$

we have

$$\mathbf{M}_t = \mathbf{M}_{t-1} + \mathbf{f}_t\mathbf{f}_t^T \tag{13}$$

$$\mathbf{P}_t^T\mathbf{y}_t = \mathbf{P}_{t-1}^T\mathbf{y}_{t-1} + \mathbf{f}_t\mathbf{y}_t \tag{14}$$

Finally, the new estimate α_t can be calculated using the previous estimate $\boldsymbol{\alpha}_{t-1}$ the inversion of $\mathbf{M}_{t-1}$ and the new training data $\{f_t, y_t\}$, we have

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \lambda\mathbf{M}_t^{-1}\mathbf{f}_t(\mathbf{y}_t - \mathbf{f}_t^T\boldsymbol{\alpha}_{t-1}) \tag{15}$$

The parameters can be computed automatically. The combination of the color (circle) and depth (square) in our approach can be explained using Fig. 1.

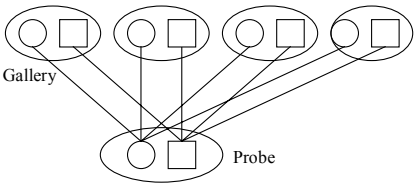


Fig. 1. Combination of color (circle) and depth (square)

3 Normalizations of Color and Disparity Images

Most researchers believe that frontal views are the best pose for recognition. In our system, a face is detected and tracked using stereo vision as the person moves in front

of the stereo camera [16]. The frontal face pose is automatically searched and detected from the captured color and depth images. These information are subsequently used for face recognition. Meanwhile, we evaluate our algorithm on XM2VTS database [12] to show the gain of recognition rate obtained with help of depth.

Face recognition is based on depth and intensity information from a binocular stereo vision system [8], which outputs the disparity/range information automatically (with ≥16 mm lens, range resolution is ≤2mm when the distance from the object to the stereo head is 1m; This assumes a baseline of 90mm, and a pixel size of 7.5 μm, with subpixel resolution of 1/16 pixel).

Using the image coordinates of the two eye centers the image is rotated and scaled to occupy a fixed size array of pixels (88 × 64). In the stereo vision system, the coordinates of pixel are consistent with the coordinates in the left image. The feature points, two eye centers can hence be located in the disparity image. The tip of the nose can be detected in the disparity image using template matching of Gordon [6]. From stereo vision, we have,

$$d' = bf/d \tag{16}$$

where d',represents the depth, d is the disparity, b is the baseline and f is the focal length of the calibrated stereo camera. Hence we can get depth image from disparity image with (16). The depth image is normalized using the depth of the nose tip, i.e. the nose tip of every subject is translated to the same point in 3D relative to the sensor. After that, the depth image is further normalized by the two eye centers.

Problems with the 3D data are alleviated to some degree by preprocessing to fill in holes (a region where there is missing 3D data during sensing) and spikes. We adopt the method in [13] to detect the spike, and then remove the holes by linear interpolation of missing values from good values around the edges of the hole.

In our approach, the color images are changed to the grey-level image by averaging three channels:

$$I = (R+G+B)/3 \tag{17}$$

4 Experiments

4.1 Experiment on Face Database

The proposed work uses the XM2VTS face database. The database consists of color frontal, color profile and 3D VRML models of 295 subjects [11]. The following tests were conducted for performance evaluation.

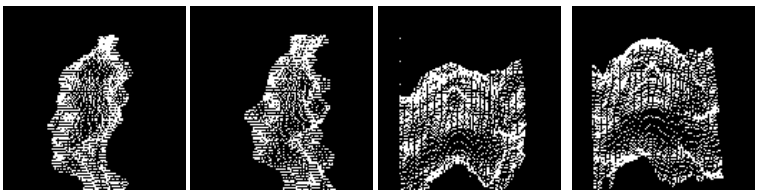


Fig. 2. 3D VRML face models

In this paper, XM2VTS database is adopted to evaluate the algorithm. XM2VTS consists of the frontal and profile views of 295 subjects. The reason for us to adopt this database is that 3D VRML model of the subjects besides 2D face images are provided in the database, which can be used to generate the depth map for our algorithm.

Generation of Depth Images. Depth image is an image where the intensity of a pixel represents the depth of the correspondent point with respect to the 3D VRML model coordinate system. 3D VRML model of a face in XM2VTS database is displayed in Fig 2. There are about 4000 points in a 3D face model to represent the face. The face surface is triangulated with these points. In order to generate a depth image, a virtual camera is put in front of the 3D VRML model, see Fig. 3. The coordinate system of the camera is defined: the image plane is defined as the X-Y plane and the Z-axis is along the optical axis of the camera and pointing toward the frontal object. The initial plane of Y_c-Z_c is positioned parallel to Y_m-X_m plane of the 3D VRML model. The projective image can be obtained using the perspective transform matrix of the camera. Z_c coincides with Z_m; however in reverse directions. X_c is parallel to X_m and Y_c parallel is to Y_m; however they are with reverse directions.

The intrinsic parameters of the camera must be properly defined in order to generate depth image from 3D VRML model. The parameters include (u_0, v_0), the coordinates of the image-center point (principle point), and f_u and f_v, scale factors of the camera along the u- and v-axis respectively. The position of the origin of the camera system, (x_0, y_0, z_0), under the 3D VRML model coordinate system is also set.

Perspective projection is assumed, i.e. for a point $F(x_m, y_m, z_m)$ in a 3D VRML model of a subject, the 2D coordinates of F in its depth image is computed as follows:

$$u = u_0 + f_u (x_m/(z_0-z_m)) \quad (18)$$

$$v = v_0 - f_v (y_m/(z_0-z_m)) \quad (19)$$

In our approach, z-buffer algorithm is applied to handle the face-self occlusion for generating the depth images.

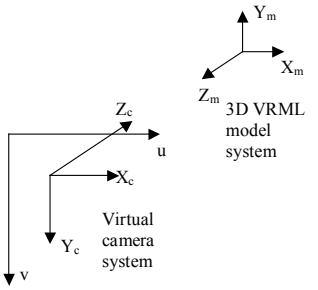


Fig. 3. Relationship among the virtual camera, 3D VRML model and the image plane

Gallery and Probe of Color and Depth Images. 512 images of 128 subjects, frontal view session of X2MVTS database are used. There are 4 images for each subject. We use any two of them for the learning gallery, while the remainder two are used as probes. There is only one 3D model for each subject. In order to generate more than one view for learning and testing, some new views are obtained by rotating the 3D

coordinates of VRML model away the frontal (Y_m) by some degrees. In our experiments, the new views obtained at 0^0, 2^0 respectively comprise the gallery, while the views at -2^0 and 4^0 comprise the probe.

Some normalized samples from XM2VTS database are shown in Fig. 4. The first 40 Eigenfaces of 2D and 3D training samples are shown in Fig. 5(a) and (b) respectively.



Fig. 4. Some normalized samples (color and corresponding depth images) from XM2VTS database



Fig. 5. The first 40 Eigenfaces of the color gallery and the first 40 Eigenfaces of the depth gallery

Recognition. Using the gallery and probe described above, the evaluation of the recognition algorithm has been done, include the recognition when the number of the eigenfaces varies from 20 to 80 with a step increment of 10. The recognition results 2D appearance, 3D depth and linear combinations of 2D+3D are given in Fig. 6 and Table 1. The order of the RMPM, r, is set to be 2. b is set to be (e^{-4}). The highest recognition rate is 98% for 2D+3D. This supports our hypothesis that the combined modality outperforms the individual modalities of appearance and depth. It also shows that each contains independent information from the other. In this experiment, the recognition rate for 3D alone is higher than the one on 2D.

Table 1. Recognition rates for 2D, 3D and 2D+3D vs. number of eigenfaces

Number of eigenfaces	Recognition rates on 2D+3D	Recognition rates on 2D	Recognition rates on 3D
20	0.957	0.621	0.676
30	0.972	0.699	0.789
40	0.973	0.785	0.855
50	0.977	0.801	0.875
60	0.977	0.863	0.895
70	0.980	0.859	0.910
80	0.977	0.863	0.922

We can see that the recognition rate has been improved by fusing appearance and depth, although the depth probe is with larger difference from the learning gallery. In our experiment, the difference is $\geq 4^0$.

The improvement of the recognition rate is between 7% (vs. 3D) and 10% (vs. 2D).

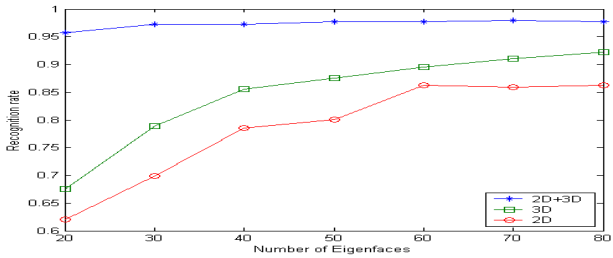


Fig. 6. Recognition rates for 2D, 3D and 2D+3D vs. number of eigenfaces

4.2 Experiment on Stereo Vision System

Encouraged by the good performance of the recognition algorithm on common database, we implemented the algorithms on a stereo vision system. We aim at identify a face by fusing disparity/depth and intensity information from a binocular stereo vision system [8], which outputs the disparity/range information automatically in real-time. Small Vision System provides 2 *mm* range resolution with ≥16 *mm* lens when the distance from the object to the stereo head is 1*m* [8]. The resolution makes it possible to combine range for face recognition. Our experiments verified this.

″new user registration″ is done automatically. When a subject is rejected by the system, i.e. when the output of the reduced model is less than 0.5, the user will be registered automatically.

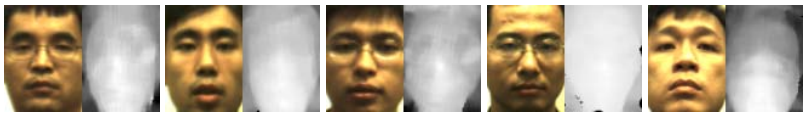


Fig. 7. Normalized color and disparity images from the stereo vision system

In order to test the recognition rate of the system, a database is built to include 30 subjects, eight images (four color and for disparity images) for each subject, two of them for training, remaining two for testing the recognition rate. Some normalized color and disparity images are shown in Figure 7. On the other hand, 40 images for five subjects, which are not included in the database, are used to test the False Acceptance Rate (FAR). The recognition rate on the database is: 95% (appearance + disparity), 81% (appearance) and 87% (disparity). The FAR is 0. The new users that are not included in the database can be registered automatically using the method described in Section 2.4.

5 Conclusion

In this paper, we contributed a face recognition formulation that combines appearance and disparity/depth. We evaluated the performance of such fusion on XM2VTS face

database. The evaluation results, which include the results from appearance alone, depth alone and fusion of them respectively, using XM2VTS database, showed the improvement of the recognition rate by combining 3D information and 2D information. The performance using fused depth and color is the best among the three tests. The improvements varies depend on the data used (appearance, depth or fusion of them) and varies between 7% and 10%.

In order to implement the algorithm on a real-time stereo vision system, near-frontal views are selected from stereo sequence for recognition. The normalization of 2D and 3D images are given.

The face recognition approach is projected to be useful for some on-line applications, such as visitor identification, ATM, HCI. Prior to such implementations in physical systems, the performance on larger database should be investigated regarding the verification accuracy. This is our future work.

References

1. C. Beumier and M. Acheroy, ″Face verification from 3D and grey level clues,″ Pattern Recognition Letters, Vol. 22, (2001), 1321-1329.
2. C. Beumier and M. Acheroy, ″Automatic face verification from 3D surface,″ in Proceedings of British Machine Vision Conference, (1998) 449-458.
3. R. Bruneli and D. Falavingna, ″Person identification using multiple cues, ″IEEE Trans. on PAMI, vol. 17, no. 10, (1995) 955-966.
4. R. Chellappa, C.L. Wilson and S. Sirohey, ″Human and machine recognition of faces,″ in Proceedings of the IEEE, Vol. 83, No. 5, May (1995), 705-740.
5. T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, Multimodal, ″Person recognition using unconstrained audio and video,″ in Proc. of AVBPA (1999) 176-181.
6. G. Gordon, ″Face Recognition Based on Depth Maps and Surface Curvature,″ in Proc. Of FGR (1996), 176-181.
7. J. Kittler, M. Hatef, R. P.W. Duin and J. Matas, ″On combining classifiers,″ in IEEE Trans. on PAMI, Vol. 20, No. 3, March (1998) 226-239.
8. Videre Design. MEGA-D Megapixel Digital Stereo Head. http://www.ai.sri.com/~konolige/svs, 2000.
9. D.A. Socolinsky, A. Selinger and J.D.Neuheisel, ″Face recognition with visible and thermal infrared imagery,″ Computer Vision and Image Understanding, Vol. 91, (2003).
10. K.-A. Toh, Q.-L. Tran and D. Srinivasan, Benchmarking a reduced multivariate polynomial pattern classifier, IEEE Trans. On PAMI, Vol. 26, No. 6, June (2004) 740-755.
11. S. Taylor and N. Cristianini, ″Kernel methods for pattern analysis″, Cambridge University Press, 2004.
12. K. Masser, J. Matas, J. Kittler, J. Luettin, and G. Maitre. ″XM2VTSDB: The extended M2VTS Database,″ In AVBPA (1999), 72-77.
13. K. I. Chang, K.W. Bowyer and P.J. Flynn, ″Multi-biometrics using facial appearance, shape and temperature,″ in Proc. FGR, (2004), 43-48.
14. F. Tsalakanidou, D. Tzovaras and M. G. Strintzis, ″Use of depth and color eigenfaces for face recognition,″ in Pattern Recognition Letters, Vol. 24 (2003) 1427-1435.
15. J.-G. Wang, H. Kong and R. Venkateswarlu, ″Improving face recognition rates by combining color and depth Fisherfaces,″ in 6th Asian Conference on Computer Vision, (2004), 126-131.
16. J.-G. Wang, E. T. Lim and R. Venkateswarlu, ″Stereo face detection/tracking and recognition,″ in IEEE International Conference on Image Processing, (2004), 638-644.

Gabor Feature Based Classification Using 2D Linear Discriminant Analysis for Face Recognition

Ming Li, Baozong Yuan, and Xiaofang Tang

Institute of Information Science, Beijing Jiaotong University, Beijing, 100044 China
liming@mail.edu.cn

Abstract. This paper introduces a novel 2D Gabor-Fisher Classifier for face recognition. The 2D-GFC method applies the 2D Fisher Linear Discriminant Analysis (2D-LDA) to the gaborfaces which is derived from the Gabor wavelets representation of face images. In our method, Gabor wavelets first derive desirable facial features characterized by spatial frequency, spatial locality, and orientation selectivity to cope with the variations due to illumination and facial expression changes. 2D-LDA is then used to enhance the face recognition performance by maximizing the Fisher's linear projection criterion. To evaluate the performance of 2D-GFC, experiments were conducted on FERET database with several other methods.

1 Introduction

Face recognition is an active research field and numerous algorithms are developed. The detailed introduction to this field can be found in paper [1][2][3]. Face recognition's task is to compare an input image (probe) against a database (gallery) then reports a match.

The Gabor wavelets, whose kernels are similar to the two-dimensional (2-D) receptive field profiles of the mammalian cortical simple cells, exhibit desirable characteristics of spatial locality and orientation selectivity [4]. It is also one of the most successful approaches for face recognition [5–9]. Despite Gabor wavelet is a very powerful method for feature extraction, there still exists a common drawback in it as in most other methods. The Gabor feature is a very high-dimensional data. For example, if the size of face images is , and five scales and eight orientations Gabor filters are applied on these images, a dimension data is used to represent a face image. Considering each dimension of Gabor representation is combined by real part and imagery part, we need to use more than 1310720-D data to store and compute a face image. To overcome this problem, several statistical methods for feature extraction are used, such as PCA [7], the Enhanced Fisher linear discriminant Model (EFM) [8] and the Kernel Direct Discriminant Analysis (KDDA) [9]. Yang et al.'s [10] work is very interesting: they used Ada-boost algorithm to select the most disciminant Gabor features to recognize human faces.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 929–936, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Traditional statistical feature extraction approaches such as principal component analysis (PCA) and linear discriminant analysis (LDA) are based on the analysis of vectors. When applied these method to the Gabor wavelets representation of face images, the high dimensionality will become a disaster.

To resolve above problems, we first mosaic the gabor faces to a Gabor feature matrix, then apply 2D-LDA [14] to the Gabor feature matrices, which calculates the between-class scatter matrix and the within-class scatter matrix directly based on the Gabor feature matrices. This strategy makes our classifier more efficient and more robust in large lighting and facial expression variation cases.

The organization of this paper is as follows: In Section 2, the Gabor wavelets representation of face images is introduced. Section 3 describes the 2D-LDA algorithm. Experiments and analysis are conducted in Section 4. Some conclusions are given in Section 5.

2 Gaborfaces

Gabor wavelets model quite well the receptive field profiles of cortical simple cells [11]. The Gabor wavelet representation, therefore, captures salient visual properties such as spatial localization, orientation selectivity, spatial frequency characteristic. The Gabor wavelets (kernels, filters) can be defined as follows [5] [8]:

$$\psi_{u,v}(z) = \frac{\| k_{u,v} \|^2}{\sigma^2} e^{\left(-\|k_{u,v}\|^2 \|z\|^2 / 2\sigma^2\right)} \left[e^{jk_{u,v}z} - e^{-\sigma^2/2} \right] \tag{1}$$

where u and v define the orientation and scale of the Gabor kernels, $z = (x, y)$, $\|\cdot\|$ denotes the norm operator, and the wave vector is defined as follows:

$$k_{u,v} = k_v e^{j\phi_u} \tag{2}$$

where $k_v = k_{max}/f^v$ and $phi_u = u\pi/8$, $\phi_u \in [0, \pi)$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain.

In most cases [6][8][10], Gabor wavelets with five different scales and eight orientations $u = \{0, 1, 2, 3, 4, 5, 6, 7\}$ is used. We also choose other parameters as: $\sigma = 2\pi$, $k_max = \pi/2$, and $f = \sqrt{2}$.

The Gabor wavelet representation of an image is the convolution of the image with a family of Gabor kernels as defined by (1).

Fig. 1 (a) shows the Gabor wavelets at different scales and orientations. Fig. 1 (b) is a 128×128 face image, and (c) is its Gaborfaces.

In traditional methods, the Gabor feature is obtained by following steps: transform each gaborface to a gabor feature vector, then concatenat all these gabor feature vector to derive an augmented feature vector. Then, based on these augmented gabor feature vector, a classifier can be obtained [8][9][10]. The Gabor feature vector is a very high-dimension pattern. Different to traditional approach, we mosaic the gaborfaces to a large *Gabor feature matrix*. Then, directly based on this Gabor feature matrix, we apply the 2DLDA algorithm. Fig.2 shows is an example of Gabor feature matrix.

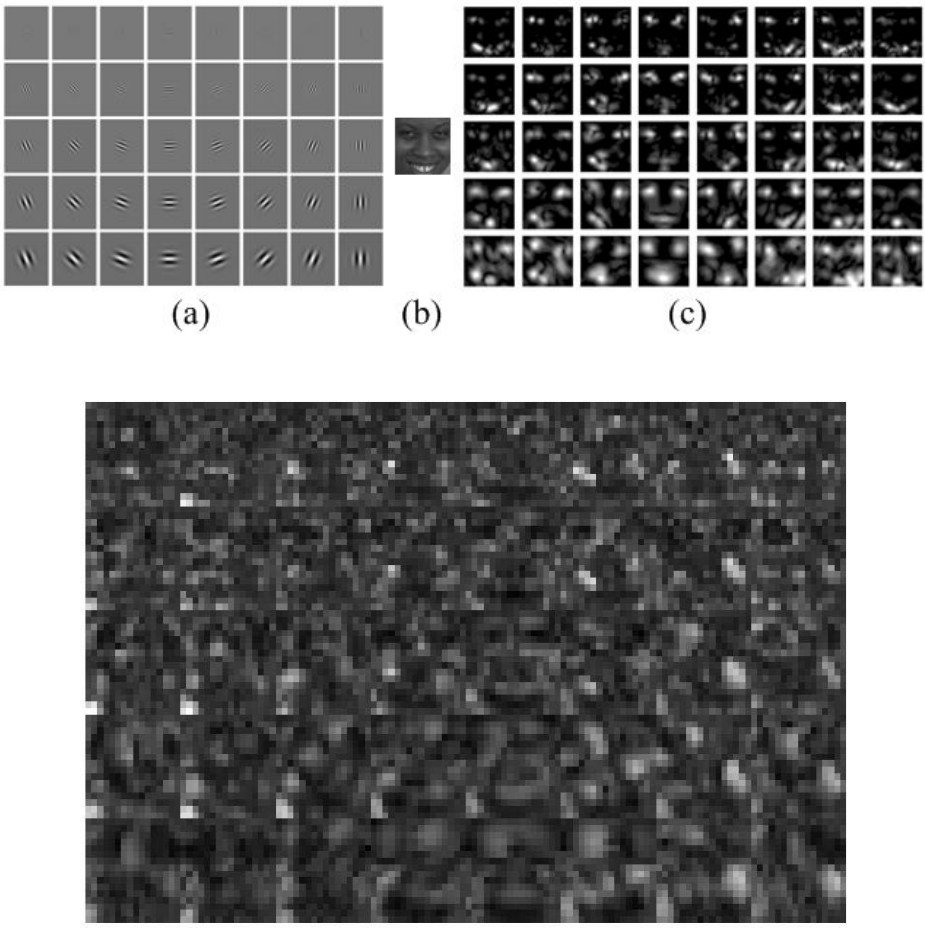


Fig. 2. Gabor wavelets and Gabor wavelets representation of a face image

3 2D Gabor-Fisher Classifier

This paper presents a novel method which applies the 2D Fisher Discriminant Analysis to the Gabor feature matrix **A** derived in Section 2. We can see that the Gabor feature matrix **A** is a very high-dimensional data. If we classify the face images in such high dimension directly, *Curse of Dimensionality* [12] is the problem we must face to. So, dimension reduction on these raw Gabor feature data is necessary.

3.1 Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) [13] is a very famous method for feature extraction. It tries to find the subspace that best discriminates the samples coming from difference classes by maximizing the between-class scatter matrix

$\mathbf{S}_b$, while minimizing the within-class scatter matrix $\mathbf{S}_w$ in the projective subspace. $\mathbf{S}_b$ and $\mathbf{S}_w$ are defined as

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{\mathbf{x}_k \in P_i} (\mathbf{x}_k - \bar{\mathbf{x}}_i)(\mathbf{x}_k - \bar{\mathbf{x}}_i)^T \tag{3}$$

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T \tag{4}$$

where $\bar{\mathbf{x}}_i$ is the mean vector for class P_i and N_i is the number of samples in class P_i. The optimal classification matrix $\mathbf{W}$ satisfies

$$\mathbf{W} = \arg\max \left| \frac{\mathbf{W}^T \mathbf{S}_b \mathbf{W}}{\mathbf{W}^T \mathbf{S}_w \mathbf{W}} \right| \tag{5}$$

$\mathbf{W}$ can therefore be constructed by the eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

While this procedure can be realized easily, problems arise when dealing with high-dimensional data, such as images. The main difficulty in this case lies in the fact that the within-class scatter matrix is almost always singular. The second difficulty is that the traditional LDA is based on the analysis of the very high-dimensional vectors , which are transformed from the image data. This means that doing LDA on such a high-dimensional vector is very inefficient [14].

3.2 2D Gabor Fisher Classifier

From above analysis, we can see that the LDA can only handle the vectors. So, we must transform a image into vector before applying it. This strategy will bring the Curse of Dimensionality problem, especially for Gabor feature vector. To overcome this problem, we propose a novel approach, 2D Gabor-Fisher Classifier (2D-GFC), which is directly based on the analysis of Gabor feature matrices to find out an optimal projective subspace.

In literature [14], the idea of 2D Linear Discriminant Analysis (2D-LDA) is proposed. Based on this work, we we use an n-dimensional column vector $\mathbf{x}$ to project the given $m \times n$ random Gabor feature matrix $\mathbf{A}$ onto a m-dimensional feature vector $\mathbf{y}$:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{6}$$

So, we have the between-class scatter matrix $\mathbf{S}_B$ and within-class scatter matrix $\mathbf{S}_W$ as follow,

$$\mathbf{S}_B = \sum_{i=1}^{C} N_i \left[(\bar{\mathbf{A}}_i - \bar{\mathbf{A}})\mathbf{x}\right] \left[(\bar{\mathbf{A}}_i - \bar{\mathbf{A}})\mathbf{x}\right]^T, \tag{7}$$

$$\mathbf{S}_W = \sum_{i=1}^{C} \sum_{\mathbf{A}_k \in P_i} \left[(\mathbf{A}_k - \bar{\mathbf{A}}_i)\mathbf{x}\right] \left[(\mathbf{A}_k - \bar{\mathbf{A}}_i)\mathbf{x}\right]^T, \tag{8}$$

The optimal projection $\mathbf{x}_{opt}$ is chosen when the Fisher's linear projection criterion is maximized, i.e.,

$$\mathbf{x}_{opt} = \arg\max_{\mathbf{x}} \frac{\mathbf{x}^T \mathbf{S}'_B \mathbf{x}}{\mathbf{x}^T \mathbf{S}'_W \mathbf{x}}, \tag{9}$$

where,

$$\mathbf{S}'_B = \sum_{i=1}^{C} N_i \left(\bar{\mathbf{A}}_i - \bar{\mathbf{A}}\right)^T \left(\bar{\mathbf{A}}_i - \bar{\mathbf{A}}\right), \tag{10}$$

$$\mathbf{S}'_W = \sum_{i=1}^{C} \sum_{\mathbf{A}_k \in P_i} \left(\mathbf{A}_k - \bar{\mathbf{A}}_i\right)^T \left(\mathbf{A}_k - \bar{\mathbf{A}}_i\right), \tag{11}$$

So, equation (9) is equivalent to solve the generalized eigenvalue problem:

$$\mathbf{S}'_B \mathbf{x}_{opt} = \lambda \mathbf{S}'_W \mathbf{x}_{opt} \tag{12}$$

In above equation, λ is the maximal eigenvalue of $\mathbf{S}'^{-1}_W \mathbf{S}'_B$.

4 Experiments and Analysis

In this section, we analyze the performance of the proposed 2D-GFC method using the FERET database. Comparative results are also given for other methods, such as PCA, LDA, Kernel PCA (KPCA), Generalized Discriminant Analysis (GDA), and Kernel Direct Discriminant Analysis (KDDA). The FERET database is a standard testbed for face recognition technologies [15]. The data set used in our experments consists of 600 FERET frontal face images corresponding to 200 subjects, so that each subject has three images with size 256×384 and gray-scale 256 levels. The face images are acquired under varying illumination and facial expressions. Data preparation is needed because the performance of face recognition is maybe affected by the factors unrelated to face such as hairstyles. The following procedures are applied to normalize the face images prior to further experiments: first, the center of the eye in each image is marked manually; second, each image is rotated and scaled to align the center of the eyes; finally, each face image is cropped to the size 128×128 to exact facial region, which is normalized to zero mean and unit variance. Fig. 2 shows some example FERET images used in our experiments that are already cropped to the size of 128×128 to extract the face region. The first two rows are sample images from the FERET database, while the third row shows the examples of our test images.

To prove the advantage of the 2D-GFC method, we compare our method with some other well-known methods: Gabor+PCA, Gabor+LDA, Gabor+KPCA, Gabor+GDA, and Gabor+KDDA [9]. Table 1 shows the experimental results of these six methods. From Table 1, we can see that the 2D-GFC method achieves the best recognition accuracy, 95.5%. We can also see that kernel methods achieve better performance than the linear ones in the 1D situation; and the methods with applies the discriminant analysis are more powerful.

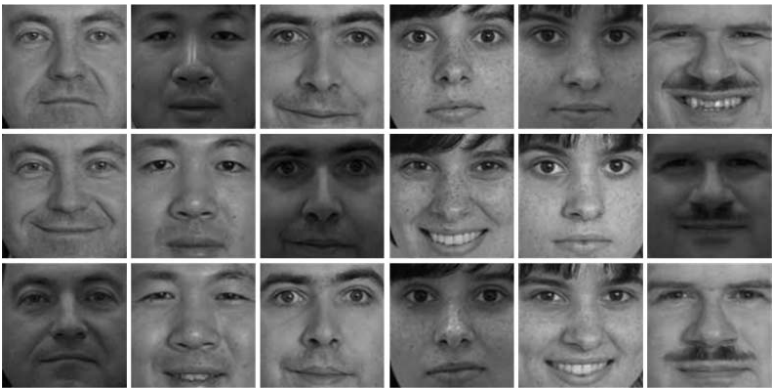


Fig. 3. Gabor wavelets and Gabor wavelets representation of a face image

Table 1. Comparative Recognition performance

Classifier	Recognition Rate(%)
Gabor+PCA	57
Gabor+LDA	76.5
Gabor+KPCA	67
Gabor+GDA	92
Gabor+KDDA	95
2D-GFC	95.5

And from [14] we know that the 2D-LDA is more efficient than 1D ones, and the nonlinear methods which applies kernel function are much more time-consuming than the linear ones. Table 2 shows the comparative computation-cost of these six methods. From Table 2, we can see that 2D-GFC is the most efficient method. This is because that 2D-GFC is only need to handle a matrix rather than a matrix in PCA and LDA.

So, we can conclude that the 2D-GFC is an efficient and high performance method.

Table 2. Comparative Computation Cost

Classifier	Time cost (s)
Gabor+PCA	47.750
Gabor+LDA	54.368
Gabor+KPCA	76.406
Gabor+GDA	308.141
Gabor+KDDA	243.359
2D-GFC	5.285

5 Conclusions

This paper introduces a novel 2D Gabor-Fisher Classifier for face recognition. The 2D-GFC method applies the 2D Fisher Linear Discriminant Analysis (2D-LDA) to the gaborfaces which is derived from the Gabor wavelets representation of face images. Gabor wavelets first derive desirable facial features characterized by spatial frequency, spatial locality, and orientation selectivity to cope with the variations due to illumination and facial expression changes. 2D-LDA is then used to enhance the face recognition performance by maximizing the Fisher's linear projection criterion. The feasibility of the new 2D-GFC on face recognition has been successfully demonstrated using a data set from the FERET database. Our 2D-GFC method has achieved the best performance on the FERET database: 95.5%. Our experiments also prove that the 2D-GFC is very efficient.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 60441002) and the University Key Research Project (No. 2003SZ002).

References

1. A. Samal and P. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition*, vol. 25, pp. 65-77, 1992.
2. R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol.83, no. 5, pp 705-740, 1995.
3. W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face Recognition: A Literature Survey," http://citeseer.nj.nec.com/cs., 2000.
4. R. DeValois and K.DeVolois, *Spatial Vision*, Oxford Press, 1998.
5. M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.*, vol. 42, pp. 300-311, 1993.
6. L. Wiskott, J.M. Fellous, N.Kruger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, July 1997, 775-779.
7. C. Ki-Chung, K. S. Cheol, K. S. Ryong, "Face Recognition Using Principal Component Analysis of Gabor Filter Responses", *Proceedings of International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pp. 53-57, 1999.
8. C. Liu, and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, Apr. 2002, 467-476.
9. L. Shen, L. Bai, "Gabor wavelets and kernel direct discriminant analysis for face recognition," *Proc. 17th Int'l Conf on Pattern Recognition* (ICPR'04), 2004, 284-287.
10. P. Yang, S. Shan, W. Gao, S.Z. Li, D. Zhang, "Face Recognition Using Ada-Boosted Gabor Features," *Proc. 6th IEEE Int'l Conf on Automatic Face and Gesture Recognition*(FGR'04), 2004, 356-361.

11. D.J. Field, "What is the goal of sensory coding," *Neural Comput.*, vol. 6, pp. 559-601, 1994.
12. R. Bellman, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.
13. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.
14. M. Li, B. Yuan, "A Novel Statistical Linear Discriminant Analysis for Image Matrix: Two-Dimensional Fisherfaces," *Proc. 7th Int'l Conf on Signal Processing*(ICSP'04), 2004.
15. P. Phililips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face-Recognition Algorithms," I*mage and Vision Computing*, Vol. 16, pp. 295-306, 1998.

Multi-resolution Histograms of Local Variation Patterns (MHLVP) for Robust Face Recognition

Wenchao Zhang[1], Shiguang Shan[2], Hongming Zhang[1], Wen Gao[2], and Xilin Chen[1,2]

[1] School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, P.R. China
{wczhang,hmzhang,xlchen}@jdl.ac.cn
[2] ICT-ISVISION Joint R&D Laboratory for Face Recognition, CAS, 100080 Beijing, P.R. China
{sgshan,wgao}@jdl.ac.cn

Abstract. This paper presents a novel approach to face recognition, named Multi-resolution Histograms of Local Variation Patterns (MHLVP), in which face images are represented as the concatenation of the local spatial histogram of local variation patterns computed from the multi-resolution Gabor features. For a face image with abundant texture and shape information, a Gabor feature map(GFM) is computed by convolving the image with each of the forty multi-scale and multi-orientation Gabor filters. Each GFM is then divided into small non-overlapping regions to enhance its shape information, and then Local Binary Pattern (LBP) histograms are extracted for each region and concatenated into a feature histogram to enhance the texture information in the specific GFM. Further more, all of the feature histograms extracted from the forty GFMs are further concatenated into a single feature histogram as the final representation of the given face image. Eventually, the identification is achieved by histogram intersection operation. Our experimental results on FERET face databases show that the proposed method performs terrifically better than the performance of some classical results including the best results in FERET'97.

1 Introduction

As one of the most successful applications of image analysis and understanding, face recognition has received significant attention in both the wide range of potential applications [1] and scientific challenges [2][3]. Although many commercial face recognition systems have emerged, it is still an active topic in computer vision community. This is partly due to the fact that face recognition is still very challenging in uncontrolled environments with rich variations of pose, illumination etc. Therefore, the goal of the on-going research is to increase the robustness of face recognition systems to these variations. Moreover, to evaluate the effectiveness of different algorithms of face recognition, some evaluation methodologies and face databases have been created, such as the FERET database and protocol [4], which attract more and more researchers to make the further progress on algorithm.

Most of the prevalent approaches for face recognition are based on statistic analysis, such as eigenfaces [5], Fisherfaces [6][7] and Bayesian methods [8]. The methods often need a great deal of data to train, but usually the data available for training are very few. In addition, if the distributions of the test examples are different from that of the training examples, the generality of the method will be weakened.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 937–944, 2005.
© Springer-Verlag Berlin Heidelberg 2005

In this paper, we propose a novel approach for face recognition based on non-statistical model, which named Multi-resolution Histograms of Local Variation Patterns (MHLVP), in which face images are represented as the concatenation of the local feature histogram of multi-resolution Gabor features.

Histograms have been widely used to represent, analyze and characterize images because they could be computed easily and efficiently and their robustness to noise and local image transformations [9]. But they do not capture spatial image information. The multi-resolution histograms [10], however, not only preserves the efficiency, simplicity and robustness of the plain histogram, but also combines intensity with spatial information. Moreover, the Gabor filters combine the spatial and frequency localization [11], which are effective to face image representations [12]. In this work, given a face image, we get forty GFMs computed by convolving the image with each of the multi-scale and multi-orientation Gabor filters. The multi-resolution histograms are extracted from the GFMs represented by the magnitude of the Gabor filters' response, which can provide a measure of the image's local properties [13]. For a face image with abundant texture and shape information, we divided the face image into some certain regions to improve the shape information of the face image represented by the multi-resolution histograms. Also we extracted LBP histograms [14][15] from the GFMs to improve the texture information. Thus, the MHLVP can be used as a texture and shape descriptor for face image.

This paper is organized as follows: In Section 2, the face description with multi-resolution Gabor representations is briefly described firstly, then the local feature histogram is introduced. Experimental design and results in Section 3. Finally the discussion and the conclusion can be found in Section 4.

2 Face Description with Multi-resolution Histograms of Local Variation Patterns

The overall framework of the proposed representation approached MHLVP-based is illustrated in Fig.1. In this method, a face image is modeled by the following procedure: (1) An input face image is normalized and transformed to obtain multiple GFMs in frequency domain by applying multi-scale and multi-orientation Gabor filters; (2) Each GFM is labeled with LBP operator; (3) Each transformation of the GFM is further divided into non-overlapping rectangle regions with specific size, and histogram is computed for each region; (4) All the histograms are concatenated to form the final histogram sequence as the model of the face. The following sub-sections will describe the procedure in detail.

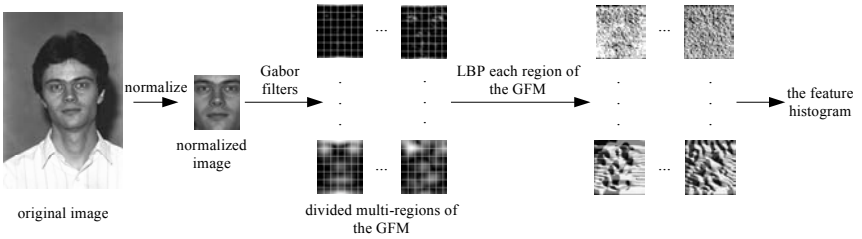


Fig. 1. The framework of Multi-resolution Histograms of Local Variation Patterns

2.1 Face Description with Multi-resolution Gabor Representations

The Gabor feature is effective to face image representation. The multi-resolution description of an image is computed with Gabor filters. The Gabor wavelets (filters, kernels) can be defined as follows, assuming that $\sigma_x = \sigma_y = \sigma$ [11][13]:

$$\Psi(x,y,\varpi_0,\theta) = \frac{1}{2\pi\sigma^2} e^{-\left((x\cos\theta+y\sin\theta)^2+(-x\sin\theta+y\cos\theta)^2\right)/2\sigma^2} \times \left[e^{i(\varpi_0 x\cos\theta+\varpi_0 y\sin\theta)} - e^{-\varpi_0^2\sigma^2/2}\right], \quad (1)$$

where x, y define the pixel position in the spatial domain, ϖ_0 the radial center frequency, θ the orientation of the Gabor wavelet, and σ the standard deviation of the Gaussian function along the x-and y-axes. In addition, the second term of the Gabor wavelet, $e^{-\varpi_0^2\sigma^2/2}$, compensates for the DC value because the cosine component has nonzero mean (DC response) while the sine component has zero mean.

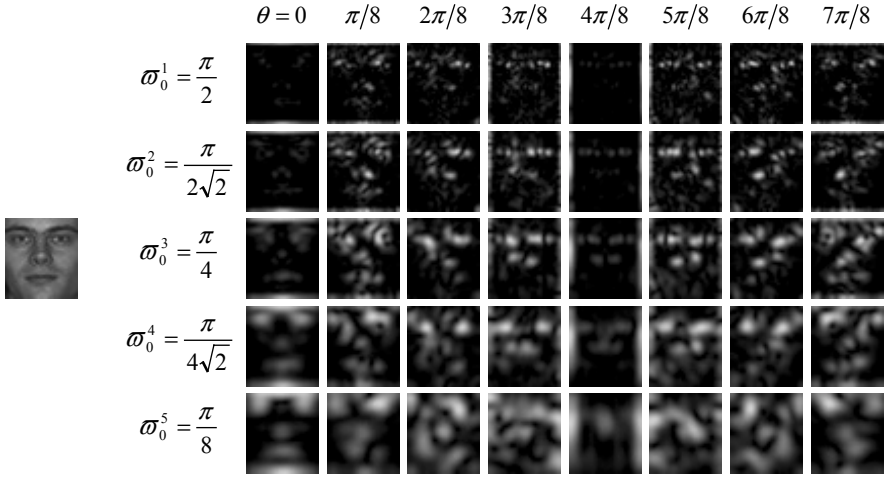
The Gabor representation of an image can be derived by convolving the image and the Gabor wavelets. Let $f(x,y)$ denote the gray level distribution of an image, then the convolution of an image $f(x,y)$ and a Gabor kernel $\Psi(x,y,\varpi_0,\theta)$ is defined as follows:

$$G_{\Psi f}(x,y,\varpi_0,\theta) = f(x,y) * \Psi(x,y,\varpi_0,\theta), \quad (2)$$

where $*$ denotes the convolution operator. With a set of ϖ_0 and θ, a multi-hierarchical Gabor wavelets representation of the face image $f(x,y)$ is composed. The visualizations of the magnitudes of the GFMs in this paper are shown in Figure 2. We select five scales and eight orientations in the Gabor filters.

2.2 Local Variation Patterns

To improve the local feature information, we divided the image into small non-overlapping regions from which LBP histograms are extracted.



(a) Face image (b) The GFMs

Fig. 2. The visualizations of the magnitudes of the GFMs

The original LBP operator, introduced in [14], labels the pixels of an image by thresholding the 3×3 -nieghborhood of each pixel $f_p(p=0,1,\ldots,7)$ with the center value f_c and considering the result as a binary number (3).

$$S(f_p - f_c)=\begin{cases}1, & f_p \geq f_c \\ 0, & f_p < f_c\end{cases}. \tag{3}$$

Then, by assigning a binomial factor 2^p for each $S(f_p - f_c)$, the LBP number is achieved as

$$LBP = \sum_{p=0}^{7} S(f_p - f_c)2^p, \tag{4}$$

which characterizes the spatial structure of the local image texture. Figure 3 shows an example of the basic LBP operator, and the transform result of an image is shown in Figure 4. The LBP images of the GFMs are shown in Figure 5.

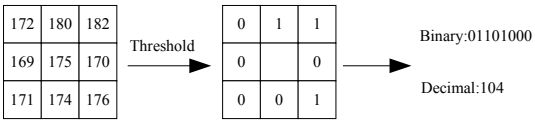


Fig. 3. Example of the basic LBP operator

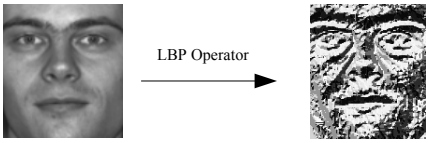


Fig. 4. The face image and the LBP image

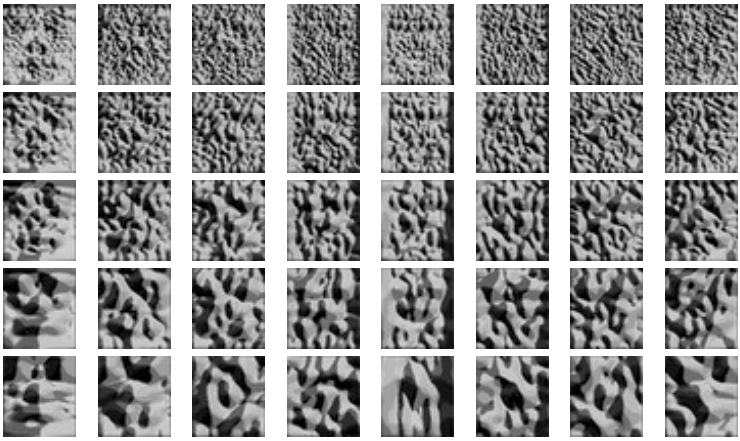


Fig. 5. The LBP images of the GFMs

2.3 Multi-resolution Histograms of Local Variation Patterns Description

The histogram h_f of the image $f(x,y)$ could be defined by (5) which provides the frequency of different values in the image.

$$h_{i,f} = \sum_{x,y} I\{f(x,y) = i\}, i = 0, 1, \ldots, n-1, \tag{5}$$

n is the number of different values of the image and

$$I\{A\} = \begin{cases} 1, A \text{ is true} \\ 0, A \text{ is false}. \end{cases} \tag{6}$$

From the description above, we can get the multi-resolution histograms of local variation patterns description, and each of the Gabor filter image $G_{\psi f}(x,y,\varpi_0,\theta)$ is divided into m regions of $R_0, R_1, \ldots, R_{m-1}$ from which the LBP histograms are extracted (7).

$$h_{i,j,G_{\psi f}} = \sum_{x,y} I\{G_{\psi f}(x,y,\varpi_0,\theta) = i\} I\{(x,y) \in R_j\} i = 0, 1, \ldots, n-1; \; j = 0, 1, \ldots, m-1. \tag{7}$$

Many dissimilarity measures have been proposed for histogram. In this paper, Histogram intersection $D(Sh_1, Sh_2)$ is used as the similarity matching between two histograms [17].

$$D(Sh_1, Sh_2) = \sum_{i=1}^{k} \min(Sh_1^i, Sh_2^i), \tag{8}$$

where Sh_1 and Sh_2 are two histograms, and k is the number of bins in the histogram.

3 Experimental Design and Results

To achieve a fair comparison, we test our face recognition algorithm on the FERET face database and protocol, which have been used widely to evaluate face recognition algorithms and are a de faceto standard in face recognition research field [20].

FERET provided a training set containing 1002 images, and a gallery consisted of images of 1,196 people with one image per person. Four probe sets, fb, fc, Dup.I, and Dup.II are provided for testing. In the fb probe set, there are 1195 images with different facial expressions from the gallery set. In the fc probe set, there are 194 images taken under different lighting condition from that of fa. In the probe categories Dup.I and Dup.II, there are 722 and 234 images taken respectively a month and a year later than the gallery.

To test the robustness of the method against different facial expression, lighting and aging, we do the experiments on all the four probe sets. In addition, to demonstrate the validity of our method, we compared the performance of our method with Correlation [18], LDA [19] and LBP. The similarity measure used in LDA and Correlation algorithms in our experiments to compare feature vectors is the normalized correlation, which is defined as (9).

$$\delta(\mathbf{x}_1, \mathbf{x}_2) = \frac{-\mathbf{x}_1^T \mathbf{x}_2}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}. \tag{9}$$

In the LBP method, we use the basic LBP operator and the same divided region form as that of the MHLVP in Figure 5. The nearest neighbor rule is then used to classify the face images. The recognition results of our proposed method are shown in Table 1 and the Cumulative Match Score (CMS) curves are plotted in Figures 6 (a)-(d). The MHLVP-based method outperforms clearly the other methods on all of the test sets. In the three test sets of fc, duplicate I and duplicate II, the results are even better than the best results reported in the FERET'97 evaluation [20] and that of [21]. It should be noted that the results of the basic LBP method might be different from the results mentioned in [21] due to the different forms of image division and LBP operators.

Table 1. The rank-1 recognition rates of different algorithms for the FERET probe sets

Method	fb	fc	duplicate I	duplicate II
MHLVP	0.942	0.959	0.676	0.594
LBP	0.947	0.294	0.536	0.269
LDA	0.934	0.727	0.551	0.312
Correlation	0.700	0.268	0.276	0.094
Best Results of FERET'97[20]	0.96	0.82	0.59	0.52

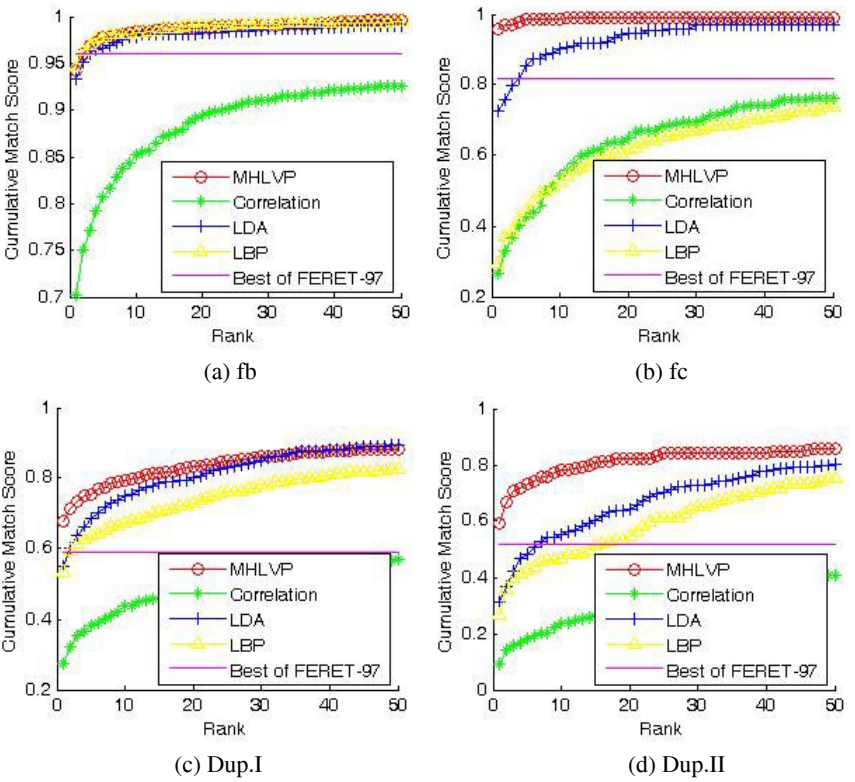


Fig. 6. Comparison of several methods on the FERET fb, fc, Dup.I, and Dup.II probe sets. Note that the FERET line just illustrated the best rank-1 recognition rate in the FERET'97 results

4 Discussion and Conclusion

This paper proposes a novel face representation, MHLVP, which is impressively insentative to appearance variations due to lighting, expression and aging. Moreover, the modeling procedure of MHLVP does not involve in any learning process, that is, it is non-statistical learning based. The effectiveness of the MHLVP comes from several aspects including the multi-resolution and multi-orientation Gabor decomposition, the LBP and the local spatial histogram modeling. The experimental results on FERET face database have evidently shown that the MHLVP method performs terrifically better than other approaches for all four standard probe sets. In the three test sets of fc, duplicate I and duplicate II, the results are even better than that of the best reported on the FERET evaluation.

Some improvements may further be made by optimizing the bins of the histogram to reduce the dimension the feature vector, and/or selecting the different scales and orientations Gabor filters to represent the shape and texture information of the face image. In addition, we could divide images into different amount of regions according to the variations of different scales.

Acknowledgements

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, "100 Talents Program" of CAS, ShangHai Municipal Sciences and Technology Committee (No.03DZ15013), and ISVISION Technologies Co., Ltd.

References

1. Phillips, P.J., Grother, P., Micheals, R.J., Blackburn, D.M., Tabassi, E., Bone, J.M.: Face recognition vendor test 2002 results. Technical report (2003)
2. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and Machine Recognition of faces: A survey. Proc. of the IEEE 83 (1995) 705-740
3. Zhao, W.Y., Chellappa, R., Phillips, P.J. and Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Survey (2003) 399-458
4. Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face recognition algorithms. Image and Vision Computing 16 (1998) 295-306
5. Turk, M. and Pentland, A.: Face recognition using eigenfaces. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (1991) 586-591
6. Belhumer, P., Hespanha, P., and Kriegman, D.: Eigenfaecs vs fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 711-720
7. Swets, D.L. and Weng, J.: Using Discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 831-836
8. Moghaddam, B., Nastar, C., Pentland, A.: A Bayesian similarity measure for direct image matching. In: 13th International Conference on Pattern Recognition (1996) II: 350-358
9. Hadjidemetriou, E., Grossberg, M.D. and Nayar, S. K.: Multiresolution Histograms and Their Use for Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 831-847
10. Hadjidemetriou, E., Grossberg, M. D., and Nayar, S. K.: Spatial Information in Multiresolution Histograms. Proc. Computer Vision and Pattern Recognition Conf. (2001) I: 702-709

11. Porat, M. and Zeevi, Y.: The Generalized Gabor Scheme of Image Representation in Biological and Machine Vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (1988) 452-468
12. Liu, C.J., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant modal for face recognition image processing. IEEE Transactions on Image Process 11 (2002) 467-476
13. Lades, M., Vorbrüggen, J.C., Buhmann, J., Lange, J., Malsburg, C., Würtz, R.P., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers 42 (1993) 300-311
14. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29 (1996) 51-59
15. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 971-987
16. Daugman, J.G.: Two-dimensional spectral analysis of cortical receptive field profile. Vision Research 20 (1980) 847-856
17. Swain, M., Ballard, D.: Color indexing. Int.J.Computer Vision 7 (1991) 11-32
18. Brunelli, R., and Poggio, T.: Face Recognition: Features vs. Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (1993) 1042-1053
19. Duda, R., Hart, P., Stork, D., Pattern Classification. Wiley Interscience, USA, Second Edition (2001)
20. Phillips, P.J. Syed, H.M., Rizvi, A. and Rauss, P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1090-1104
21. Timo, A., Abdenour, H. and Matti, P.: Face Recognition with Local Binary Patterns. ECCV 2004 Proceeding, Lecture Notes in Computer Science 3021, Springer (2004) 469-481

Analysis of Response Performance Characteristics for Identification Using a Matching Score Generation Model

Takuji Maeda[1], Masahito Matsushita[1], Koichi Sasakawa[1], and Yasushi Yagi[2]

[1] Advanced Technology R&D Center, Mitsubishi Electric Corporation, Japan
[2] The Institute of Scientific and Industrial Research, Osaka University, Japan

Abstract. To make person authentication systems be more useful and practical, we have developed an identification algorithm, and also showed that the authentication accuracy depends on response performance. Current identification algorithm employs a comparison computation function that is optimized for one-to-one comparison. By optimizing a comparison computation function, however, it might be possible to improve response performance. In this paper, we describe design guidelines for a comparison computation function for improving the response performance of the identification. To show the guidelines, we clarify the relation between the characteristics of a matching score distribution and response performance using a matching score generation model, and also demonstrate the effectiveness of the design guidelines with a simulation using an example of another comparison computation function.

1 Introduction

To make biometrics authentication systems be more useful, we have developed a general algorithm for person identification that controls the search order only on the basis of the comparison of the matching score of sets of biometrics data[1]. Its response performance is better than that of conventional methods. There are some approaches to obtain quick response in an identificatoin. One is realized by clustering fingerprint data[2–6]. Those approaches have problems that clustering method should be defined for each kind of biometrics, and mis-clustering causes slow response. Another is by decrease the match candidate data by some evaluation function[7, 8]. Its problem is that evaluation function depends on the quality, distortion, and translation of biometrics images.

The response performance of identification techniques affects more than user convenience, however; we have earlier shown that it also greatly affects authentication accuracy[9], so it promises a faster method of searching for an individual's enrollment data. The comparison computation function used in current identification algorithm is optimized for authentication accuracy in one-to-one comparison, and thus is not necessarily optimum for identification response performance. Here, we describe comparison computation function design guidelines for improving the response performance of the identification operation.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 945–955, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Identification Algorithm

2.1 Overview of the Algorithm

At first it is described that the principle of our identification algorithm. To reduce the response time of the identification operation, we employ a matching score matrix. It is constructed at the time of enrollment. The matrix consists of the round-robin selection other-person matching scores of the enrollment data.

An unknown user data is input at the time of identification. The initial search involves finding suitable data from the enrollment data (the data at the top, for example) to serve as a matching candidate, and then doing one-to-one comparison on the data. If the matching score thus obtained is equal to or greater than the predetermined threshold value, the candidate data is taken to be the enrollment data for the unknown user and the identification processing is ended. If, on the other hand, the matching score does not meet the threshold, it is necessary to select the next match candidate. The procedure for doing that is described below.

In the mth round of the search process, the $\boldsymbol{y_u}(m)$ defined by Eq.(1) and the $\boldsymbol{x_i}(m)$ defined by Eq.(2) are used to obtain the $z_i(m)$ defined by Eq.(3) for all i except those for which comparison has already been done. And the next candidate will be determined as i_{max} ($r(m+1) = i_{max}$) which maximizes $z_i(m)$ for i. Here, the term $r(m)$ is the mth match candidate, $y_{u,r(m)}$ is a matching score for the input unknown user u and the mth match candidate $r(m)$, and $x_{i,r(m)}$ is $(i, r(m))$ component of the matching score matrix.

$$\boldsymbol{y_u}(m) = \{y_{u,r(1)}, y_{u,r(2)}, \ldots, y_{u,r(m)}\} \tag{1}$$

$$\boldsymbol{x_i}(m) = \{x_{i,r(1)}, x_{i,r(2)}, \ldots, x_{i,r(m)}\} \tag{2}$$

$$z_i(m) = \frac{2\boldsymbol{x_i}(m) \cdot \boldsymbol{y_u}(m)}{\boldsymbol{x_i}(m) \cdot \boldsymbol{x_i}(m) + \boldsymbol{y_u}(m) \cdot \boldsymbol{y_u}(m)} \tag{3}$$

By repeating the above procedure until $y_{u,r(m)}$ satisfies the threshold, it is possible to determine the enrollment data which corresponds to the input unknown user.

2.2 Simulation Example

The result of an identification simulation is shown in Fig.1 as the distribution of the number of search steps. In this simulation, 2000 fingerprint data is enrolled. The horizontal axis is the number of search steps. The vertical axis on the left is the relative frequency presented as a bar graph and the vertical axis on the right is the cumulative frequency depicted by a solid line. The horizontal and right vertical axes have a log scale. Here, the number of search steps refers to how many search rounds are required until the enrollment data that corresponds to the unknown input data appears as a match candidate in the identification process. We can evaluate response performance by this distribution.

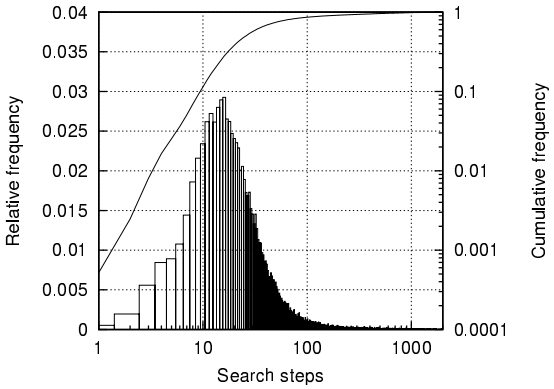


Fig. 1. Search steps distribution for fingerprint

As we see in Fig.1, in most cases the number of search steps is around 10 to 20. This result means that response is very fast. And also, the cumulative distribution saturates to 100% at around 200 steps. If we make use of this characteristic to set a search cut-off at 200 steps of search, it is possible to determine that the user who is not in the database is not enrolled by the 200-th step.

3 Response Performance with a Matching Score Generation Model

In the proposed identification algorithm, the enrollment data of the person inputting the data becomes a match candidate at an early round in the search process. This characteristic contributes to accuracy in an identification[9]. That is to say, improvement in response performance is related to higher authentication accuracy, so even further improvement in response performance is a necessity. If the relation between the matching score generation characteristics and the response performance were clarified, it would be possible to design a comparison computation function for improving the response performance. We therefore assume a hypothetical biometrics and analyze the relation between the matching score generation characteristics and the response performance.

3.1 A Matching Score Generation Model

As was explained in previous section, our proposed search algorithm decides the next match candidate according to the matching score with other-person data. Thus, if the matching score of two sets of data can be defined, it could in principle be applied to any type of biometrics. Furthermore, we believe generalizing this idea would make it applicable to a hypothesis model for generating matching scores.

We believe that identification response performance is affected by the occurrence characteristics of other-person matching scores. To show what special characteristics are associated with the occurrence of other-person matching scores in

actual biometrics, the frequency distributions of other-person matching scores are presented in Fig.2. In that figure, the bar graph represents the frequency distribution of the other-person matching scores and the broken line represents a normal distribution. It was found that the other-person matching score distribution fits the normal distribution by chi-square test. We therefore chose a normal distribution model to serve as an other-person matching score generation model.

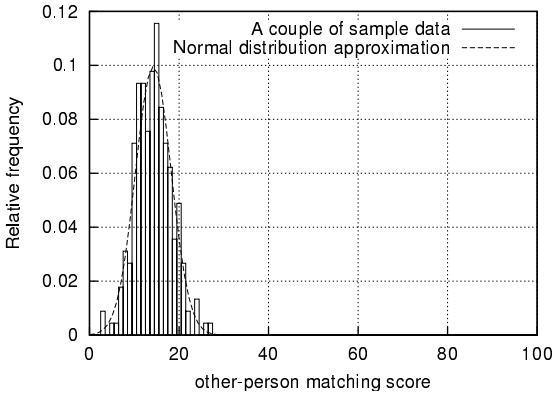


Fig. 2. Example of other-person score distribution for fingerprint

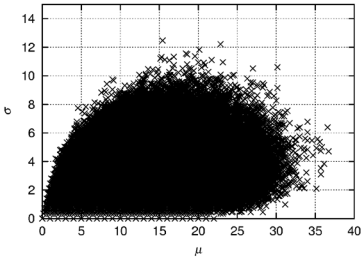
The other-person matching score generation characteristics are considered to vary from user to user. Therefore, for other-person pairs user i and user j, the characteristic changes with the combination (i, j) and we assume a model for generating other-person matching scores that follow a normal distribution with mean μ_{ij} and standard deviation σ_{ij}.

3.2 Parameter Model of the Matching Score Generation Model

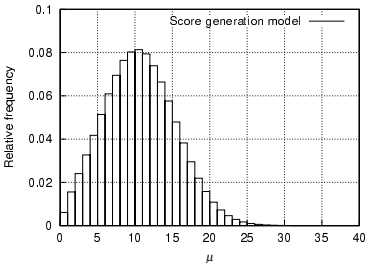
Because the purpose of the matching score generation model as a hypothesis biometrics that we are concerned with here is to understand the relation between the model parameters and response performance, we simply assume that μ_{ij} and σ_{ij} can be approximated by normal distributions.

Here shows an example of a parameter model generated by normal distributions in Fig.3. Fig.3(a) shows the relation of μ_{ij} and σ_{ij} of the other-person matching score, and Fig.3(b) and Fig.3(c) respectively show the frequency distributions of μ_{ij} and σ_{ij}. This example is used as a reference parameter of simulation experiments in the next section.

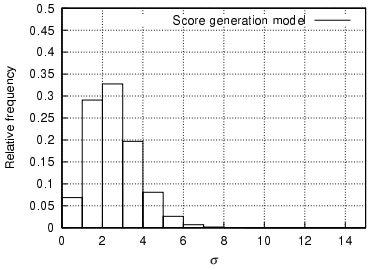
Using this matching score generation model, we ran an identification simulation. The distribution of the number of search steps is shown in Fig.4. Even though this is a hypothesis biometrics that employs a very simplified matching score generation model using normal distributions, the obtained distribution (Fig.4) has a peak at a low number of searches in the same way as for Fig.1, and this is characteristic of the proposed algorithm. What this means is that fast response can be achieved even for a hypothesis biometrics.



(a) Relation of μ_{ij} and σ_{ij}



(b) μ_{ij} frequency distribution



(c) σ_{ij} frequency distribution

Fig. 3. μ_{ij} and σ_{ij} for the matching score generation model

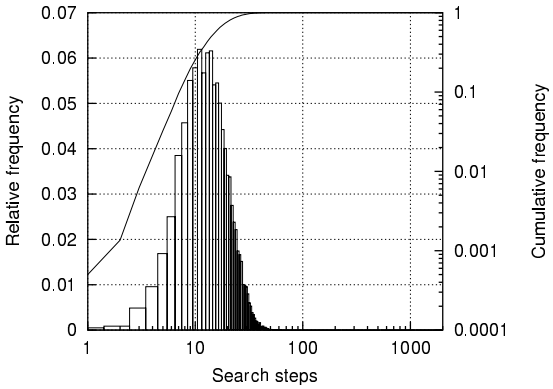
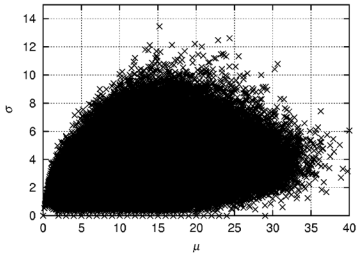


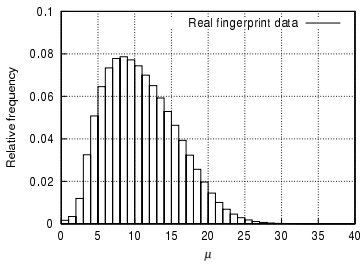
Fig. 4. Search steps distribution for the matching score generation model

Note that there are, for example, differences in the values obtained with the actual data (Fig.5) and the model (Fig.3). It is because that the matching score generation model employs a simple normal distribution, but the actual data is calculated in more complicated procedure.

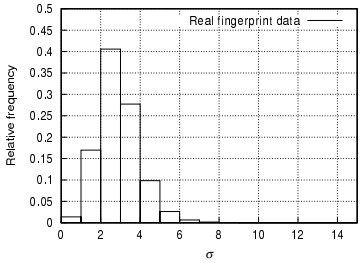
The parameter model illustrated above can be determined however we like, but these parameters were obtained by referring the actual fingerprint data in Fig.5. It is our future work to make a model that completely reproduces the actual data.



(a) Relation of μ_{ij} and σ_{ij}



(b) μ_{ij} frequency distribution



(c) σ_{ij} frequency distribution

Fig. 5. μ_{ij} and σ_{ij} for fingerprint

For reference, the procedure to obtain the reference parameters described above is shown here.

1. In Fig.5(a), there is a tendency for the distribution to skew to the right, and we obtain the main component direction of this two-dimensional distribution and the angle (θ) with respect to the μ axis.
2. Rotate $\boldsymbol{p}(i,j) = (\mu_{ij}, \sigma_{ij})$ by $-\theta$ according to Eq.(4) to obtain $\tilde{\boldsymbol{p}}(i,j) = (\tilde{\mu}_{ij}, \tilde{\sigma}_{ij})$. In Here, $\boldsymbol{T}(-\theta)$ is a rotated matrix that rotates $\boldsymbol{p}(i,j)$ by $-\theta$.

$$\tilde{\boldsymbol{p}}^t(i,j) = \boldsymbol{T}(-\theta)\boldsymbol{p}^t(i,j) \tag{4}$$

3. Obtain the mean (the mean of $\tilde{\mu}_{ij}$) and the standard deviation (the standard deviation of $\tilde{\mu}_{ij}$) of the orthogonal projection of $\tilde{\boldsymbol{p}}(i,j)$ on the μ axis. Similarly, obtain the mean (the mean of $\tilde{\sigma}_{ij}$) and the standard deviation (the standard deviation of $\tilde{\sigma}_{ij}$) of the orthogonal projection of $\tilde{\boldsymbol{p}}(i,j)$ on the σ axis.
4. For each other-person data pair, (i,j), obtain $\tilde{\mu}_{ij}$ from the mean of $\tilde{\mu}$ and the standard deviation of $\tilde{\mu}$ and obtain $\tilde{\sigma}_{ij}$ from the mean of $\tilde{\sigma}$ and the standard deviation of $\tilde{\sigma}$ and obtain $\tilde{\boldsymbol{q}}(i,j)$ that occurs in their respective normal distributions.
5. Rotate the obtained $\tilde{\boldsymbol{q}}(i,j)$ by $\boldsymbol{T}(\theta)$ and then obtain $\boldsymbol{q}(i,j)$ in Eq.(5).

$$\boldsymbol{q}^t(i,j) = \boldsymbol{T}(\theta)\tilde{\boldsymbol{q}}^t(i,j) \tag{5}$$

6. Take $\boldsymbol{q}(i,j)$ the obtained by the above steps as the μ_{ij} and σ_{ij} of the model for generating the matching scores.

3.3 Effect of Matching Score Characteristics on Response Performance

To investigate what effects changes in the characteristics of other-person matching score has on response performance, we ran the identification simulation for various values of the μ_{ij}, σ_{ij} occurrence parameters of the matching score generation model described in previous section. Specifically, taking the matching score generation parameters used in section 3.2 as reference values, the mean of μ_{ij} and the standard deviation of μ_{ij}, and the mean of σ_{ij} and the standard deviation of σ_{ij} were varied according to Table 1. When one parameter was varied, the other was fixed at the reference value.

Table 1. Varying of the model parameters

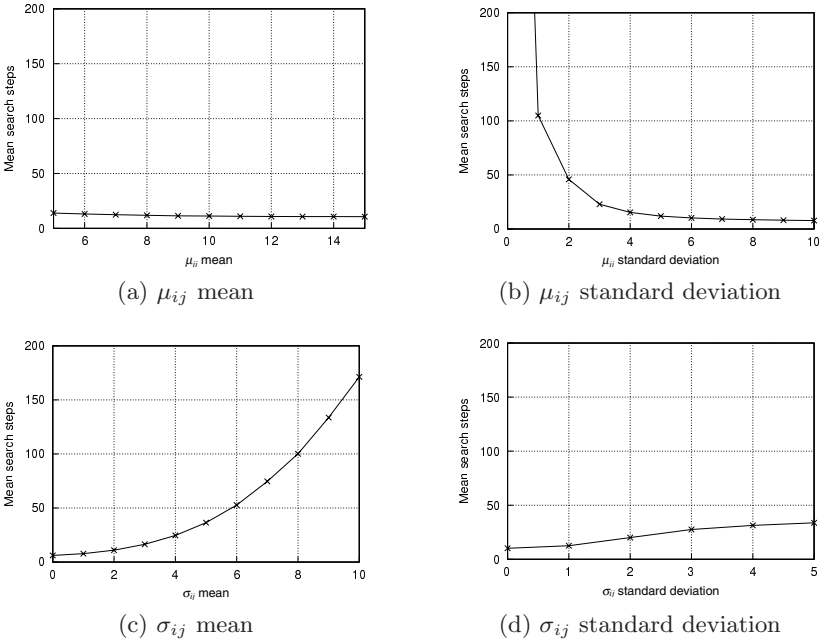
Parameters	Variation Range	Reference Value
μ_{ij} mean	5 to 15	10.87
μ_{ij} standard deviation	0 to 10	4.85
σ_{ij} mean	0 to 10	2.89
σ_{ij} standard deviation	0 to 5	1.03

The results for when the μ_{ij} mean, μ_{ij} standard deviation, σ_{ij} mean, and σ_{ij} standard deviation are varied are presented in Fig.6. In these figures, the mean number of search steps is used as an index of response performance. From Fig.6(a) we can see that the μ_{ij} mean does not greatly affect the response performance. The μ_{ij} mean is an element for shifting the entire distribution in parallel, and can be considered to have no effect on response performance.

In Fig.6(b), we can see that the response performance is getting worse as the μ_{ij} standard deviation becomes small. This is believed to be a result of the other-person matching score taking same value for any pair of data sets, thus not contributing to the classification of other-person. Conversely, the larger the μ_{ij} standard deviation, the better the response performance. It is because that the individuality emerges according to various values of μ_{ij}. Therefore, if the other-person matching score takes on diverse values reflecting that individuality, then the classification as other-person can be arrived at easily, meaning that the person's data will quickly become a match candidate.

In addition, we believe that smaller σ_{ij} mean values are associated with better response performance (Fig.6(c)) because the other-person matching score becomes stable as the result of the same values always appearing for certain other-person combinations. Conversely, larger σ_{ij} mean value bring lower response performance. This means that, if the matching score becomes disperse and unstable, it is difficult for the person's data to become a match candidate.

Smaller values of the σ_{ij} standard deviation (Fig.6(d)) are associated with better response performance. It is believed to have somewhat better response performance if few cases include large variance.



(a) μ_{ij} mean (b) μ_{ij} standard deviation

(c) σ_{ij} mean (d) σ_{ij} standard deviation

Fig. 6. Relation of parameter and response performance for the matching score generation model

Thus, these identification simulations using a matching score generation model have clarified the following points. The parameters that have large effects on the response performance are the μ_{ij} standard deviation and the σ_{ij} mean. When the variance in the mean value of the other-person matching score distribution is large and there are various values for each user, response performance is better. Also, when the standard deviation of the other-person matching score distribution is small, and a stable other-person matching score for each other-person is generated, response performance is better.

3.4 A Simulation Result Using an Example of Another Comparison Computation Function

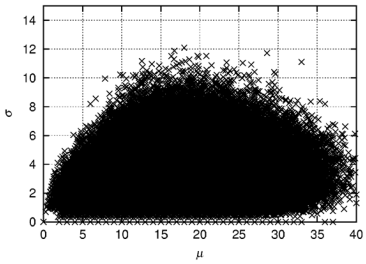
Here, we present the simulation results for another comparison computation function that has different other-person matching score distribution characteristics. This is used to verify the relationship of the comparison computation function and response performance described in previous section.

According to the guidelines in previous section, we introduce an example of another comparison computation function. Here, we define $y_{u,c}$ as the maximum value among the matching scores for the input data u and one set of enrollment data of the current match candidate c: $y_{u,c_1}, y_{u,c_2}, y_{u,c_3}, \ldots, y_{u,c_l}$. It is often used in biometric authentication systems. If we will use Eq.(6) as another comparison computation function, response performance can be better because $max()$

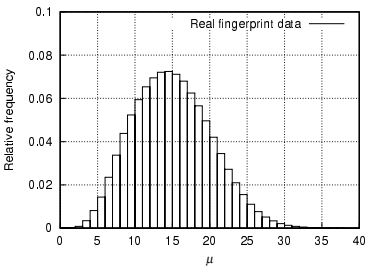
function is expected to make μ_{ij} standard deviation be larger and σ_{ij} mean be smaller.

$$y_{u,c} = max(y_{u,c_1}, y_{u,c_2}, y_{u,c_3}, \ldots, y_{u,c_l}) \tag{6}$$

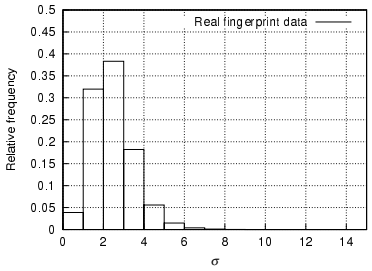
For when Eq.(6) is used as the comparison computation function, the relation of other-person matching score μ_{ij} and σ_{ij}, and the frequency distributions of μ_{ij} and σ_{ij} are respectively shown in Fig.7. The distribution of the number of search steps for when Eq.(6) is used as the comparison computation function is shown in Fig.8.



(a) Relation of μ_{ij} and σ_{ij}



(b) μ_{ij} frequency distribution



(c) σ_{ij} frequency distribution

Fig. 7. μ_{ij} and σ_{ij} for fingerprint with new comparison computation function

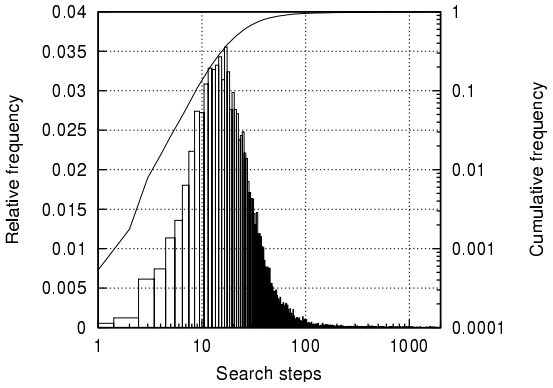


Fig. 8. Search steps distribution for fingerprint with new comparison function

Table 2. Model parameters and response performance

Parameters	Fig.1	Fig.8
μ_{ij} mean	10.87	14.90
μ_{ij} standard deviation	4.85	5.19
σ_{ij} mean	2.89	2.46
σ_{ij} standard deviation	1.03	1.02
Mean search steps	83.70	40.96

The contents of Fig.1 and Fig.8 are compared in Table 2. We see in the table that the mean number of search steps is smaller in Fig.8, where the new comparison computation function is used, indicating good response performance. This result is explained by the fact that μ_{ij} standard deviation and σ_{ij} mean were changed in the direction of better response performance as we expected in the new comparison computation function.

In this section, we introduced an example of another comparison computation function, however, it is necessary to revise the definition of the matching score for more response performance improvement.

4 Conclusion

We clarified the relation between the characteristics of a matching score distribution and response performance using a matching score generation model, and described design guidelines for a comparison computation function to improve response performance in the identification operation. And also, simulations using an example of another comparison computation function demonstrated the effectiveness of these guidelines.

The matching scores calclated in the identification operation up to now have employed a comparison function that is optimized for one-to-one comparison. However, if the comparison computation function for the identification is designed according to this paper, it should be possible to achieve both high speed and highly accurate individual authentication.

References

1. T.Maeda, M.Matsushita, and K.Sasakawa, “Identification algorithm using a matching score matrix,” IEICE Transactions on Information and Systems, vol.E84-D, no.7, pp.819-824, 2001
2. Dario Maio, and Davide Maltoni, “A structural approach to fingerprint classification”, Proc. of ICPR’96, pp.578-585, 1996
3. A.K. Jain, S. Prabhakar, and L. Hong, “A multichannel approach to fingerprint classification”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol.21, no. 4, pp.348-359, 1999
4. Yuan Yao, G.L. Marcialis, M. Pontil, P. Frasconi, and F. Roli, “A new machine learning approach to fingerprint classification,” Proc. of the 7th Congress of the Italian Association for Artificial Intelligence on Advances in Artificial Intelligence, pp.57-63, 2001

5. Sen Wang, Wei Wei Zhang, and Yang Sheng Wang, "Fingerprint classification by directional fields," Proc. of the IEEE 4th International Conference on Multimodal Interface, pp.395-398, 2002
6. A.Lumini, D.Maio, and D.Maltoni, "Continuous versus exclusive classification for fingerprint retrieval," Pattern Recognition Letters, vol.18, no.10, pp.1027-1034, 1997
7. R.S. Germain, A. Califano, and S. Colville, "Fingerprint matching using transformation parameter clustering," IEEE Computational Science and Engineering, vol.4, no.4, pp.42-49, 1997
8. A. Califano, B. Germain, and S. Colville, "A high-dimensional indexing scheme for scalable fingerprint-based identification," Proc. of the Third Asian Conference on Computer Vision, vol.1, pp.32-39, 1998
9. T.Maeda, M.Matsushita, and K.Sasakawa, "Quantitative Performance Analysis of Biometrics Identification for Authentication Systems Design," WSEAS Transactions on Computer, issue 5, vol.3, pp.1281-1289, 2004

Pose Invariant Face Recognition Under Arbitrary Illumination Based on 3D Face Reconstruction

Xiujuan Chai[1], Laiyun Qing[2], Shiguang Shan[2], Xilin Chen[1,2], and Wen Gao[1,2]

[1] School of Computer Science and Technology, Harbin Institute of Technology, 150001 Harbin, China
{xjchai,xlchen,wgao}@jdl.ac.cn
[2] ICT-ISVISION Joint R&D Lab for Face Recognition, ICT, CAS, 100080 Beijing, China
{lyqing,sgshan}@jdl.ac.cn

Abstract. Pose and illumination changes from picture to picture are two main barriers toward full automatic face recognition. In this paper, a novel method to handle both pose and lighting condition simultaneously is proposed, which calibrates the pose and lighting condition to a pre-set reference condition through an illumination invariant 3D face reconstruction. First, some located facial landmarks and a priori statistical deformable 3D model are used to recover an elaborate 3D shape. Based on the recovered 3D shape, the "texture image" calibrated to a standard illumination is generated by spherical harmonics ratio image and finally the illumination independent 3D face is reconstructed completely. The proposed method combines the strength of statistical deformable model to describe the shape information and the compact representations of the illumination in spherical frequency space, and handle both the pose and illumination variation simultaneously. This algorithm can be used to synthesize virtual views of a given face image and enhance the performance of face recognition. The experimental results on CMU PIE database show that this method can significantly improve the accuracy of the existed face recognition method when pose and illumination are inconsistent between gallery and probe sets.

1 Introduction

The face recognition problem has been studied for more than three decades. Currently, the accuracy of face recognition for frontal face under uniform lighting condition is pretty high [12]. However, in some more complicated cases, the recognition tasks suffer from the variations of poses and illuminations.

The appearance of faces may looks quite different when pose or illumination change, and this issues an imperfect task for face recognition when only the 2-D appearance-based method is applied. Although some 2-D-based methods are proposed to tackle pose or illumination variation problem, we believe that 3-D-based method is the final killer of both pose and illumination blending problem.

In the early years, using the low dimensional representation is the mainstream to tackle both pose and illumination problem in face recognition. Eigenfaces [11] and Fisherfaces [2] apply statistical learning to get the empirical low dimensional pose or illumination space of the faces. These methods have demonstrated their easy implementation and accuracy, but the performance decreased dramatically when the imaging condition is dissimilar to those of the training images. The Fisher light-fields algorithm [7] proposed by Gross etc tackled the pose and illumination problem by

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 956–965, 2005.
© Springer-Verlag Berlin Heidelberg 2005

estimating the eigen light-fields of the subject's head from the gallery or probe images, which was used as the set of features to do recognition finally. Extended this work, Kevin Zhou presented an illuminating light field algorithm [14], in which a Lambertian reflectance model was used to handle the illumination variation. This leads to a more powerful generalization to novel illuminations than the Fisher light field. However, lots of images under multi-poses and multi-lights are needed for the training of this algorithm.

Since the pose and illumination variations are all related to the 3D face structure, the pose and illumination invariant face recognition can be easily achieved once the 3D face is known. Some model-based approaches were proposed to treat the extrinsic parameters as separate variables and model their functional role explicitly. These methods commonly build an explicit generative model of the variations of the face images, to recover the intrinsic features of the face: shape and/or albedo. Georghiades proposed the Illumination Cone [6] to solve face recognition under varying lightings and poses. Sampling across pose changing, the corresponding illumination cone is approximated by a low-dimensional linear subspace whose basis vectors are estimated using generative model. This method needs at least seven images under different lighting condition for each subject, which is impractical for the most of the applications. Zhao introduce the symmetric constraint to shape from shading for 3D face reconstruction and proposed the SSFS (Symmetric Shape from Shading) method [13]. The most successful face recognition system across pose and lighting is the 3D morphable model [3]. In this method, the shape and the texture of a face are expressed as the barycentric coordinates as a linear combination of the shapes and textures of the exemplar faces respectively. The 3D faces can be generated automatically from one or more photographs by optimizing the shape parameters, the texture parameters and the mapping parameters. This morphable method has been used in FRVT 2002 for its good performance [12]. However, the iterative optimization cost too much computational power and the fitting processing takes about 4.5 minutes on a workstation with a 2Ghz P4 processor.

Inspired by the work of the 3D morphable model [3], we also take the 3D statistical deformable model to represent the 3D shape space of human face. But differ from it, only 2D shape vector of the given facial image and a sparse version of the 3D deformable model are used to get the optimal shape coefficients, which can recover the whole elaborate 3D shape. The face region is extracted directly from the input image. Based on the recovered 3D shape, we approximate the "texture image" by relighting the face region to the standard illumination. Then the illumination independent 3D face is recovered completely only from single face image under any pose with arbitrary illumination. This strategy is based on the assumption that the pose is relevant to the relative locations of some key feature points and independent to the intensity of the image. Then the complicated optimal procedure is avoided by separating the shape and texture. The finally match is performed between the pose and illumination normalized facial images and the gallery images which have also been done the same normalization.

The remaining parts of the paper are organized as follows: In Section 2, how to realize the pose and illumination invariant face recognition is described in detail, in which two parts are included. In subsection 2.1, the 3D shape reconstruction algorithm based on the sparse statistical deformable model is described. In subsection 2.2

the illumination independent "texture image" generation with spherical harmonic ratio image is presented. Some synthesized examples based on our algorithm and the experimental results of face recognition across pose and illumination are presented in Section 3, followed by short conclusion and discussion in the last section.

2 Face Recognition Across Pose and Illumination

The whole framework of pose and illumination calibration for face recognition is given in Fig. 1. First, the irises are located by a region growing searching algorithm [4] and the rude pose class is defined for labeling the sparse feature points in the given facial image. Then 3D shape is reconstructed based on a 2D geometry driven statistical deformable model. Recurring to the recovered 3D shape of the specific person, the illumination independent "texture image" is obtained by relighting the face region extracted from the given image with spherical harmonic ratio image strategy. The pose and lighting calibrated image are used as the input of face recognition and get the identity result. Our algorithm can be regarded as a pre-process step of any face recognition system.

In the following subsections, we will explain the two key issues of the proposed framework – the 3D shape reconstruction and the illumination independent "texture image" generation.

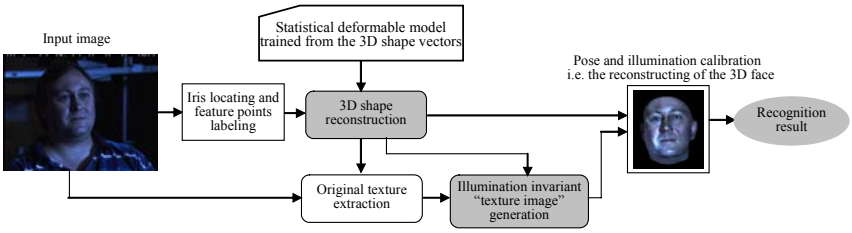


Fig. 1. The framework of pose and illumination calibration for face recognition

2.1 3D Shape Reconstruction from Single View

It is well known that the most direct solution to do pose normalization for a single non-frontal face image is to recover the 3D structure of the specific face. However, without any assumptions, recovering 3D shape from single image is a typical ill-posed problem. The minimal number of the images necessary to reconstruct the 3D face is three [8]. To overcome this, we use the prior knowledge of the 3D face class to describe the specific 3D shape of any novel face. A 3D face data set is used for training to get the statistical deformable model. This training set is formed by 100 laser-scanned 3D faces selected from the USF Human ID 3-D database [3]. All these faces are normalized to a standard orientation and position in space. The geometry of a face is represented by 75,972 vertices and down-sampled to 8,955 vertices in order to predigest computation. In the following paragraphs, the whole 3D facial shape reconstruction procedure will be explained in detail.

We represent the 3D geometry of a face with a shape-vector that is composed by concatenating the X , Y , and Z coordinates of the n vertices as:

$\mathbf{S} = (X_1, Y_1, Z_1, \quad, X_n, Y_n, Z_n)^T \in \Re^{3n}$. Supposing the number of the 3D face training collection is m, each face vector can then be written as $\mathbf{S_i}$, where $i = 1, \quad, m$. These 3D shape vectors have been full correspondence. Each novel 3D shape can be represented as the linear combination of the m exemplar faces shapes by: $\mathbf{S} = \sum_{i=1}^{m} w_i \mathbf{S_i}$. Because all face shapes are similar in holistic with some small differences, PCA (Principle Component Analysis) is appropriate for capturing the variance in terms of the principle components and filtering the noise among these shape vectors. Performing an eigen-decomposition to the matrix composed by these 3D shapes using PCA and we obtain $d \le (m-1)$ eigen shape vectors according to the descending order, which constitute the projection matrix $\mathbf{P}$. Therefore, the statistical deformable model is formed: $\mathbf{S} = \overline{\mathbf{S}} + \mathbf{P}\boldsymbol{\alpha}$, where $\overline{\mathbf{S}}$ is the mean shape and $\boldsymbol{\alpha}$ is the coefficient vector corresponding to the projection matrix $\mathbf{P}$, whose dimension is d.

Expanding this denotation, if the face takes some rotation variation, then the above formulation can be written as:

$$\mathbf{SR} = \overline{\mathbf{S}}\mathbf{R} + \mathbf{PR}\boldsymbol{\alpha}, \tag{1}$$

where $\mathbf{R}$ is the rotation matrix, relevant with the three rotation angles around the corresponding three coordinate axes. $\mathbf{SR}$ is the 3D face shape rotated around the 3D face coordinate center. We import a denotation $\mathbf{V}^{\mathbf{R}}$, which represents the operator performing a transformation to a 3D vector $\mathbf{V}$ by right multiplying a rotation matrix $\mathbf{R}$. Therefore, equation (1) can be rewritten as:

$$\mathbf{S}^{\mathbf{R}} = \overline{\mathbf{S}}^{\mathbf{R}} + \mathbf{P}^{\mathbf{R}}\boldsymbol{\alpha} \tag{2}$$

Similarly, the vector concatenating the coordinates of k landmarks ($(x_i, y_i), i = 1,2, \quad, k$) in 2D image is denoted as $\mathbf{S_I}$. Each 2D landmark corresponds to a fixed point in 3D shape vector with the coincident mapping relation. These corresponding 3D points constitute the sparse version of the 3D shape. The x and y coordinates of this sparse 3D shape concatenated to a 2D shape vector called $\mathbf{S_f}$, that is $\mathbf{S_f} = (x_1, y_1, \quad, x_k, y_k)^T \in \Re^{2k}$. Because the $\mathbf{S_f}$ can be regarded as the partial segment of the 3D shape $\mathbf{S}$, the following equation approximately holds: $\mathbf{S_f} = \overline{\mathbf{S}}_{\mathbf{f}} + \mathbf{P_f}\boldsymbol{\alpha}$. Here $\mathbf{V_f}$ is imported to denote the 2D shape vector comes from extracting the x and y from the 3D vector $\mathbf{V}$. So, $\overline{\mathbf{S}}_{\mathbf{f}}$ and $\mathbf{P}_f$ describe the corresponding parts to the 2D landmarks extracted from the 3D mean shape $\overline{\mathbf{S}}$ and projection matrix $\mathbf{P}$ respectively. $\mathbf{S}_{\mathbf{f}}^{\mathbf{R}}$, which denotes the sparse 2D shape vector extracted from the 3D face under $\mathbf{R}$ pose, can be represented inferentially by the following formula:

$$\mathbf{S}_f^{\mathbf{R}} = \overline{\mathbf{S}}_f^{\mathbf{R}} + \mathbf{P}_f^{\mathbf{R}}\boldsymbol{\alpha}. \tag{3}$$

Our aim is to reconstruct the whole 3D shape information with the coefficient vector $\boldsymbol{\alpha}$, which can be computed from the following equation:

$$\boldsymbol{\alpha} = (\mathbf{P}_f^{\mathbf{R}})^{+}(\mathbf{S}_f^{\mathbf{R}} - \overline{\mathbf{S}}_f^{\mathbf{R}}), \tag{4}$$

where $(\mathbf{P}_f^{\mathbf{R}})^+$ is the pseudo-inverse matrix, which can be computed by $(\mathbf{P}_f^{\mathbf{R}})^+ = \left((\mathbf{P}_f^{\mathbf{R}})^T(\mathbf{P}_f^{\mathbf{R}})\right)^{-1}(\mathbf{P}_f^{\mathbf{R}})^T$. So the crucial element is to compute the accurate $\mathbf{S}_f^{\mathbf{R}}$ of the specific person from the feature landmarks $\mathbf{S}_I$. The relation between $\mathbf{S}_I$ and $\mathbf{S}_f^{\mathbf{R}}$ can be represented by:

$$\mathbf{S}_I = \left(\mathbf{S}_f^{\mathbf{R}} + \mathbf{T}\right)c \,. \tag{5}$$

For the rotation matrix $\mathbf{R}$, we define it with the three rotation angles of the corresponding coordinate axis. Making use of the 5 key landmarks in a face image and the corresponding 3D facial points of its 3D shape model $\mathbf{S}$, the three rotation angle parameters can be inferred by projection computation. The 5 landmarks used to compute the rotation matrix $\mathbf{R}$ are the left and right iris, the nose tip, the left and right mouth corner respectively.

In the following, we will describe the iterative algorithm to compute the optimal shape coefficients vector $\boldsymbol{\alpha}$. In the first iteration, we set the $\overline{\mathbf{S}}_f$ to be the initial value of $\mathbf{S}_f$, and set $\overline{\mathbf{S}}$ to be initial 3D shape $\mathbf{S}$ of a specific person to get the initial values of the pose parameters. The iterative optimization procedure is given below:

(a) Compute the rotation matrix $\mathbf{R}$ by erecting equation group according to 5 points projection computation.

(b) Then the translation $\mathbf{T}$ and scale factor c for the landmarks between $\mathbf{S}_I$ and $\mathbf{S}_f^{\mathbf{R}}$ are calculated based on the computation between these two 2D shape vectors.

(c) Refine $\mathbf{S}_f^{\mathbf{R}}$ through equation (5) using the $\mathbf{T}$ and c gained above.

(d) Having the new $\mathbf{S}_f^{\mathbf{R}}$, we can get the coefficient vector $\boldsymbol{\alpha}$ easily by equation (4).

(e) Reconstruct the 3D face shape $\mathbf{S}$ for the specific person by equation $\mathbf{S} = \overline{\mathbf{S}} + \mathbf{P}\boldsymbol{\alpha}$

(f) Repeat the step (a) to (e) until the coefficients vector converges or a limit on the iteration times is reached.

Finally, we get the optimal 3D shape for the given face. Then we transform the result 3D shape by multiplying the pose matrix $\mathbf{R}$ to get the rotated 3D shape, which has the same pose to the input face. To get more elaborate 3D shape solution, we regulate the x and y coordinates of the vertices in 3D face according to the corresponding landmarks in the given 2D image.

2.2 Illumination Independent "Texture Image" Generation with Spherical Harmonic Ratio Image

With the recovered 3D shape and the pose parameters in subsection 2.1, the face region is extracted from the given image. However, this face region image is influenced by the illumination condition. For the difficulty to get the really intrinsic texture, we transformed to compute the calibrated "texture image" under some standard illumination. Finally, this calibrated "texture image" can be mapped to the 3D shape and the illumination independent 3D face is reconstructed completely.

Since the reflection equation can be viewed as a convolution, it is natural to analyze it in frequency-space domain. With spherical harmonics, Basri et al [1] proved

that most energy of the irradiance was constrained in the three low order frequency components and got its frequency formula as

$$\begin{aligned} E(\alpha,\beta) &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l} E_{lm}Y_{lm}(\alpha,\beta), \\ &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l} A_l L_{lm}Y_{lm}(\alpha,\beta), \\ &\approx \sum_{l=0}^{2}\sum_{m=-l}^{l} A_l L_{lm}Y_{lm}(\alpha,\beta), \end{aligned} \tag{6}$$

where $A_l (A_0 = \pi, A_1 = 2\pi/3, A_2 = \pi/4)$ [1] are the spherical harmonic coefficients of Lambertian reflectance, L_{lm} are the coefficients of the incident light, and Y_{lm} are the spherical harmonic functions. The spherical harmonics has already been used in face recognition across illumination, such as [15].

Given a face region image I, for each pixel (x, y), this equation always holds up: $I(x,y) = \rho(x,y)E\big(\alpha(x,y),\beta(x,y)\big)$. Here, the $\alpha(x,y)$ and $\beta(x,y)$ can be gotten from the normal vector of the 3D face shape. We also assume the albedo ρ is a constant. Let $\mathbf{E_{lm}} = A_l \mathbf{Y}_{lm}$ denote the harmonic irradiance image and $\mathbf{E}$ is a $n\times 9$ matrix of $\mathbf{E_{lm}}$, where n is the pixel number of the texture image. Then the coefficients of the illumination $\mathbf{L}$ can be gotten by solving the least squares problem:

$$\hat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left\| \mathbf{E}(\rho\mathbf{L}) - \mathbf{I} \right\|, \tag{7}$$

Once we have estimated the lighting condition of the given image, relighting it to a standard illumination is straightforward [16].

For any given point P at position (x, y) on the image, whose normal is (α, β), and albedo is $\rho(x, y)$, then the intensities at P in the original image and the canonical image are respectively:

$$\begin{aligned} I_{org}(x,y) &= \rho(x,y)\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l \hat{L}_{lm}Y_{lm}(\alpha,\beta), \\ I_{can}(x,y) &= \rho(x,y)\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l L_{lm}^{can}Y_{lm}(\alpha,\beta), \end{aligned} \tag{8}$$

where (x, y) ranges over the whole image.

The ratio image of the two different illuminations is defined as:

$$R(x,y) = \frac{I_{can}(x,y)}{I_{org}(x,y)} = \frac{\rho(x,y)\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l L_{lm}^{can}Y_{lm}(\alpha,\beta)}{\rho(x,y)\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l \hat{L}_{lm}Y_{lm}(\alpha,\beta)} = \frac{\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l L_{lm}^{can}Y_{lm}(\alpha,\beta)}{\sum_{l=0}^{2}\sum_{m=-l}^{l} A_l \hat{L}_{lm}Y_{lm}(\alpha,\beta)}. \tag{9}$$

Therefore, with the original image and the ratio image, the illumination canonical image is:

$$I_{can}(x,y) = R(x,y)\times I_{org}(x,y). \tag{10}$$

After the elaborated 3D shape and illumination calibrated "texture image" are recovered, we can reconstruct the whole 3D face of the specific person. For the invisible points in the texture, the interpolation strategy is exploited. And the pose normalization can be achieved by rotating the 3D face model to any predefined standard pose.

3 Experiments and Results

In this section, we evaluate the performance of the proposed algorithm through pose and illumination invariant face recognition. For a given non-frontal image under arbitrary illumination, we reconstruct its illumination independent 3D face. Pose normalization is achieved by rotating the 3D face to a predefined (frontal) pose. Then the calibrated face image is used as the input of the general face recognition system to perform recognition.

3.1 Experimental Results for Face Recognition Across Pose Only

First, the experiment on face recognition across pose only is carried out on 4 pose subsets of CMU PIE database [10], which are pose set 05 (turn right 22.5 degree), pose set 29 (turn left 22.5 degree), pose set 37 (turn right 45 degree) and 11 (turn left 45 degree) respectively, and the gallery images are from the pose set 27, which are all frontal images. Our face recognition method is Gabor PCA plus LDA, whose idea is similar to the GFC [9]. The training images are selected from the CAS-PEAL Database [5], totally 300 persons, and each person has 6 pose images, 10 frontal images averagely. In our experiment the feature points are labeled manually. Some pose normalization results based on the 3D face reconstruction are presented in Fig. 2 to give a visualize evaluation. The recognition results are listed in Fig. 3, which has intensively shows the good performance of the pose normalization based on our 3D face reconstruction. The recognition match scores for the 4 pose sets are improved significantly compared with the original recognitions, and the rank-1 recognition rate reaches to 94.85% averagely after pose normalization.

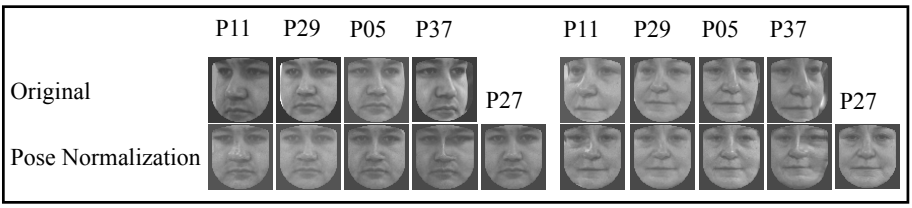


Fig. 2. The pose normalized images. The first row is the original masked images. The second row is the corresponding pose normalized images, and right to which are the gallery images in 27 to be references

3.2 Experimental Results for Face Recognition Across Pose and Illumination

We verify the simultaneous effect of the pose and illumination normalization in this section. In our experiments, we used the "illum" subsets of the CMU PIE database, which provides the facial images under well-controlled poses and lightings. We take the experiment on 2856 images from the 2 pose subsets, 05 and 29, each subset including 21 different kinds of illuminations and the flash numbers are 02-21. The frontal pose set 27 under flash "11" is taken as the gallery, and the other probe images are all aligned to the frontal pose and the standard light as flash number "11". Some examples of the pose and illumination normalized images are given in Fig.4. The nor-

malized images in (b) are more similar to the gallery image as (c) in vision than the original images shown in (a). The experimental results of face recognition with correlation matching strategy across pose and lighting are listed in Table 1.

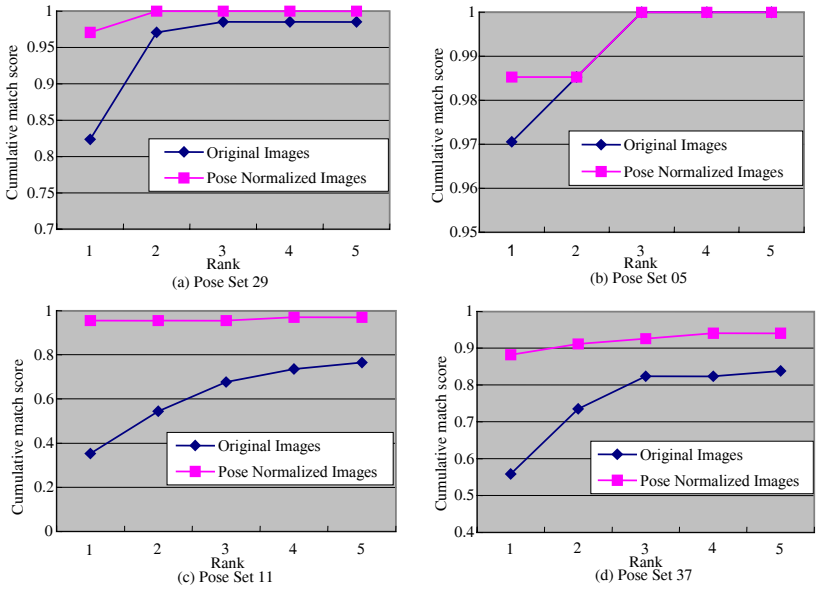


Fig. 3. The recognition results on the original and the pose normalized images in the 4 different pose sets of CMU PIE database with Gabor PCA plus LDA recognition strategy



Fig. 4. The pose and illumination calibrated results. (a) the original images. (b) the corresponding pose and lighting calibrated results. (c) the gallery image

4 Conclusion

In this paper a novel illumination independent 3D face reconstruction is proposed to recognize facial images across pose and illumination. The 3D shape is recovered from single non-frontal facial image based on a statistical deformable model regressed through 2D geometry formed by some facial landmarks. Recurring to the reconstructed 3D shape, the illumination independent facial "texture image" is achieved with spherical harmonic ratio image. The experimental results show that the pose and illumination calibrating strategy largely improves the performance of the general face recognition for the probe images under uncontrolled pose and lighting.

Table 1. Recognition results on 2 pose subsets under 21 different lightings in CMU PIE Database with the correlation matching strategy

F\C	Pose 05 (original)	Pose 05 (calibrated)	Increase	Pose 29 (original)	Pose 29 (calibrated)	Increase
02	0.044	0.206	0.162	0.015	0.235	0.230
03	0.059	0.412	0.353	0.029	0.324	0.295
04	0.103	0.735	0.632	0.059	0.612	0.553
05	0.397	0.897	0.500	0.103	0.882	0.779
06	0.735	0.882	0.147	0.162	0.926	0.764
07	0.676	0.912	0.236	0.118	0.912	0.794
08	0.544	0.897	0.353	0.588	0.956	0.368
09	0.235	0.897	0.662	0.676	0.985	0.309
10	0.324	0.912	0.588	0.088	0.838	0.750
11	0.676	0.912	0.236	0.838	0.971	0.133
12	0.309	0.926	0.617	0.765	0.941	0.176
13	0.074	0.868	0.794	0.221	0.882	0.661
14	0.088	0.897	0.809	0.235	0.912	0.677
15	0.029	0.750	0.721	0.059	0.750	0.691
16	0.029	0.368	0.339	0.044	0.471	0.427
17	0.015	0.221	0.206	0.029	0.279	0.250
18	0.250	0.838	0.588	0.074	0.750	0.676
19	0.647	0.912	0.265	0.118	0.926	0.808
20	0.662	0.912	0.250	0.838	0.971	0.133
21	0.265	0.926	0.661	0.706	0.941	0.235
22	0.074	0.838	0.764	0.118	0.824	0.706
Average	0.296	0.768	0.472	0.280	0.776	0.496

Accurate alignment would facilitate the 3D shape recovery and the subsequent recognition. Therefore, one of our future efforts will be the accurate alignment, especially under the non-ideal lighting environment.

Acknowledgements

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, "100 Talents Program" of CAS, ShangHai Municipal Sciences and Technology Committee (No.03DZ15013), and ISVISION Technologies Co., Ltd.

References

1. R. Basria and D. Jacobs, "Lambertian Reflectance and Linear Subspaces", Proc. ICCV' 2001, pp. 383-390.
2. P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs Fisherfaces: recognition using class specific linear projection". IEEE Trans. on PAMI, 1997.7, vol.20, No.7.
3. V. Blanz and T. Vetter, "Face Recognition based on Fitting a 3D Morphable Model", IEEE Transactions on PAMI 2003, vol. 25, pp. 1063-1074.
4. B.Cao, S.Shan and W.Gao, "Localizing the Iris Center by Region Growing Search", Proceeding of the ICME2002. 2002, vol. 2, pp. 129-132.

5. W. Gao, B. Cao, S.G. Shan, D.L. Zhou, X.H. Zhang and D.B. Zhao, “The CAS-PEAL Large-Scale Chinese Face DataBase and Evaluation Protocols”, Technique Report No. JDL-TR_04_FR_001, Joint Research & Development Laboratory, CAS, 2004.
6. A.S. Georghiades, P.N. Belhumeur and D.J. Keiegman, “From Few to Many: Illuminition Cone Models for Face Recognition under Variable Lighting and Poses”, IEEE Transactions on PAMI, 2001, vol. 23, pp.643-660.
7. R. Gross, I. Matthews and S. Baker, “Eigen Light-Fields and Face Recognition Across Pose”, Proc. FG’02, 2002.
8. T.S. Huang and C.H. Lee, “Motion and Structure from Orthorgraphic Projections”, IEEE Transactions on PAMI, 1989, vol. 2, no. 5: pp. 536-540.
9. C. Liu and H. Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition", IEEE Trans. Image Processing, 2002, vol. 11, no. 4, pp. 467-476.
10. T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database", Proc. the 5th IEEE International Conference on Automatic Face and Gesture Recognition (FG'02), Washington, DC, May 2002.
11. M. Turk and A. Pentland, “Eigenfaces for Recognition” Journal of cognitive neuroscience, 1991, vol. 3, no. 1, pp. 71-86.
12. P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi and M. Bone, “Face Recognition Vendor Test 2002: Evaluation Report”, FRVT, 2002.
13. W. Zhao and R. Chellappa, “SFS Based View Synthesis for Robust Face Recognition”, In Proceedings of the International Conference on Automatic Face and Gesture Recognition, Grenoble, 2000, pp. 285-292.
14. K. Zhou and R. Chellappa, “Illuminating Light Field: Image-based Face Recognition across Illuminations and Poses”, Proc. FG’04, 2004.
15. L. Zhang, D. Samaras. “Face Recognition Under Variable Lighting using Harmonic Image Exemplars”, In Proc. CVPR 2003, pp. I: 19-25.
16. L. Y. Qing, S. G. Shan, W. Gao, “Face Recognition under Generic Illumination based on Harmonic Relighting”, International Journal of Pattern Recognition and Artificial Intelligence, 2005 (To Appear).

Discriminant Analysis Based on Kernelized Decision Boundary for Face Recognition

Baochang Zhang[1], Xilin Chen[1,2], and Wen Gao[1,2]

[1] Computer School, Harbin Institute of Technology, China
{Bczhang}@jdl.ac.cn
[2] ICT-ISVISION Joint R&D Lab for face recognition, ICT, CAS, China
{Bczhang,Xlchen,Wgao}@jdl.ac.cn

Abstract. A novel nonlinear discriminant analysis method, Kernelized Decision Boundary Analysis (KDBA), is proposed in our paper, whose Decision Boundary feature vectors are the normal vector of the optimal Decision Boundary in terms of the Structure Risk Minimization principle. We also use a simple method to prove a property of Support Vector Machine (SVM) algorithm, which is combined with the optimal Decision Boundary Feature matrix to make our method consistent with the Kernel Fisher method(KFD). Moreover, KDBA is easily used in its applications, and the traditional Decision Boundary Analysis implementations are computationally expensive and sensitive to the size of the problem. Text classification problem is first used to testify the effectiveness of KDBA. Then experiments on the large-scale face database, the CAS-PEAL database, have illustrated its excellent performance compared with some popular face recognition methods such as Eigenface, Fisherface, and KFD.

Keywords: Face Recognition, Kernel Fisher, Support Vector Machine

1 Introduction

Feature extraction has long been a hot topic in the pattern recognition field. Principle Component Analysis (PCA) and Fisher Linear Discriminant Analysis (FDA) are two classical techniques for the linear feature extraction. In many applications, both methods have been proven to be very powerful. PCA is designed to capture the variance in a dataset in terms of the principle components, and FDA is a well-known discriminant tool, which maximizes the so-called Fisher criterion $J(w) = tr[(wS_w w^T)^{-1}(wS_b w^T)]$, where S_w is the within-class scatter matrix, S_b is the between-class scatter matrix. Since the FDA is mainly based on a single class center, i.e.. the mean sample of the class, the feature vector calculated by which is not reliable if mean vectors can not reflect the distribution of the data set. Moreover, PCA and FDA are inadequate to describe the complex nonlinear variations in the training dataset. In recent years, the kernelized feature extraction methods have been paid much attention, such as Kernel Principal Component Analysis (KPCA)[1] and Kernel Fisher Discriminant analysis (KFD) [1,2,3], which are well-known nonlinear extensions to PCA and FDA respectively. However, the KFD cannot be easily used in real applications. The reason is that the projection directions of KFD often lie in the span of all the samples [4], therefore, the dimension of the feature often becomes very large, when the input space is mapped to a feature space through a kernel function. As a result, the scatter matrices become singular, which is the so-called "Small Sam-

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 966–976, 2005.
© Springer-Verlag Berlin Heidelberg 2005

ple Size problem" (SSS). Similar to [5], KFD simply adds a perturbation to within-class scatter matrix. Of course, it has the same stability problem as that in [5], because eigenvectors are sensitive to the small perturbation, moreover, the influence of which is not yet understood.

In this paper, we propose a new algorithm for feature extraction based on the proposed Kernelized Decision Boundary Analysis(KDBA). The algorithm tries to extract the necessary feature vectors to achieve the same classification accuracy as in the original space. The decision boundary theory was first proposed by Fukunaga et al [6, 7] and further developed by Lee [8] et al. However, the existing implementations of calculating the Decision Boundary Feature Matrix are computationally expensive and sensitive to the sample size. Moreover, the decision boundary method is a linear method, and fails to find the nonlinear structure from the training set. We proposed the KDBA method, which is a non-linear extension to the traditional one. Furthermore, the proposed method is easily implemented and strongly related to the theory of Structure Risk Minimization (SRM).

The rest of the paper is organized as following. In Section 2, we briefly introduce the Kernel Fisher analysis method. In Section 3, we define the Kernelized Decision Boundary Feature Matrix (KDBFM), which is based on the optimal Decision Boundary vector in terms of SRM. In section 4, we proposed the algorithm of KDBA, which is easily realized in its application. In section 5, we will give some experiment results on the Text classification problem and face recognition in the large-scale CAS-PEAL face database [14,18]. In the last section, we will make some conclusions about the proposed method.

2 Kernel Fisher Discriminant Analysis

We first describe, in this paper, the Kernel Fisher analysis method, which is a well-known extension to FDA. Moreover, many definitions will be used in our paper later, such as the kernel within-class and between-class scatter matrices. It is also the baseline algorithm in our paper, and we will make some comparative experiments with the proposed method.

The idea of Kernel FDA is to yield a nonlinear discriminant analysis in a higher dimensional space. The input data is first projected into an implicit feature space F by the nonlinear mapping $\Phi : x \in R^N -> f \in F$, and then seek to find a nonlinear transformation, which can maximize the between-class scatter and minimize the within-class scatter in F [1-4]. In its implementation, Φ is implicit and we will just compute the inner product of two vectors in F by using a kernel function:

$$k(x, y) = (\Phi(x) \cdot \Phi(y)). \tag{1}$$

We define between-class scatter matrix $\mathbf{S}_b$ and within-class scatter matrix $\mathbf{S}_w$ in F as following:

$$\mathbf{S}_b = \sum_{i=1}^{C} p(\varpi_i)(u_i - u)(u_i - u)^T \ , \tag{2}$$

$$\mathbf{S}_W = \sum_{i=1}^{C} p(\varpi_i) E\{((\Phi(x_i) - u_i)(\Phi(x_i) - u_i)^T) \mid \varpi_i\}, \tag{3}$$

$u_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\phi(x_{ij})$ denotes the sample mean of the class i, and u is the mean of all training images in F, $p(\varpi_i)$ is the prior probability. To perform FDA in a higher dimensional space F, it is equal to maximize Eq.4.

$$J(w) = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} = \frac{tr(\mathbf{S}_b)}{tr(\mathbf{S}_w)}. \tag{4}$$

Because any solution $\mathbf{w} \in F$ should lie in the span of all the samples in F [9,10], there exists:

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i\phi(x_i), \boldsymbol{\alpha}_i, i = 1,2...n.' \tag{5}$$

Then we will get the following Maximizing Criterion:

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T\mathbf{K}_b\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\mathbf{K}_w\boldsymbol{\alpha}}, \tag{6}$$

where $\mathbf{K}_w$ and $\mathbf{K}_b$ are defined as following:

$$\mathbf{K}_w = \sum_{i=1}^{C} p(\varpi_i)E(\eta_j - m_i)(\eta_j - m_i)^T, \tag{7}$$

$$\mathbf{K}_b = \sum_{i=1}^{C} p(\varpi_i)(m_i - \bar{m})(m_i - \bar{m})^T, \tag{8}$$

where $\eta_j = (k(x_1, x_j), k(x_2, x_j), ..., k(x_n, x_j))^T$,

$m_i = \left(\frac{1}{n_i}\sum_{j=1}^{n_i}k(x_1, x_j), \frac{1}{n_i}\sum_{j=1}^{n_i}k(x_2, x_j), .., \frac{1}{n_i}\sum_{j=1}^{n_i}k(x_n, x_j)\right)$, and $\bar{m}$ is the mean of all η_j.

Similar to FDA [5], this problem can be solved by finding the leading eigenvectors of $\mathbf{K}_w^{-1}\mathbf{K}_b$ used by Liu [9] and Baudat (GDA) [3], which is the Generalized Kernel Fisher Discriminant (GKFD) criterion. In our paper, we use the technique of the pseudo inverse of the within-class scatter matrix, and then perform PCA on $\mathbf{K}_w^{-1}\mathbf{K}_b$ to get the transformation matrix $\boldsymbol{\alpha}$. The projection of a data point x onto $\mathbf{w}$ in F is given by:

$$v = (\mathbf{w} \cdot \Phi(x)) = \sum_{i=1}^{n}\boldsymbol{\alpha}_i k(x_i, x). \tag{9}$$

3 Decision Boundary Feature Matrix

In this part, we will first briefly review some basic properties of the discirminant subspaces, for a detailed description and proofs, we can refer to[8].

Property 1. If a vector V is orthogonal to the vector normal to decision boundary at every point on decision boundary, V contains no information useful in discriminating classes, i.e., vectors discriminantly redundant. If a vector is normal to the decision boundary at least one point on the decision boundary, the vector contains information

useful in discriminating classes, i.e., the vector is discriminantly informative. And we can also refer to Fig.1.

Property 2. The Decision Boundary Feature Matrix (DBFM): let $N(x)$ be the unit vector normal to the decision boundary at a point x on the decision boundary for a given pattern classification problem, $p(x)$ is the data density. Then the DBFM is defined as following

$$\sum_{DBFM} = \int_S N(x)N^t(x)p(x)dx \,. \tag{10}$$

Property 3. The eigenvectors of the Decision Boundary Feature Matrix of a pattern recognition problem corresponding to non-zero eigenvectors are the necessary feature vectors to achieve the same classification accuracy as in the original space for the pattern recognition problem.

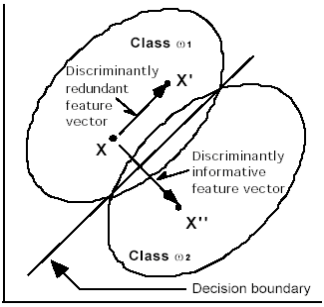


Fig. 1. $X^{'}$ is discriminantly redundant feature vector, $X^{''}$ is discriminantly informative feature vector

3.1 The Property of SVM

Support Vector Machines (SVM) is a state-of-the-art pattern recognition technique, whose foundations stem from the statistical learning theory. However, the scope of SVM is beyond pattern recognition because they can also handle another two learning problems, i.e., regression estimation and density estimation. SVM is a general algorithm based on guaranteed risk bounds of statistical learning, the so-called structural risk minimization principle. And we can refer to the tutorials [11] about the SVM. The success of SVM in face recognition [12, 13] as a recognizer provides us with further motivations to utilize SVM to enhance the performance of our system. However, we did not construct SVM classifier, and just used it to find the support vectors shown in Fig.2.

Here, we will use a simple way to prove the property of the support vectors based within-class scatter matrix for two-class problem, which shows that the SVM is strongly related to the Kernel Fisher analysis method.

In a higher dimensional space, x_1, x_2 are represented as $\Phi(x_1), \Phi(x_2)$. The SVM aims to optimize the following objective function [11]:

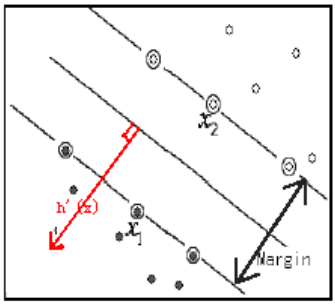


Fig. 2. Support Vectors are circled such as x_1, x_2

$$\min_{w} \frac{1}{2} w^T w \,, \tag{11}$$

$$\text{Subject to: } y_i(w^T \Phi(x_i) + b) - 1 \geq 0 \,. \tag{12}$$

The decision hyperplane function is linear style in a higher dimensional space as following:

$$h(\Phi(x)) = w^T \Phi(x_i) + b \,. \tag{13}$$

If $\Phi(x_i)$ is the support vector, we can know that [9]:

$$y_i(w^T \Phi(x_i) + b) - 1 = 0, y_i \in \{-1,1\} \,. \tag{14}$$

Thus, for $\Phi(x_1), \Phi(x_2)$ are the support vectors, we have:

$$\begin{aligned} &w^T \Phi(x_1) + b = 1, \Phi(x_1) \in S_1, \\ &S_1 = \{\Phi(x_i) \mid y_i = 1, w^T \Phi(x_i) + b = 1\} \end{aligned} \,, \tag{15}$$

$$\begin{aligned} &w^T \Phi(x_2) + b = -1, \Phi(x_2) \in S_2, \\ &S_2 = \{\Phi(x_i) \mid y_i = -1, w^T \Phi(x_i) + b = -1\} \end{aligned} \,. \tag{16}$$

The elements of S_1 and S_2 are support vectors, and it is easy for us to prove the following equations:

$$\begin{aligned} &w^T \overline{\Phi(x_1)} = 1 - b, \\ &\overline{\Phi(x_1)} = \frac{1}{n_1} \sum \Phi(x), \Phi(x) \in S_1 \end{aligned} \,, \tag{17}$$

$$\begin{aligned} &w^T \overline{\Phi(x_2)} = -1 - b, \\ &\overline{\Phi(x_2)} = \frac{1}{n_2} \sum \Phi(x), \Phi(x) \in S_2 \end{aligned} \,, \tag{18}$$

n_1 is the size of the S_1, n_2 is the size of the S_2. For two-class problem, the within-class scatter matrix is defined as following:

$$\begin{aligned} \mathbf{S}_w^{'} = &\frac{n_1}{n_2 + n_1} \sum_{i=1, \Phi(x_i) \in S_1}^{n_1} (\Phi(x_i) - \overline{\Phi(x_1)})(\Phi(x_i) - \overline{\Phi(x_1)})^T + \\ &\frac{n_2}{n_2 + n_1} \sum_{i=1, \Phi(x_i) \in S_2}^{n_2} (\Phi(x_i) - \overline{\Phi(x_2)})(\Phi(x_i) - \overline{\Phi(x_2)})^T \end{aligned} \,. \tag{19}$$

Therefore, we can know that:

$$\mathbf{w}^T \mathbf{S}_w^{'} \mathbf{w} = 0 \,, \tag{20}$$

where $\mathbf{S}_w^{'}$ is the within-class scatter matrix calculated by using the support vectors. $h_{\Phi(x)}^{'}$ is the normal vector of the decision hyperplane in a higher dimensional space, which is the optimal decision boundary in terms of the SRM. We called w the Kernelized Decision Bounday Feature vector calculated as following:

$$w = h_{\Phi(x)}^{'} = \sum_{i}^{n} \alpha_i y_i \Phi(x_i), \tag{21}$$

α_i is calculated in the SVM algorithm, and $\Phi(x_i)$ is the support vector.

3.2 Kernelized Decision Boundary Feature Matrix

In this part, we will define the KDBFM for the multi-class problem. The face Samples of S_i are from the class C_i, and the dataset $S_i^{'}$ includes samples from $C_l, l \neq i$. We will use SVM algorithm to find the Decision Boundary Feature vectors by using Eq.21, and first divide the dataset $S_i^{'}$ into k, $1 <= k <= C-1$ groups, represented by $S_{ij}^{'}, j = 1,...,k$. Now for any pair of data sets, such as $(S_i, S_{ij}^{'}), j = 1,...,k$, we use the two-class SVM algorithm to find the Decision Boundary Feature vector $\mathbf{w}_{ij}$ as in Eq.21. Therefore, we define the KDBFM as following:

$$\sum_{KDBFM} = \frac{1}{kC} \sum_{i=1}^{C} \sum_{j=1}^{k} w_{ij} w_{ij}^{T}. \tag{22}$$

From the Eq.20, we know that support vector has the excellent property, and we will redefine the within-class scatter matrix based on the support vector set calculated in SVM algorithm for a pair of data sets $(S_i, S_{ij}^{'})$ as following:

$$\mathbf{S}_w^{'}(i,j) = \sum_{i=1}^{c} p(\varpi_i) \sum_{m=1, C_m \in SV_i^j | C_m | > 1, x_m^{'} \in C_m}^{c} p(\varpi_m | \varpi_i) E((\Phi(x_m^{'}) - u_m^{'})(\Phi(x_m^{'}) - u_m^{'})^T). \tag{23}$$

In the case that only samples of C_i are included in the positive set, for two-class SVM algorithm, we will explain the parameters used in Eq.23 as following. $SV_i^j, j = 1,...,k$ Includes all the support vectors after performing SVM on the pair of data sets $(S_i, S_{ij}^{'})$, which is divided into two sub-sets, one of which is SV_{i1}^j, whose elements are the support vectors belonging to the class C_i, and the other is SV_{i2}^j including all other support vectors. $u_i^{'}$ denotes the sample mean of the set SV_{i1}^j, and $p(\varpi_i)$ is the prior probability. $u_{i0}^{'}$ denotes the sample mean of SV_{i2}^j, whose samples come from different classes. The number of classes in SV_{i2}^j is n_i, and the center of each class is represented as the mean vector $u_{ih}^{'}, h = 1,..,n_i$ (if only one sample

for one class is contained in SV_{i2}^{j}, then the sample is the class center). And now SV_i^j can be represented by a multi-center vector $(u_{\mathrm{i}}^{'}, u_{i0}^{'}, u_{i1}^{'}, ..., u_{in_{\mathrm{i}}}^{'})$, so we can know $u_m^{'} \in \{u_{\mathrm{i}}^{'}, u_{i0}^{'}, u_{i1}^{'}, ..., u_{in_{\mathrm{i}}}^{'}\}$. We will calculate the within-class scatter **sub-matrix** for the class, if more than one samples of which are contained in SV_i^j ($|C_m| > 1$). In our paper, the kernel function is polynomial style used in SVM and the proposed method, $k(x,y) = (\frac{x \cdot y}{|x| \cdot |y|} + 1)^r$, r is a constant integer.

Here, we will define the whole within-class scatter matrix based on all the support vector sets as following:

$$\mathbf{S}_w = \sum_{i=1}^{c} p(\varpi_i) \sum_{j=1}^{k} \frac{1}{k} \mathbf{S}_w^{'}(i,j) \quad . \tag{24}$$

To be concluded, in this part, we propose a method to construct KDBFM and the new kernelized within-class scatter matrix based on the support vector set, which is represented by a multi-center vector, $(u_{\mathrm{i}}^{'}, u_{j0}^{'}, u_{j1}^{'}, ..., u_{jn_{\mathrm{i}}}^{'})$.

4 Kernelized Decision Boundary Analysis

From the above discussion, we know that the eigenvectors of $\sum_{KDBFM}$ corresponding to non-zero eigenvalues are the necessary feature vectors. Moreover, we also hope that Decision Boundary Feature vector satisfies the Eq.20. Now, we show that our method is closely related to the Kernel Fisher method. If the $\sum_{KDBFM}$ is thought of as the between-class scatter matrix, which means that the trace of which is maximized by choosing the nonzero eigenvalues, and the proposed method will be consistent with the Fisher criterion, since the Eq.20 means minimizing the trace of the within-class scatter matrix. In the kernelized version of our method, $\mathbf{w}$ is also defined as Eq.5, and $\mathbf{W} \sum_{KDBFM} \mathbf{W}^T = \boldsymbol{\alpha} \mathbf{K}_b \boldsymbol{\alpha}^T$. $\mathbf{K}_w$ is the kernel within-class scatter matrix corresponding to the new within-class scatter matrix defined in Eq.24. Now, we will use a simple method to implement the KDBA.

4.1 The Algorithm of KDBA

The KDBA improves the generalization capability by decomposing its procedure into a simultaneous diagonalization of two matrices. We can refer to Nullspace Linear Discriminant Analysis (NLDA) method about the discriminant procedure [10]. The simultaneous diagonalization is stepwisely equivalent to two operations, and we first whiten $\mathbf{K}_t = \mathbf{K}_b + \mathbf{K}_w$ as following:

$$\mathbf{K}_t \boldsymbol{\Xi} = \boldsymbol{\Xi} \boldsymbol{\Gamma} \text{ and } \boldsymbol{\Xi}^T \boldsymbol{\Xi} = \mathbf{I}, \tag{25}$$

$$\boldsymbol{\Gamma}^{-1/2} \boldsymbol{\Xi}^T \mathbf{K}_t \boldsymbol{\Xi} \boldsymbol{\Gamma}^{-1/2} = \mathbf{I}, \tag{26}$$

where Ξ, Γ are the eigenvector and the diagonal eigenvalue matrices of $\mathbf{K}_t$. We can get the eigenvectors matrix Ξ', whose eigenvalues are bigger than zero (Corresponding diagonal eigenvalue matrix is $\Gamma'^{-1/2}$). The new within-class scatter matrix is computed by using the following method:

$$\Gamma'^{-1/2}\Xi'^{T}\mathbf{K}_w\Xi'\Gamma'^{-1/2} = \Xi_w. \tag{27}$$

Diagonalizing now the new within-class scatter matrix Ξ_w.

$$\Xi_w\theta = \theta\gamma \text{ and } \theta^T\theta = \mathbf{I}, \tag{28}$$

where θ, γ are the eigenvector and the diagonal eigenvalue matrices of Ξ_w in an increasing order. We remove the Eigenvectors, whose eigenvalues are far from zero, and the remained Eigenvectors construct the transformation matrix θ'.

The overall transformation matrix is now defined as following.

$$\alpha' = \Xi'\Gamma'^{-1/2}\theta'. \tag{29}$$

We use $\mathbf{w}'$ as the transform matrix, v is the extracted feature calculated by using Eq.30

$$v = \mathbf{w}'\Phi(x) = \sum_{i=1}^{n} \alpha_i' k(x_i, x). \tag{30}$$

4.2 Similarity Measure for KDBA

If v_1, v_2 are the feature vectors corresponding to two face images x_1, x_2, which are calculated by using the Eq.30, then the similarity rule is based on the cross correlation between the corresponding extracted feature vectors as following:

$$d(x_1, x_2) = \frac{v_1 \cdot v_2}{\| v_1 \| \cdot \| v_2 \|}. \tag{31}$$

The first experiment is tested on the Text classification problem. The other experiments are performed on the large CAS-PEAL database, and the comparative performance is carried out against the Eigenface, Fisherface and GKFD.

5 Experiment

5.1 Text Classification

We first make an experiment on the two-class Text classification problem. The dataset is the Example2(Details of the dataset can be referred to http://svmlight.joachims .org/), the training database of which contains 5 positive examples and 5 negative examples, and the test database contains 600 samples. In the case of k=1, KDBA achieves 84.5% accuracy rate, and the SVM-Light package is 84.3%.

Fig.3 shows the decision value calculated by SVM-Light, and the first feature found by KDBA for all 600 test examples.

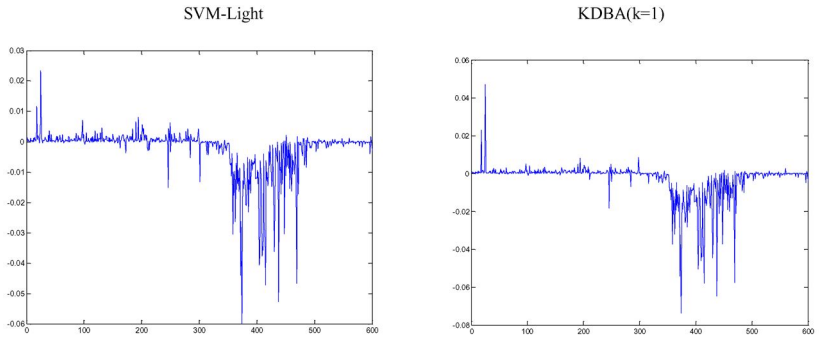


Fig. 3. Experiment on the Text classification problem

5.2 CAS-PEAL Face Database

In our experiments, the face image is cropped to size of 64X64 and overlapped with a mask to eliminate the background and hair. For all images concerned in the experiments, no preprocessing is exploited. To speed up the system, we first make PCA on the face images, and the lower dimensional vector in the PCA space is used in our experiments to capture the most expressive features of the original data. The CAS-PEAL face database was constructed under the sponsors of National Hi-Tech Program and ISVISION [14]. The goals to create the CAS-PEAL face database include: providing the worldwide researchers of FR community a large-scale face database for training and evaluating their algorithms; facilitating the development of FR by providing large-scale face images with different sources of variations, especially Pose, Expression, Accessories, and Lighting (PEAL); advancing the state-of-the-art face recognition technologies aiming at practical applications especially for the oriental. Currently, the CAS-PEAL face database contains 99,594 images of 1040 individuals (595 males and 445 females) with varying Pose, Expression, Accessory, and Lighting (PEAL). Gallery set contains one image for each person. One sample person was shown in Fig.4, and the size of the face image is 360X480. In this experiment, only one face image for each person was used as Gallery database. Details of the face database are shown at http://jdl.ac.cn [14].



Fig. 4. Sample of Face Images in CAS-PEAL database

Table 1. Experiment Result on CAS-PEAL database (Accurate rate)

	Eigenface	Fisherface*	GKFD	KDBA(k=1)	KDBA(k=5)	KDBA(k=10)
Accessory	37.1	61	58.7	62.4	63.1	64
Aging	50	72.7	77.3	84.8	87.9	87.9
Distance	74.2	93.5	94.9	96	96	96
Expression	53.7	71.3	78.2	78.2	78.5	79.9
Background	80.5	94.4	91.7	94.1	94.3	94.4

* Fisherface method refers to[15]

From above experiments, the KDBA method has achieved better performance than other popular face recognition schemes. The Kernelized Decision Boundary Feature vector can be thought of as the optimal direction in terms of SRM theory. We also know that, from the experiments, GKFD is just a little better than the Fisherface method, since the Fisherface method is based on the Enhanced Fisher Model[15], which can reserve the useful discriminant information by performing PCA on the within-class scatter matrix, and then modify between-class scatter matrix to get the whole transformation matrix. We also can know that the bigger k is used here, and a litter better performance of the face recognition system will be achieved. However, the complexity will also be increased accordingly, so we often choose a small value for k .

6 Conclusion and Future Work

We have proposed a novel discriminant analysis method named by KDBA. The contributions of our method include: (1) the proposed nonlinear discriminant approach, a kernelized extension to the Decision Boundary Analysis, is easily implemented and suitable to the large sample size problem by dividing the training database into several sub-datasets. (2) The KDBFM is constructed based on the normal vector of the optimal Decision Boundary (Decision Hyperplane of SVM) in terms of the SRM, and the PCs of which is the intrinsic discriminant subspace. (3) We also utilize a simple method to prove the property of the SVM, which is combined with KDBFM to make our method consistent with the Fisher Criterion. The feasibility of the new method has also been successfully tested on Text classification problem, and the face recognition task using data sets from the CAS-PEAL database, which is a very large one. The effectiveness of the method is shown in terms of accurate rate against some popular face recognition schemes, such as Eigenface, Fisherface, GKFD, and so on.

Gabor wavelet feature has been combined with some discriminant methods and successfully used in the face recognition problem [16, 17]. Therefore, we try to make full use of Gabor wavelet representation of face images before using KDBA to get the transformation matrix.

Acknowledgement

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, "100 Talents Program" of CAS, ShangHai Municipal Sciences and Technology Committee (No. 03DZ15013), and ISVISION Technologies Co., Ltd. Specially, thanks for Dr.Shiguang shan's advice.

References

1. B. Scholkopf, A. Smola, K.R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol.10, pp.1299-1319, 1998.
2. S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K.R. Muller, "Fisher discriminant analysis with kernels, " IEEE International Workshop on Neural Networks for Signal Processing, pp.41-48, 1999.

3. G. Baudat, F. Anouar, "Generalized discriminant analysis using a kernel approach, " Neural Computation, vol.12, no.10, pp.2385-2404, 2000.
4. Jian yang, Alejandro, "A new Kernel fisher discriminant algorithm with application to face recognition," Letters of Neurocomputinting, 2003.
5. Z. hong and J. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," Pattern recognition, vol.24, no.4, pp.317-324, 1991.
6. K. Fukunaga, J. M. Mantock, "nonparametric Discriminant Analysis," PAMI, Vol5(6), pp.671-678, 1983
7. Rik Fransens, Jan De Prins, "SVM-based Nonparametric Discriminant Analysis, An application to Face Detection," ICCV2003.
8. C.Lee, D.D. Langrebe, "Feature Extraction Based on Decision Boundaries," PAMI, Vol.15(4). Pp.388-400, 1993.
9. Qingshan Liu, Rui Huang, "Face Recognition Using Kernel Based Fisher Discriminant Analysis, " The 5th International Conference on Automatic Face and Gesture Recognition, pp.187-191, 2002.
10. Wei Liu, Yunhong Wang, "Null space based Kernel Fisher Discriminant analysis for face recognition," The 6th International Conference on Automatic Face and Gesture Recognition, 2004.
11. C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," Knowledge Discovery and Data Mining, vol.2, pp.121-167, 1998.
12. G. Guo, S.Z. Li, and C. Kapluk. "Face recognition by support vector machines," Image and Vision Computing, vol.19, pp.631--638, 2001.
13. A. Tefas, C. Kotropoulos, and L. Pitas, "Using Support Vector Machines to Enhance the Perfornance of Elastic Graph Matching for Frontal Face Authentication, " IEEE Trans. on PAMI, Vol. 23, No. 7, pp.735-746, 2001.
14. Wen Gao, Bo Cao, Shiguang Shan, "The CAS-PEAL Large-Scale Face Database and Evaluation Protocols," Technical Report No. JDL_TR_04_FR_001, JDL, CAS, 2004.
15. Chenjun liu, Harry Wechsler, "Enhanced Fisher Linear Discriminant Models for face Recognition," 14th International Conference on Pattern Recognition, vol.2, pp.1368-1372, 1998.
16. Chengjun Liu and Harry Wechsler, "Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition, " IEEE Trans. Image Processing vol.11 no.4, pp.467-476, 2002.
17. Baochang Zhang, Wen Gao, Shiguang Shan, Yang Peng, "Discriminant Gaborfaces and Support Vector Machines Classifier for Face Recognition, " Asian Conference on Computer Vision, pp.37-42, 2004.
18. Stan Z. Li and Anil K. Jain (Eds). "Handbook of Face Recognition," Springer, to be published in 2004.

A Probabilistic Approach to Semantic Face Retrieval System

Karthik Sridharan, Sankalp Nayak, Sharat Chikkerur, and Venu Govindaraju

Center for Unified Biometrics and Sensors University at Buffalo, NY, USA
{ks236,snayak,ssc5,govind}@buffalo.edu

Abstract. A probabilistic system that retrieves face images based on verbal descriptions given by users is proposed. This interactive system prompts the user at each stage of query to provide a description about a facial feature that will help it to retrieve the required face image. It is a soft biometric system that groups people by the description of their faces. The method proposed automates the process of extracting general verbal descriptions of faces like 'long nosed' or 'blonde haired' and performs queries on them. The proposed method uses Bayesian learning for the query process and hence is more immune to errors in both the extraction of semantic descriptions and user given information. It was found that the required face image appeared 77.6% of the time within the top 5 and 90.4% of the time within the top 10 retrieved face images.

1 Introduction

Face Recognition has beeen proven to be very popular in several biometric and law enforcement applications. There have been many algorithms proposed for face detection and recognition [11]. However all these system require require a pictorial description of the person to be searched or verified. However, in law enforcement applications the photograph of the suspect is not usually present and the only available evidence is in the form of verbal description by the witness. Performing face retrieval against verbal queries is a challenging problem. Existing face retieval systems used in law enforcement are based on synthesizing a face image and performing a traditional face recognition process. We propose a system that automatically retrieves the semantic description of each face and stores it in a meta database. For a user query which is semantic, we retrieve all images that match the description from the meta database. Thus, our system is very useful in retrieving images of suspects from a database based on queries constructed from the verbal descriptions given by witnesses. The system can also be used to speed up the process of identification of a face in large databases by first extracting the semantic features of the given face and then matching with the faces in the database in the order in which images are retrieved for the query.

A snapshot of the proposed system is shown in Figure 1. The user can select a particular feature on which he wants to query and choose one of the descriptions of the feature from a list. For example, the user can select the feature 'mustache' and provide the description that the person has a mustache. For relative queries

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 977–986, 2005.
© Springer-Verlag Berlin Heidelberg 2005



Fig. 1. Snapshot of the System

like 'lip thickness', the user can view the top few retrieved images and with respect to these images can select a description like 'normal', 'thick' or 'thin' for the lip of the required person. The user can base his query on the top few images retrieved at each step. Further the system also prompts the user to provide a description about the feature that is most discriminant amongst images that are more likely to be the required face image.

2 Related Work

Law enforcement agencies have been using Identikits [1, 12] for composing a rough sketch of the face. Identikit consists of a set of transparencies of various facial features that can be combined to build up a picture of the person sought. Forensic artists transform the verbal description of the suspect given by the witness into rough sketch of the suspects by putting together these transparencies to fit the description given by user. Once the rough sketch is made, large image databases are searched manually to find faces that resemble the sketch. The process is iterative and laborious.

Many systems such as Evofit [2] have tried to automate the traditional identikit concept by evolving the required face based on user feedback from faces present in the database. The Phantomas [3] is an automated facial database search system using Elastic Graph Matching to recognize faces. The system takes as input the composed face images and retrieves images from the database in the order of their similarity to the composed picture. However in these systems there is a need for composing the required face realistically before retrieval.

The Computer Aided Facial Image Identification, Retrieval and Inference System (CAFIIRIS) [4] for criminal identification stores and manages facial images and criminal records and provide necessary image and text processing and editing tools. The system uses a combination of feature-based PCA coefficients, facial landmarks, and text descriptions to construct index keys for an image. The advantage of this system is that it is interactive since it uses relevance feedback from user to retrieve relevant images. Photobook [5] is a content-based retrieval system that provides methods for searching several types of related image databases including faces. One potential problem in this system is that user

can keep on cycling through the same set of images if it gets stuck in local maxima. Also, since the eigenfaces store the most common components, it would be difficult if the desired face is quite different from those present in the database

In both these systems and many other recent systems that use PCA, semantic description of the face is captured indirectly by eigenfaces, as PCA capture only the most significant features. However the problem of synthetic face composition and the fact that user description in many cases can be limited to a verbal description limits these systems. Often, simple verbal descriptions of a person like "thick lipped" or "blonde haired" help in narrowing down the potential faces efficiently. Further, during the enrollment or data population stage we are generally not constrained by time but during query stage, we need the results immediately. The proposed system extracts simple verbal descriptions of the face saves them in a meta database and thus speeds up query results by a large magnitude.

3 System Overview

The proposed system can be broken down into Enrollment sub-system and Query sub-system. Figure 2 shows the model of the system. The Enrollment sub-system accepts as input mugshot face images with frontal view. It outputs semantic descriptions of the face like whether the person was wearing spectacles, whether the person had a long nose etc. The Query sub-system accepts the semantic descriptions of a person's face given by the user and retrieves images in the order of how well they fit the given description. This sub-system also prompts the user at each stage about which feature that would result in an optimal query.

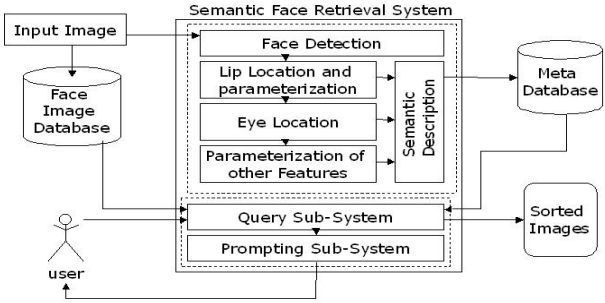


Fig. 2. System Overview

The enrollment sub-system first performs face detection based on skin color based image segmentation. Further, to extract the semantic descriptions of the face, we first need to localize facial features and parameterize them. Since the enrollment process is offline, we can afford to parameterize facial features by exact contours but for effectively describing features, such precision is not necessary. Hence to parameterize these features, we fit the lowest order polygon that

can effectively describe them. For example, fitting a triangle for the nose can describe its length width and size. When a new person is added to the database, the sub-system automatically extracts the semantic description of the face from the image and stores this in a database of semantic descriptions. During the facial features localization, the system first performs lip and eye detection and uses their location to reduce the space in which we search for the remaining features.

The Query sub-system accepts as input from the user the semantic description of the face of the person and retrieves the images in the order of how well they fit the description. For the ordering, the probability that the face is the one being queried about is found using bayesian approach for each image. This probability is used to sort the images. At each stage, the sub-system also finds which feature has the most discriminating power in the database and prompts the user to provide information about that feature.

4 Enrollment Sub-system

It is the responsibility of the enrollment sub-system which works offline to create the database of semantic descriptions of the faces.

4.1 Face Detection

Since all the images considered are frontal images of faces with non-intensive background and little illumination variance, skin color based face detection works very well. Each pixel in the image can be viewed as a vector in the 3D RGB color space. Skin color based region segmentation is done by thresholding the unit color vectors [8] . Since the intensity holds no color information, unit vectors are used for the color segmentation.

4.2 Lip Localization and Parameterization

From [9] we see that efficient lip detection can be done by segmenting image based on lip color on the face image region. Since lip color is distinct in the face region, we can effectively perform lip detection based on lip color. However the problem with this method is that the lip corners are not captured as they have very low intensity. We use histogram based object segmentation [10] method to overcome this problem. Since we do face detection based on skin color, once we locate the face we subtract the skin color regions in the image. Thus we are left with the image of the face with only facial features like lips, eyes, eyebrows etc. Now by lip color detection, we find the approximate location of the lips. Next we find the histogram of this region of the image to perform segmentation. To segment the image, the histogram is first smoothed. Next the points at which the slope of histogram changes sign is termed as valleys and these points are used as threshold to segment the image. Once the quantization of the lip region of the image is performed, based on the quantized image the lip width is altered



Fig. 3. Lip Localization

to fit the lip well. Figure 3, shows the initial color based detection results the histogram based segmented image and final detection result.

4.3 Eye Localization and Parameterization

Once lips are located we can reduce the area in which we are to search for eyes. We use the detail that the eyes are above the lips and hence we search for eyes only in a rectangular strip above the lips. We use circle detection based on hough transform as proposed in [7] for eye localization. However, to use hough transform for circle detection, we need to know the radius of the circle to be searched. In the database used, the radius of pupil of all the eyes varied from 9 to 13 pixels and the mean radius was 11 pixels. Using the set of 5 radii for each point in the image we get a set of 5 accumulator values. Usually to find out the location of the eyes, the point that has maximum sum of accumulator values for all the radii is chosen as the center for the pupil. However we know that in the database most pupils have radius of 11. Hence to improve the accuracy of eye detection, we use a probabilistic model to locate the most probable eye center. We convert the accumulator scores to probability of the point being the center of circle of radius r_j by dividing it by $2\pi r_j$. Since the accumulator can have a maximum value of $2\pi r_i$ the circumference representing a complete circle. Now we need to find out for each point $point_i$, the marginal probability of it being the center of pupil. This is calculated as

$$P(point_i) = \sum_{j=1}^{5} (P(point_i|r_j)P(r_j))$$

where $P(point_i|r_j) = \frac{Accumulator(point_i, r_j)}{2\pi r_j}$ and $r_{1..5} = [9..13]$. The priors, that is the probability of a particular eye having a particular radius was set by trial and error as $P(r_1..r_5) = [0.1, 0.2, 0.4, 0.2, 0.1]$. Finally the point in the image with maximum marginal probability $P(point_i)$ was set as the center for the pupil. The process is done for both left and right eyes separately to locate the two eye centers.

4.4 Localization and Parameterization of Other Features

Once the eye centers and lip center triangle is formed we can use face proportions to locate all other features based on knowledge about the face that we have. For instance, from anthropological studies we know that distance between midpoint between the eyes and nose is about two-third of the distance between eyes and lips. Thus we get the approximate line of nose-ending. Then we use color based

edge detection to trace out the curves of the nostrils and using these we fit a triangle for the nose. The color-based edge detection is performed by assuming that each pixel is a vector in 3D RGB space. Thus by thresholding angle between the color vectors of two pixels, using a sobel mask we detect edges [13]. By using color based edge detection, we can select threshold to detect even soft edges like the ones formed by nose. As nose edges are formed by transition of pixels from skin color to skin color, we can only detect edges that transition from skin color to skin color. The color edge map can also be used along with knowledge about the face to detect facial hair regions- eyebrows, beard, moustache etc. On the approximate region of the feature we are looking for we can apply the color based edge detection and check if the region is present. The list of semantic features extracted and their type is summarized in Table 1.

Table 1. List of Features

Feature	Type
Spectacles, Beard , Mustache, Long Hair and Balding	Discrete (yes/no)
Hair Color	Discrete (Black/Brown/Blonde)
Nose Width, Length and Size, Lip Width and Thickness, Face Length, Darkness of Skin and Eyebrow Thickness	Continuous

5 Query Sub-system

The query sub-system performs the retrieval based on semantic description given by the user. It also prompts the user about which feature to query next.

5.1 Retrieving Images

Based on the description given by the user, the system at each stage orders the images according to their probability of being the face we are looking for. The system deliberately does not prune the images as pruning the images based on wrong information given by user would mean elimination of the required image from the list we are searching in. Initially, before user provides any information, we set the probability of a face being the required face $P(face)$to be $1/n$ where n is the number of faces in the database. Now as the user provides the description d_j about each feature f_i, we use this to update the probability using bayesian learning as

$$P(face_k|f_i = d_j) = \frac{P(f_i = d_j|face_k)P(face_k)}{\sum P(f_i = d_j|face_k)P(face_k)}$$

After each query the prior probabilities for the faces are made equal to the posteriors found. The probability $P(f_i = d_j|face)$ of the feature f_i matching the description d_j for each face is set for binary attributes like whether the person has mustache or not by 0.9 if the face has feature f_i matching the given description d_j and 0.1 otherwise. The probabilities aren't set to 0 and 1 to make

the system robust to user or enrollment system errors. For continuous valued features, the probability is set by normalizing the continuous value between 0 and 1.

5.2 Prompting the User

The system at each step prompts the user to enter a description about the feature that will help to effectively retrieve the required image. To do this the system should prompt the user to enter information about the feature having most entropy. More the entropy, more discriminative is the feature. For instance if half the people in the database wear spectacles and other half don't, it would be a better feature to query about than a feature like fair skinned which most people in the database may be.

However there are two problems in doing this. Firstly, for finding the entropy we need to discretize the continuous values of features. However descretizing the values, we may loose relative information. For instance, it may happen that when we initially descritized nose length, the required person may have had a medium nose. But after a couple of queries it may happen that the required person's face has a longer nose among the more probable faces. Thus, by descritizing we cannot capture this information. The second problem is that we can't just find entropy of each attribute assuming that all descriptions of the feature are equally probable. The probability of a feature having a particular description is governed by the probability of faces having that description for the given feature.

To overcome the first problem, we descritize the continuous features into low, medium and high using appropriate thresholds and save them separately in a table while also keeping the continuous values in a table for calculating probabilities. Further, instead of assuming equal probabilities for all descriptions of a feature during calculation of entropy, we use the probabilities of the attributes given the current probabilities of faces. Given that each face k has probability $P(face_k)$ of being the required one, the probability of some feature f_i having description d_j is given by sum of probabilities of all faces which have $f_i = d_j$. For instance, the probability of nose being long is the sum of probabilities of faces with long nose. Thus we calculate entropy Hs_i for the ith attribute as

$$Hs_i = -\sum_{j=1}^{m} P(f_i = d_j|P(face))log_2(P(f_i = d_j|P(face)))$$

where m is the number of total values the attribute can take and

$$P(f_i = d_j|P(face)) = \frac{\sum_{k:f_{i,k}=d_j} P(face_k)}{\sum_{p=1}^{n} P(face_p)}$$

where $f_{i,k}$ represents feature i of face k.

6 Performance Analysis

The database used for testing the system [6] had 55 frontal images of Females and 70 frontal images of males all in a white background. For testing the system

users were shown images of people in the database taken on a different day with different clothes.

6.1 Example

An Example query is shown in Figure 4 where the user is querying about the person marked by the rectangular box. Each row in the Figure 4 shows the top five images after each query. Initially as all images are equally possible, the first five are just the first five images in the database. The required person was at the 31st place. Now when the information that the person was wearing spectacles was provided we see that images of people wearing spectacles appear in the top 5. The required person was at the 13th position. When the information that the person had a mustache was provided he appears in the third position among the top five as seen in the third row of Figure 4. Finally when the information that the person had a large nose was provided we see that the person moves to second place among top five Figure 5 shows the probabilities of faces in sorted order for the above example query.

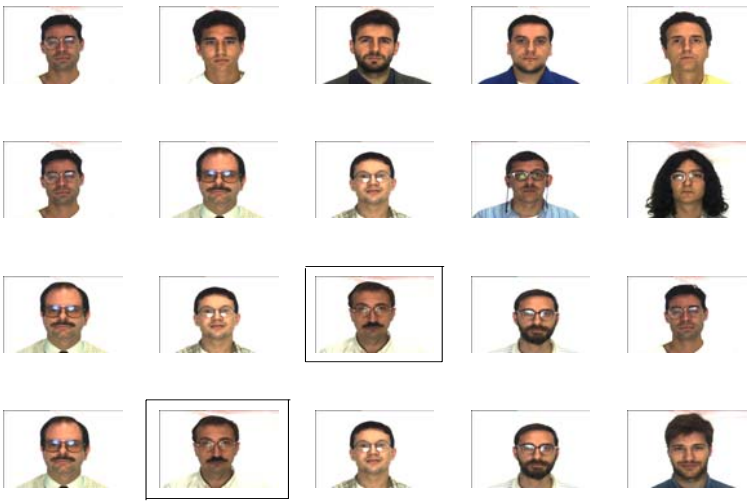


Fig. 4. Example Query

6.2 Evaluation of Enrollment Sub-system

The Table 2 summarizes the results of the individual feature extraction of the enrollment sub-system. The performance on the continuous valued features like nose width can be evaluated by how well the polygons fit the features and how easily the user can locate the required person.

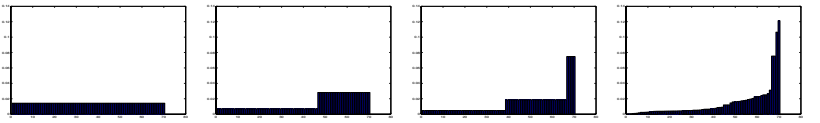


Fig. 5. Plot of Probabilities

Table 2. Performance of Query Sub-system on Discrete Valued Attributes

Feature	Number of False Accepts	Number of False Rejects
Spectacles	1	2
Mustache	2	4
Beard	4	0
Long Hair	2	8
Balding	1	0

6.3 Evaluation of Query Sub-system

Experiments were conducted to test the usability and query capability of the system. 25 users were each shown pictures of 5 people in the database taken on different days and wearing different clothes from the ones in the database. Then the users were asked to input the verbal descriptions of the 5 faces to the system. Table 3 summarizes the average number of queries required to get the person we are looking for within top five, ten and fifteen images respectively for the 125 test cases.

Table 3. Average Queries Needed for Retrieval

-	Top 5	Top 10	Top 15
Average No. of Queries	6.64	4.59	2.72

7 Conclusion

We have presented a probabilistic face retrieval system based on verbal query which is interactive in nature. From the results shown above we can see that the system performs the extraction of semantic features from images effectively. The results of query show the effectiveness of the system. The system will find application in law enforcement for picking out the image of suspect from a huge database based on verbal description given by the witness.

The system can be further improved to handle pose variations in the images. The system can also be extended to do the retrieval from not just still images but also video clips. Retrieval from video will make the system very useful in surveillance applications.

References

1. V. Bruce, *Recognizing Faces*, Faces as Patterns, pp. 37-58, Lawrence Earlbaum Associates, 1988
2. Frowd, C.D., Hancock, P.J.B. and Carson, D. *EvoFIT: A Holistic, Evolutionary Facial Imaging Technique for Creating Composites*, ACM TAP, Vol. 1 (1), 2004

3. *Phantomas Elaborate Face Recognition*,Product description: http://www.global-security-solutions.com/FaceRecognition.htm
4. J. K. Wu, Y. H. Ang, P. C. Lam, S. K. Moorthy, A. D. Narasimhalu,*Facial Image Retrieval, Identification, and Inference System*, Proceedings of the first ACM international conference on Multimedia, pp. 47-55, 1993.
5. A. Pentland, R. Picard, S. Sclaroff, *Photobook: tools for content based manipulation of image databases*, Proc. SPIE: Storage and Retrieval for Image and Video Databases II, vol. 2185
6. A.M. Martinez and R. Benavente, *The AR face database*, CVC Tech. Report No. 24, 1998.
7. David E. Benn, Mark S. Nixon, John N. Carter, *Robust Eye Centre Extraction Using the Hough Transform*, AVBPA, pp. 3-9, 1997.
8. Vezhnevets V, Sazonov V, Andreeva A, *A Survey on Pixel-Based Skin Color Detection Techniques*, Proc. Graphicon-2003, pp. 85-92, Moscow, Russia, September 2003
9. Wark, T. and Sridharan, S. ,*A syntatic approach to automatic lip feature extraction for speaker identification*, ICASSP, pp. 3693–3696, 1998.
10. S.C. Sahasrabudhe, K.S.D. Gupta, *A Valley-seeking Threshold Selection Technique*, Computer Vision and Image Processing, (A. Rosenfeld, L. Shapiro, Eds), Academic Press, pp.55-65, 1992.
11. M. Turk and A. Pentland,*Eigenfaces for recognition*,Journal of Cognitive Neuroscience, 3(1), pp. 71-86, 1991.
12. K. R. Laughery and R.H. Fowler *Sketch Artists and Identi-kit*, Procedure for Recalling Faces, Journal of Applied Psychology, 65(3), pp. 307-316, 1980.
13. R. D. Dony and S.Wesolkowski, *Edge detection on color images using RGB vector angles*, Proc. IEEE Can. Conf. Electrical and Computer Engineering, pp. 687–692, 1999.

Authenticating Corrupted Facial Images on Stand-Alone DSP System

Sang-Woong Lee, Ho-Choul Jung, and Seong-Whan Lee*

Center for Artificial Vision Research/
Department of Computer Science and Engineering, Korea University,
Anam-dong, Seongbuk-ku, Seoul 136-713, Korea
{sangwlee,hcjung,swlee}@image.korea.ac.kr

Abstract. In this paper, we propose a method of authenticating corrupted photo images based on noise parameter estimation and implement an authentication system using TMS320C6711 DSP chip. The proposed method first generates corrupted images and the noise parameters in the training phase. With a corrupted image and an original image, the noise parameters of the corrupted photo image can be estimated in the testing phase. Finally, we can make a synthesized photo image from the original photo image using the estimated noise parameters. We made some experiments on the prototype of the stand-alone system to verify the performance of the proposed method and to apply for real-life applications. The experimental results on this system show that the proposed method can estimate the noise parameters accurately and improve the performance of photo image authentication.

1 Introduction

Up to now, the photo image authentication market is expanding, and many fields, such as ID card authentication systems and biometric passport systems, are starting to use photo image authentication techniques for security reasons. Photo image authentication refers to the verification of a scanned facial image of an identification card, passport or smart card based on its comparison with an original facial image contained in a database or stored on a chip(Fig. 1). However, the scanned photograph used to be corrupted by real problems, such as scratch, blur and discoloration(Fig. 2). In fact, handling corrupted photo images is one of the most difficult and commonly occurring problems in image processing applications. Additionally, most of the current approaches to face authentication require at least two training images per person, in order to obtain good performance. Unfortunately, in real-world tasks, such a requirement cannot always be satisfied. From a different standpoint, we have witnessed the explosion in interest and progress in automatic face recognition and authentication technology during the past few years. Many systems, implemented on workstation or PC, are already deployed expensively as a component of an intelligent building system and a security system for gate control. However, hardware cost and

* To whom all correspondence should be addressed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 987–996, 2005.
© Springer-Verlag Berlin Heidelberg 2005

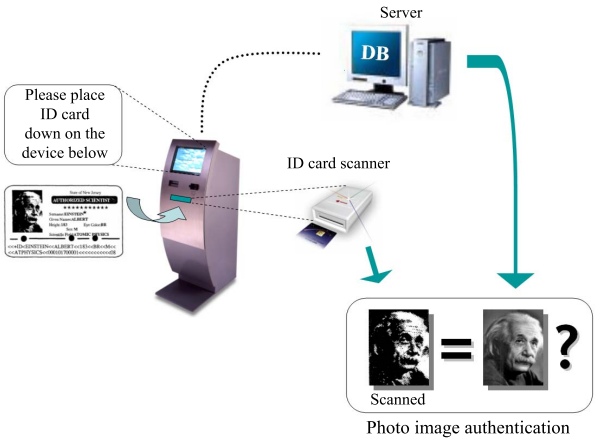


Fig. 1. Example of photo image authentication system



Fig. 2. Example of corrupted face images scanned from identification cards

volume often limit the application using facial technologies such as interactive toys and mobile devices, etc. Therefore, it is needed that the systems become smaller and contain faster algorithms. In this paper, in order to solve the above problems, namely the corruption of photo images, the requirement of multiple training images per person, and the compactness of stand-alone system, we propose an efficient photo image authentication method based on noise parameter estimation and also design and implement our method on stand-alone system using DSP chip for real applications. The utilization of this method based on noise parameter estimation is subject to two preconditions. Firstly, that the size of the original and scanned photo images are the same and that they are normalized in terms of scale, rotation and translation. Secondly, that the corruption of the images is less 30% of the whole image area.

2 Related Work

Research into face authentication has been carried out for a long time. In particular, there are several approaches which can be taken to solve the noise problem and to eliminate the requirement of multiple training images per person. Herein, we introduce two popular approaches and a series of approaches for hardware implementation.

2.1 Corrupted Image Analysis

Sanderson et al. [2] proposed a method of extracting robust features in various image conditions. To accomplish this, they proposed a new feature set, utilizing polynomial coefficients derived from 2D Discrete Cosine Transform(DCT) coefficients obtained from horizontally and vertically neighboring blocks. The proposed feature set is superior (in terms of its robustness to illumination changes and discrimination ability) to the features extracted using previous methods. This method is based on robust feature extraction against Gaussian white noise and Gaussian illumination changes, however, it does not consider the question of which features are required for the purpose of authentication in images corrupted by scratch, blur and discoloration.

2.2 Reconstruction of Partially Corrupted Image

Turk and Pentland [4] proposed a method of reconstructing noisy or missing parts of a partially corrupted face using eigenfaces based on Principal Component Analysis(PCA). However, their method showed good results only when applied to an unknown face of a person for whom multiple images are available in the training set, or a face that was itself part of the initial training set.

Takahashi et al. [3] also proposed a method of removing noise using KPCA(Ke rnel Principal Component Analysis). This method is able to remove outliers in data vectors and replace them with the values estimated via KPCA [1]. By repeating this process several times, it is possible to obtain feature components less affected by the outliers. This method is more effective at outlier removal than the standard method of PCA proposed in [4]. However, it is not efficient for real-time face authentication, because it takes too much time to remove the noise using the kernel function.

2.3 Stand-Alone DSP System for Face Authentication

In the early 1990s, Gilbert et al. introduced a real-time face recognition system using custom VLSI hardware for fast correlation in an IBM compatible PC[10]. Five years later, Yang et al. introduced a parallel implementation of face detection algorithm using a TMS320C40 chip[9]. On the other hand, IBM introduced a commercial chip, ZISC which can compute the classification in RBF(Radial Basis Function) based neural network[8]. However, these systems did not implement the whole stages of a face authentication system. Unlike these, we have implemented the entire steps using only a TMS320C6711 DSP chip. This will enables face authentication system to be applied to diverse applications.

3 Noise Model

3.1 Noise Analysis in the Case of Corrupted Images

In this paper, we assume that the corruption of the images originates from changes in contrast, brightness, Gaussian noise and Gaussian blur [7].

Firstly, we define an image whose contrast and brightness are changed, as follows:

$$I_{CB}(x, y) = c \times I_{org}(x, y) + b \tag{1}$$

where I_{CB} is the image corrupted by the change of contrast and brightness, I_{org} is the original image, c is the contrast parameter, and b is the brightness parameter.

Secondly, we define a corrupted image which is generated by applying Gaussian blur, as follows.

$$I_G(x, y) = I_{org}(x, y) * G_{blur}(x, y) \tag{2}$$

where G_{blur} is the Gaussian blur filter, G is Gaussian blur function and $*$ is the image convolution operator.

$$G_{blur}(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{3}$$

where σ in (3) is the Gaussian blur parameter. Fig. 3 shows examples of corrupted images so generated.



Fig. 3. Examples of corrupted images (a) Adjustment of contrast and brightness (b) Gaussian blur

3.2 Definition of Noise Model

In this section we will formally specify the noise model. We define the corrupted image, I^c, as follows:

$$I^c = I_{CB} * G_{blur} \tag{4}$$

Then, more formally, the noise model is defined as the combination of corrupted images, I^c, and the noise parameters, P.

$$N_i = \begin{pmatrix} I_i^c \\ P_i \end{pmatrix} \quad (i = 1, \cdots, m) \tag{5}$$

where $I^c = (x_1, \cdots, x_k)^T$, $P = (p_1, \cdots, p_l)^T$. $x_1, \cdots, x_k$ are the intensities of the pixels in the corrupted image, k is the number of pixels in the corrupted image, p is the parameter value, l is the number of noise parameters used and m is the number of corrupted images. In this paper, we used $l = 3$, $p_1 = c$, $p_2 = b$, and

$p_3 = \sigma$ since we consider the changes in contrast, brightness and Gaussian blur. Thus, the noise model, N, is represented as follows:

$$N = \bar{N} + \sum_{i=1}^{m-1} \alpha_i n_i(j), (j = 1, \cdots, k, k+1, \cdots, k+l) \tag{6}$$

where $\bar{N}$ is the mean of $N_i (i = 1, \cdots, m)$. By PCA, a basis transformation is performed to an orthogonal coordinate system formed by eigenvector n_i of the covariance matrices on the data set of m corrupted images and noise parameters. The probability for coefficients $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} \in R^{m-1}$) is defined as:

$$p(\boldsymbol{\alpha}) \sim \exp\left[-\frac{1}{2}\sum_{i=1}^{m-1}(\frac{\alpha_i}{\xi_i})^2\right] \tag{7}$$

where with ξ_i^2 being eigenvalues of the covariance matrix, C_s.

4 Photo Image Authentication

In order to authenticate corrupted photo images, the proposed method includes the following two phases, the training phase and testing phase. In the training phase, we first generate corrupted images by adjusting the parameters of contrast, brightness and Gaussian blur of an original photo image. Then, we obtain the basis vectors of the corrupted images and the noise parameters. In the testing phase, the photo image authentication procedure for the corrupted photo image is performed through several steps

4.1 Noise Parameter Estimation

Using the noise model, only an approximation of the required parameters can be obtained. The goal is to estimate the noise parameters by finding an optimal solution in such an overdetermined condition. At first, we want to find the value of α which satisfies Equation (8).

$$\tilde{N}(j) = \sum_{i=1}^{m-1} \alpha_i n_i(j), \ (j = 1, \cdots, k) \tag{8}$$

where j is the pixel in the corrupted image, k is the number of pixels in the corrupted image and the difference image is defined as $\tilde{N} = N - \bar{N}$. Generally, there may not exist any value of α that perfectly fits $\tilde{N}$. Therefore, we choose $\alpha^\star$ to minimize the error function described in Equation (8).

To do this, we first define an error function, $E(\alpha)$, in Equation (10), and set a condition to minimize the error function. The goal is to find the value of α which minimizes the error function, $E(\alpha)$, according to the following equation:

$$\alpha^\star = \arg\min_{\alpha} E(\alpha) \tag{9}$$

The error function is given as:

$$E(\alpha) = \sum_{j=1}^{k} \left(\tilde{N}(j) - \sum_{i=1}^{m-1} \alpha_i n_i(j) \right)^2 \tag{10}$$

We then find the coefficient values that minimize the error function using the least-square minimization method. According to equations (9) and (10), we can solve this problem by the least-square method. Equation (8) is equivalent to the following:

$$\begin{pmatrix} n_1(1) \cdot n_{m-1}(1) \\ \vdots \quad \ddots \quad \vdots \\ n_1(k) \cdot n_{m-1}(k) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} = \begin{pmatrix} \tilde{N}(1) \\ \vdots \\ \tilde{N}(k) \end{pmatrix} \tag{11}$$

We can rewrite (11) as:

$$\mathbf{I_N}\alpha = \tilde{\mathbf{I_N}} \tag{12}$$

where

$$\mathbf{I_N} = \begin{pmatrix} n_1(1) \cdot n_{m-1}(1) \\ \vdots \quad \ddots \quad \vdots \\ n_1(k) \cdot n_{m-1}(k) \end{pmatrix}, \alpha = (\alpha_1, \cdots, \alpha_{m-1})^T, \quad \tilde{\mathbf{I_N}} = (\tilde{N}(1), \cdots, \tilde{N}(k))^T \tag{13}$$

The least-square solution to an inconsistent $\mathbf{I_N}\alpha^\star = \tilde{\mathbf{I_N}}$ of k equation in $m-1$ unknowns satisfies ${\mathbf{I_N}}^T\mathbf{I}\alpha^\star = {\mathbf{I_N}}^T\tilde{\mathbf{I}}$. If the columns of $\mathbf{I_N}$ are linearly independent, then ${\mathbf{I_N}}^T\tilde{\mathbf{I_N}}$ has an inverse and

$$\alpha^\star = ({\mathbf{I_N}}^T\mathbf{I_N})^{-1}{\mathbf{I_N}}^T\tilde{\mathbf{I_N}}. \tag{14}$$

The projection of $\tilde{\mathbf{I_N}}$ onto the column space is therefore $\hat{\mathbf{I_N}} = \mathbf{I_N}\alpha^\star$. By using equations (6) and (14), we obtain

$$N(j) \cong \bar{N}(j) + \sum_{i=1}^{m-1} \alpha_i^\star n_i(j), \quad (j = 1, \cdots, k) \tag{15}$$

where j is the pixel in the corrupted image and k is the number of pixels in the whole region of the photo image.

We previously made the assumption that the columns of $\mathbf{I_N}$ are linearly independent in equation (12). Otherwise, Equation (14) may not be satisfied. If $\mathbf{I_N}$ has dependent columns, the solution represented by $\alpha^\star$ will not be unique, in which case we will have to choose a particular solution from among the possible ones. The optimal solution of $\hat{\mathbf{I_N}} = \mathbf{I_N}\alpha^\star$ is the one that has minimum length according to equation (7). The optimal solution in this case can be obtained by calculating the pseudoinverse of $\mathbf{I_N}$ [6]. However, in our case, where the goal is to effectively estimate the noise parameters from a corrupted photo image, this is unlikely to happen.

To estimate the noise parameters, the linear coefficients are applied to the sub-matrix of eigenvectors corresponding to the noise parameters. Therefore, the estimated noise parameters can be defined as in equation (16).

$$P = \bar{N}(k+s) + \sum_{i=1}^{m-1} \alpha_i n_i(k+s),\ (s = 1, \cdots, l) \tag{16}$$

4.2 Authentication

Given the test image and the original image, the purpose of photo image authentication is to decide if the test image is identical to the original image. In this work, we first remove the Gaussian noise from the test image. After estimating the noise parameter, P, for the test image, the synthesized image is obtained by applying P to the original image as described by Equations (1) and (2). Then, we used the normalized correlation method for the purpose of authenticating the test image with the synthesized image.

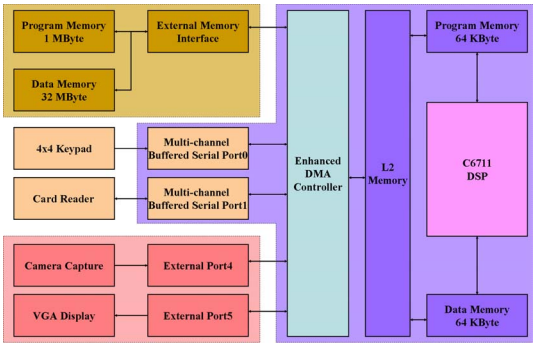


Fig. 4. Block diagram of the system

5 Stand-Alone System Design

Most real-time image processing tasks are time-critical and highly computation-intensive. We have chosen the approach of using a high performance DSP chip, namely, a float-point TMS320C6711. The implementation used C6711 DSK (DSP Starter Kit) and IDK (Image Developer's Kit) from Texas Instrument, Inc.

Because the C6711 DSK platform was developed for an evaluation, it has many limitations in developing a large application. Thus, we re-designed DSK's circuit and combined other devices. The current system is composed of a keypad module, a card reader, a main board, IDK and a host PC shown as Fig. 4. Main board contains two McBSPs (Multi-channel Buffered Serial Port) and EMIF (External Memory InterFace) to communicate with other devices, and supplies 5V power to all the modules. Users use 4x4 keypad for typing their own ID number and selecting modes. In addition to the buttons, keypad module has

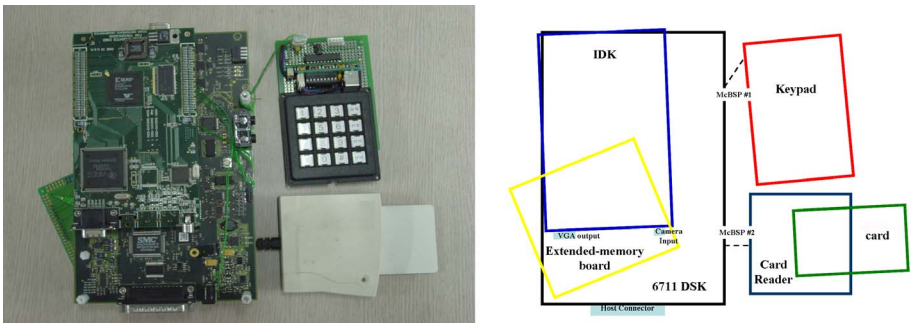


Fig. 5. The implemented face authentication system

RS-232C adapter and converter of baud rate for serial communication. IDK is a daughter board to capture and display images. We used this to monitor the process. Card reader can read user information from each card, and the host PC analyzes and stores the results from C6711 through a RS-232C port. Properly deleting or overwriting data into flash memory of this stand-alone system, we can operate whole functions without the host PC.

6 Experimental Results and Analysis

For testing the proposed method, we used the Korean Face Database(KFDB) introduced in [5] and face data scanned from identification card. We used 100 persons which have one frontal image in KFDB and generated 100 virtual corrupted images. We also tested our algorithm against 137 face images scanned from identification card. The resolution of the images was 320 by 200 pixels and the color images were converted to 8-bit gray level images. Also, these 137 face images were scanned with 300dpi. In the experiment, we performed the following two experiments. First, we compared the difference between the real parameter values and the estimated parameter values against virtual corrupted images from KFDB. We estimated the noise parameters using virtual corrupted face images not including training data. Fig. 6 represents the real parameter values and the

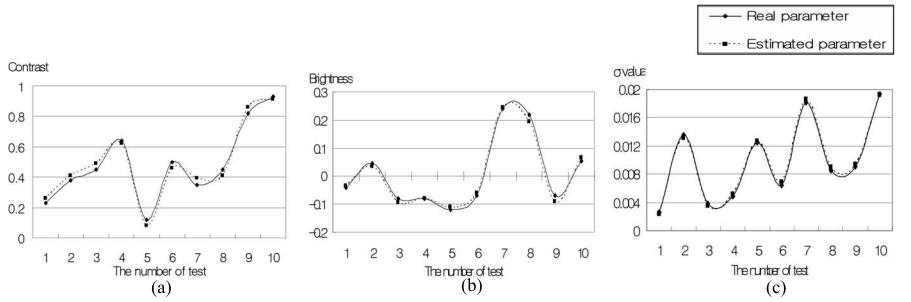


Fig. 6. The comparison of the real parameters and the estimated ones. (a) Contrast parameter. (b) Brightness parameter. (c) Gaussian blur parameter

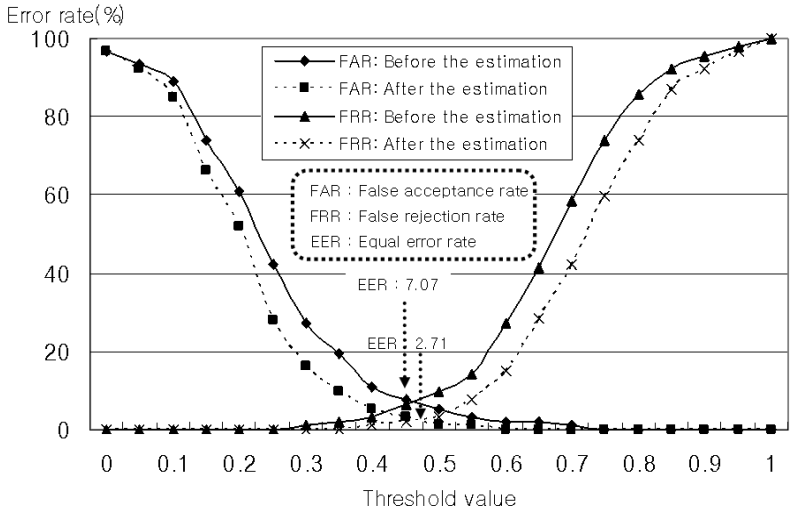


Fig. 7. ROC curve for threshold value used 137 images scanned from identification card

parameter values estimated by the proposed method. The experimental result shows that the proposed method can estimate noise parameters accurately. In this case, however, the estimation error of noise parameters for corrupted face images depends on the number and the value of noise parameter. Therefore, it is very important to use suitable parameters to make corrupted images. Second, we performed an face authentication against the photo images scanned from identification cards under real environments using our prototype system. Fig. 7 shows that the EER is improved from 7.07% down to 2.71%. As a result of this experiment, we showed that the proposed method is more robust for face authentication of corrupted face images than before.

7 Conclusion

Herein, we proposed a method of authenticating corrupted photo images based on noise parameter estimation and implemented an authentication system using TMS320C6711 DSP chip. In contrast to the previous methods, the proposed method deals with the corrupted photo images based on noise parameter estimation and uses only one image per person for training. In this paper, we proved that the estimated parameter values are very close to the real ones. With the images obtained from the KFDB and photo images scanned from identification cards, the proposed method provided for the accurate estimation of the parameters and improved the performance of photo image authentication. For applying the proposed method to real applications, we designed a stand-alone system on DSP chip and implemented our algorithms using this system.

The experimental results on this system show that the noise parameter estimation of the proposed method is quite accurate and that this method is very

useful for authentication, because of its solving the noise problem of corrupted photo images. Also, the proposed method offers good performance in the case of corrupted photo images scanned from identification cards.

Acknowledgement

This research was supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

References

1. Schölkopf, B., Smola,A., Müller, K.: Non-linear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, Vol. 10, No. 5, (1998) 1299-1319
2. Sanderson, C., Bengio, S.: Robust Features for Frontal Face Authentication in Difficult Image Condition. Proc. of Int. Conf. on Audio- and Video-based Biometric Person Authentication, Guildford, UK (2003) 495-504
3. Takahashi, T., Kurita, T.: Robust De-Noising by Kernel PCA. Proc. of Int. Conf. on Artificial Neural Networks, Madrid, Spain (2002) 739-744
4. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience, Vol. 12, No. 1 (1991) 71-86.
5. Hwang, B.-W., Byun, H., Roh, M.-C., Lee, S.-W.: Performance Evaluation of Face Recognition Algorithms on the Asian Face Database, KFDB. Proc. of Int. Conf. on Audio- and Video-based Biometric Person Authentication, Guildford, UK (2003) 557-565
6. Strang, G.: Linear Algebra and Its Applications. Harcourt Brace Jovanovich College Publishers (1988) 442-451
7. Flusser, J., Suk, T.: Degraded Image Analysis : An Invariant Approach. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 6 (1998) 590-603
8. IBM ZISC036 Data Sheet. http://www.ibm.com
9. Yang, F., Paindavoine, M., Abdi, H.: Parallel Implementation on DSPs of a Face Detection Algorithm. Proc. of International Conference on the Software Process, Chicago, USA (1998)
10. Gilbert, J., Yang, W.: A Real-Time Face Recognition System Using Custom VLSI Hardware. Proc. of Computer Architectures for Machine Perceptron Workshop, New Orleans, USA (1993) 58-66

Evaluation of 3D Face Recognition Using Registration and PCA

Theodoros Papatheodorou and Daniel Rueckert

Visual Information Processing, Department of Computing, Imperial College
180 Queen's Gate, London SW7 2BZ, UK
{tp500,dr}@doc.ic.ac.uk

Abstract. We introduce a novel technique for face recognition by using 3D face data that has been reconstructed from a stereo camera system. The face data consists of a dense 3D mesh of vertices describing the facial shape and geometry as well as a 2D texture map describing the facial appearance of each subject. We propose a recognition algorithm based on two steps: The first step involves the cleaning of the facial surface followed by a registration and normalization to a standard template face. The second step involves the creation of a PCA model and the use of a reduced dimensionality face space for the calculation of facial similarity. We use this technique on 3D surface and texture data comprising 83 subjects. Our results demonstrate the wealth of 3D information on the face as well as the importance of standardization and noise elimination in the datasets.

1 Introduction

The face is controlled by a set of 19 muscles, which are responsible for a range of facial expressions. The face is a constant source of information. Emotions as well as identity are conveyed to our social environment. Analysing this information correctly is so important that face and expression recognition seem to be a dedicated process in the brain [1, 2].

Face recognition research spans several disciplines such as computer vision, pattern recognition, machine learning and psychology. Face recognition encompasses law enforcement as well as several commercial applications [3, 4]. Crowd surveillance, electronic line-up, store security and mug shot matching are some of the security applications. Besides this, it could also assist in computerized aging simulations as in [4] where shape and texture normalized 3D-faces were judged to be more attractive and younger than the original faces. Furthermore it could assist in the reconstruction of partially damaged face images as in [5] where PCA analysis of a face database enabled researchers to fill in the information in partially occluded faces. Research in categorizing gender from biological motion of faces as in [6] could also benefit from face recognition algorithms.

Traditional face recognition relies on two-dimensional photographs. However, 2D face recognition systems tend to give a high standard of recognition only when images are of good quality and the acquisition process can be tightly controlled. Popular techniques such as Eigenfaces [7] are very sensitive to image noise. As a result variations in illumination and head position tend to affect the recognition process negatively.

3D face recognition has several advantages over traditional 2D face recognition: First, 3D face data provides absolute geometrical information about a face's shape

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 997–1009, 2005.
© Springer-Verlag Berlin Heidelberg 2005

and size. Additionally, face recognition using 3D data is more robust to angle and posture changes since the model can be rotated to any arbitrary position. Moreover, 3D face recognition can be less sensitive to illumination since it does not solely depend on pixel brightness for calculating facial similarity. Finally, it provides automatic face segmentation since the background is typically not synthesized in the reconstruction process. On the other hand, 3D data capture, is typically more complex than 2D data acquisition and currently requires more cooperation from the subject. Furthermore, the capturing equipment and process remains significantly more expensive and the acquisition process is not as straightforward as in 2D. However, in recent years the availability of affordable and easy-to-use 3D camera systems and laser scanners has increased significantly. Also, alternative ways of synthesis of 3D faces from a 2D input have recently emerged [8] and are becoming more and more popular.

Relatively little research has focused on combining 3D and 2D information. Recognition systems rely on 2D intensity images as in [7] or 3D range data [9, 10, 11]. In [9] facial surfaces are acquired using structured light technology. In order to simplify and thus speed up the comparison the 3D model is reduced to a 2D profile contour. The similarity between faces is measured by measuring the differences in curvature between two facial contours. In [10] textured 3D meshes are represented as 2D images that incorporate the 3D geometry. PCA is applied in the 2D images and the technique is robust to facial expressions. In [11] a method is proposed in which 3D information from profiles is used in linear combination with grey level cues. Tsalakanidou et. al. [12] used a multimodal PCA technique in which they got very high recognition rates (99%) using 40 subjects. [13] also used PCA analysis for doing recognition in one of the biggest studies using 3D data taken over a long period of time. They managed to achieve nearly perfect rank one recognition for 3D+2D recognition and 94% when using only 3D. A recent paper by Bronstein et al [14] presented a novel 3D face recognition approach. Their key idea was to represent the facial surface in a way that it is not affected by isometric deformations such as expressions or postures of the face. This is based the multi-dimensional scaling technique presented in [15, 16] to flatten the 3D surface while maintaining the geodesic distances between points intact.

In this paper we propose and evaluate a PCA-based approach using 3D face data which has been normalized using 3D registration techniques. We pre-process the data to eliminate, as much as possible, noise and artifacts from the datasets by adopting a spherical B-spline surface representation. Using data collected from 83 subjects we show that recognition rates of up to 100% and verification rates of up to 98% can be achieved.

2 Data Acquisition

We use a commercial stereo camera system for our 3D data acquisition. The Vision RT VRT3D stereo camera system is made up of three video cameras and a speckled pattern projector. The output is an accurate 3D surface face model. Figure 1 shows the camera system and the three images captured simultaneously from the cameras.

We have chosen this technology over laser scanners because of its speed of acquisition (up to 30 frames/sec. can be captured) and the speed of the data reconstruction (< 5 sec. on a 1GHz machine) thus allowing near real time processing in realistic scenarios. The speed of the data acquisition also prevents motion artifacts from being

introduced in the 3D acquisition process. Finally, the system is built in a cost effective fashion thus greatly reducing hardware costs compared to other capturing techniques. The accuracy of the system is high with a RMS error of under 1mm for a typical face acquisition. Studies on the camera system have been validated in clinical settings [17]. The drawbacks of the acquisition system are its limited depth of field since the faces need to be placed within a specific distance from the camera making data acquisition more intrusive than 2D video imaging. A similar problem arises when using laser range scanners.

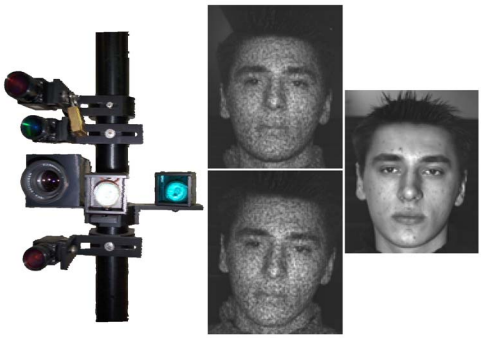


Fig. 1. The VRT3D camera system and its output files

3 Face Cleaning and Registration

PCA is sensitive to noise and requires relatively uniform data. In [7] as well as other papers researchers scale the 2D face data in order to maximize the correspondence between the faces. The section below describes a method of semi-automatic cleaning using landmarks and then regularizing the face by deforming a regular surface.

After reconstruction the 3D faces are landmarked. We selected 17 landmarks on the 3D faces of each subject as shown on the left in figure 2. Using these we generate a convex hull in the 2D texture image which is used during the cleaning process (figure 2, right).

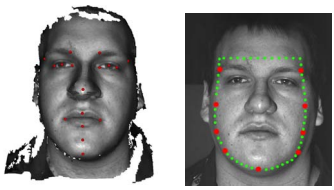


Fig. 2. A landmarked face before cleaning

A common problem using 3D faces is that the surface data may also contain portions of the hair, neck and shoulders. To avoid including these areas during the model building and recognition stages we only keep those vertices whose texture coordinates lie within the convex hull of the landmarks in the 2D texture map. Figure 3 shows the face data before and after cleaning. The non-textured area below indicates the area that has been discarded.

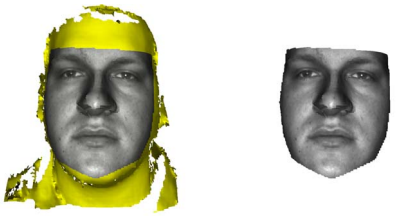


Fig. 3. A face before and after cleaning

After the faces have been cleaned they are rigidly registered to a template face. We are using rigid rather than affine registration to maintain size as an important feature for facial similarity. The rigid surface registration is carried using a well-known registration algorithm, the Iterative Closest Point algorithm (ICP) as proposed in [18] which aligns arbitrary 3D shapes. We deal with the alignment problem like an optimisation problem where we try to minimize the Euclidean distance between the template face and each dataset. ICP iterates through each point in the source or moving mesh (**S**) and finds the closest point in the template mesh (**T**). The distance we try to minimize between an individual 3-dimensional data point **s** with coordinates (s_x, s_y, s_z) and the closest 3-dimensional point on template mesh **T** with coordinates (t_x, t_y, t_z) can be denoted as:

$$d_{3D}(\mathbf{s},\mathbf{T}) = \min_{\mathbf{t}\in\mathbf{T}} [\sqrt{(s_x - t_x)^2 + (s_y - t_y)^2 + (s_z - t_z)^2}] \quad (1)$$

Figure 4 shows the colour map of the residual 3D distance between the two faces before and after registration. Blue colours indicate small distances while red colours indicate larger distances between the faces.

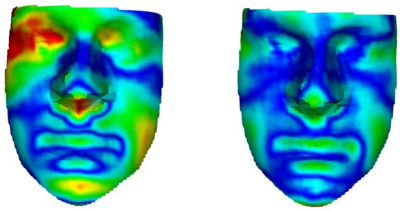


Fig. 4. Distance map on the face, before and after registration

4 Face Modelling Using B-Splines

In order to normalize the faces we have chosen to use a standard surface with a fixed number of points that we deform using B-splines as described in [19] to approximate each dataset. Instead of deforming a plane we have chosen to deform a sphere using a spherical free-form (spline) transformation. The advantage of deforming a closed surface such as a sphere is that we achieve an easy parameterisation of the data since every point on the surface can be described using two angles, φ and θ. We fit a spherical free-form transformation to approximate the facial surface as presented in [19] and shown in figure 5.

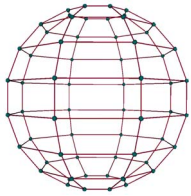


Fig. 5. The spherical B-spline control point grid

In order to fit our model to the data we use a multi-resolution approach: After approximating the data using an initial spherical control point grid, we subdivide the control point grid and we approximate again achieving greater accuracy by allowing more control points to capture a greater degree of detail. The process is repeated until the face is approximated well enough. Figure 6 below are the images of the sphere as it is deformed to approximate the dataset and figure 7 shows the high level of detail captured by the approximation.

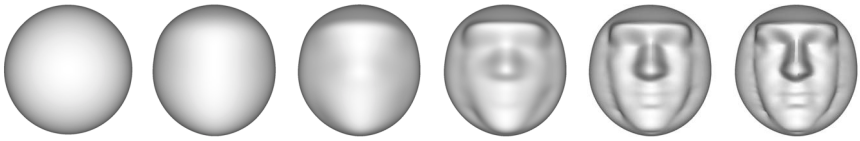


Fig. 6. Sphere being deformed

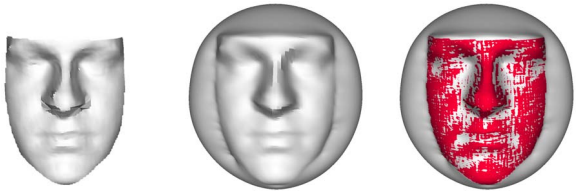


Fig. 7. Original face (left), approximated face (middle), overlap of two (right)

Apart from standardizing the faces, the B-spline approximation of the faces allows for the correction of artifacts of the surface. Holes in the face that arise from the limitations of our data capturing hardware are corrected automatically as the sphere is deformed to approximate the hole's surrounding points. The dataset in Figure 8 below shows artifacts in the areas of high curvature where occlusion might occur and in areas of low reflectance such as the eyes. After deforming the sphere these artifacts disappear as the surface on the right shows.

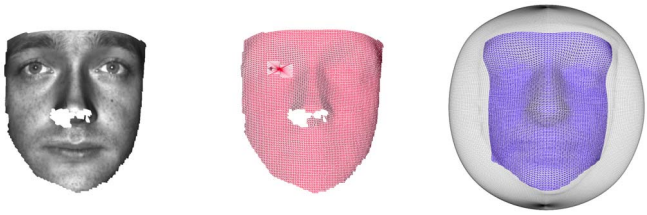


Fig. 8. Example of error correction

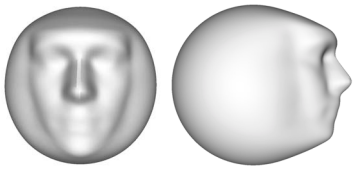


Fig. 9. The mean face after B-spline approximation

After all the faces of the database have been approximated we calculate the mean face shown in Figure 9.

Since most points in our model are not part of the face surface it would be desirable to eliminate these points. More importantly the border between the area of the sphere that is approximating the face and the sphere that is not deformed has the potential to negatively affect the recognition process. Figure 10 shows the two approximated faces of the same subject. We can see that in the border there is variation that could affect the statistical methods we will later employ.

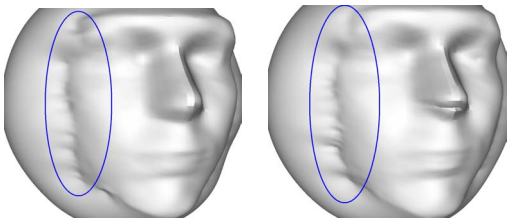


Fig. 10. The noise in the face border

Figure 11 shows a top view of the same face and one can see one face being more protruded in the borders than the other.

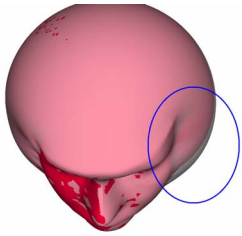


Fig. 11. Two sphered faces of the same person and the discrepancy between them

In order to eliminate this variation we chose to select a smaller area that is localized on the area of the sphere that is deformed to approximate the facial data. Since we need the same number of points for PCA we cannot use every face's landmarks to find the area on the sphere that is part of the face. Since the faces are rigidly registered before the B-spline approximation they occupy approximately the same are on the sphere. We therefore landmark the mean approximated face and for every face we keep only the points that lie within the landmarked area of the mean face. The landmarks are placed well within the border area on the mean face so that the border

points are not included in the final dataset for most faces. The datasets in Figure 12 below are examples of the datasets (7000 points) after they have been fully processed. They all contain the same number of points and they contain no surface artefacts or extreme irregularities that would have adverse effects on the recognition effort.

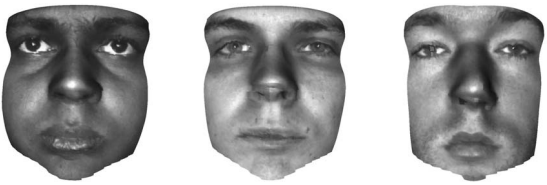


Fig. 12. Example datasets after preprecessing

Figure 13 shows the processing pipeline described in detailed above and that every face in the database goes through.

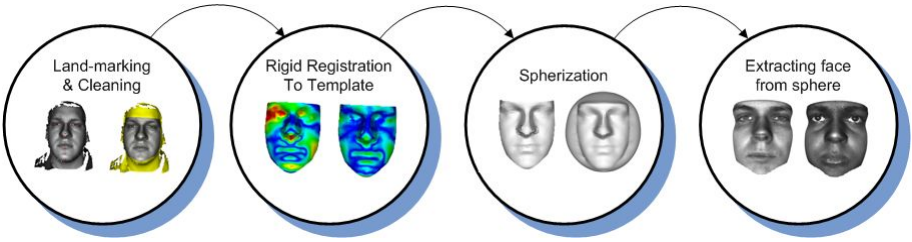


Fig. 13. The processing pipeline

5 Building the PCA Model

The correspondence problem is one of the key problems in model building. Given several datasets and knowing the correspondences between the data in essential for statistical analysis of the population. [7] show that when correlation between images drops, such as when images are of different size or orientation, the recognition rates fall as well. There are a number of different approaches for automatically establishing correspondences between shapes: In [20] an optical flow algorithm is used while [21] uses a blend of ICP and Active Shape Models to establish correspondences between points on two surfaces. In our case we do not explicitly establish correspondences. Instead correspondences are established implicitly by the registration and surface modeling steps.

Faces from a population of subjects have certain similarities, both in terms of shape and texture. Based on this idea we can subject them to a dimensionality reduction and describe any face in the population by a set of principal components or eigenvectors. One of these techniques that we can use for this is the Principal Component Analysis (PCA). PCA is a transform that chooses a new coordinate system for the data in which the greatest variance lies on the first axis, the second greatest in the second axis etc. We can therefore reduce the dimensionality by retaining those characteristics that account for most of the variance in the dataset. We can then describe any dataset as a collection of weighten eigenvectors added to the mean as shown below.

$$\mathbf{x} \approx \overline{\mathbf{x}} + \Phi\mathbf{b} \tag{2}$$

Where $\mathbf{\Phi} = (\phi_1 \mid \phi_2 \mid \ldots \mid \phi_t)$ and $\mathbf{b}$ is a weights vector that has as many dimensions as the eigenvalues. PCA generates n-1 eigenvectors where n is the size of the population. Having now regularized datasets with about 5000 points each we can run PCA on the shape information. Figure 14 shows the first four principal modes for the shape of the population. The first 16 modes are enough to describe 90% of shape variability.

The weight for each mode i is $\pm 3\sqrt{\lambda_i}$ where λ_i is the eigenvalue of the eigenvector ϕ_i.

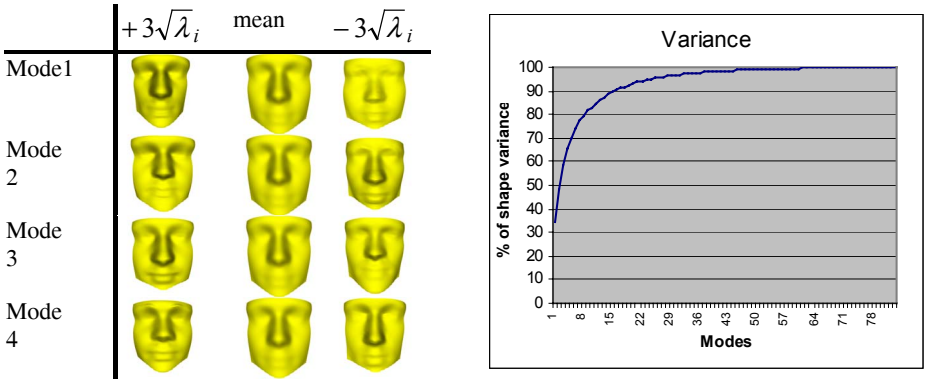


Fig. 14. First four principal modes of shape

We also run PCA just on the texture of the faces and we build texture face space. The textural information is not pre-processed and we expect it to contain illumination variations due to slight differences in pose especially since our capture system does not emit ambient lighting. Figure 15 shows the first four modes of the texture variation and the texture variance explained by various eigenmodes. We can see that the first mode shows a very light and a very dark face. This is because of largely due to illumination differences in capturing the data as well as the race differences in the population. Our population was multicultural and we had variation in skin colour from subject to subject.

Finally we joined the data vectors of texture and shape together before performing. As the shape and texture units are different we had to use a weight to scale the texture, otherwise small differences in the shape which tend to have a smaller scale are going to be overshadowed by the effect of texture. To calculate the optimal weight we calculated the ratio between the texture and shape variability as $\mathbf{W}=r\mathbf{I}$ where r^2 is the ration of the total intensity variation to the total shape variation. Figure 16 below shows the first four eigenmodes of the combined shape and texture model.

6 Results

After the creation of the face space each face in the database is described by a series of 83 parameters. Each query face is then projected in the 83-dimensional face space. The closest database face in this space is considered the closest match. We measure

the Euclidean or the Mahalanobis distance in this space. The Mahalanobis distance takes into account the correlations of the dataset therefore producing slightly different results. In order to assess how good the face space is suited for face recognition, both the recognition and verification rate are calculated. Given a query dataset, the recognition rate measures the number of correct first matches. In the case of verification we need to establish a threshold and given a similarity between the closest match from the database make a decision whether or not the person is who he claims to be. A term often associated with this measure is the False Acceptance rate (FA) which is the number of "impostors" that have been incorrectly labeled as clients. Another term is the False Rejection rate (FR) which is the number of times a client is incorrectly labeled as an "impostor" and is undeservedly barred. In our experiments we set the threshold so that FA=0, in other words there are no impostors that are incorrectly accepted into the system. We then measure FR or how many clients are incorrectly left out.

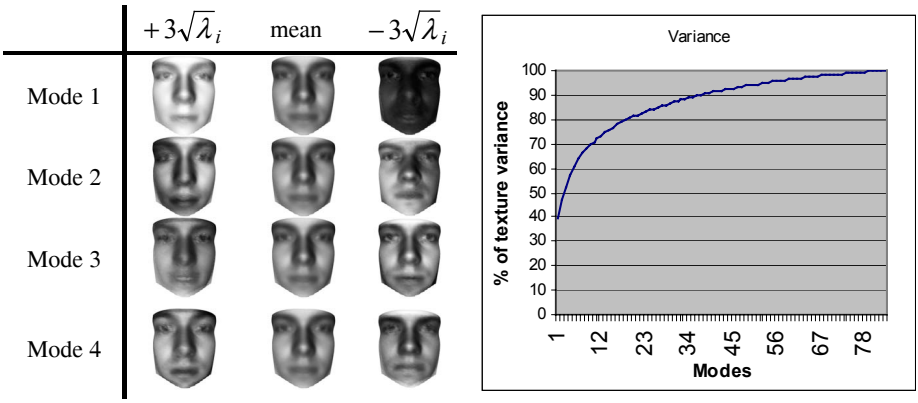


Fig. 15. First four principal modes of texture

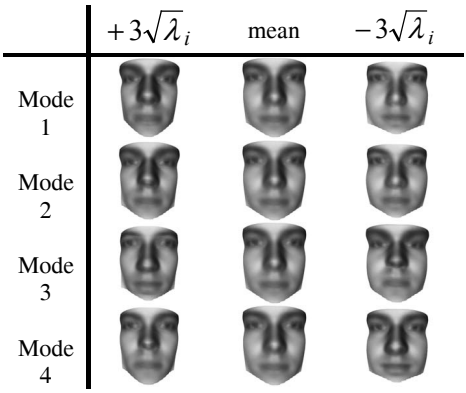


Fig. 16. First four principal modes for texture and shape

Figure 17 shows the recognition results of experiments using just the shape of the datasets and the Euclidean and Mahalanobis distance to measure facial similarity as

well as the verification results using the same measurements. With just 10 eigenvalues the recognition rate goes to 100% for the Euclidean distance and up to 97% using the Mahalanobis distance. The verification, being a stricter criterion goes up to 98% using 15 eigenvalues in Euclidean and 30 eigenvalues for the Mahalanobis. The verification threshold is given a value so that there are no false acceptances (FA=0).

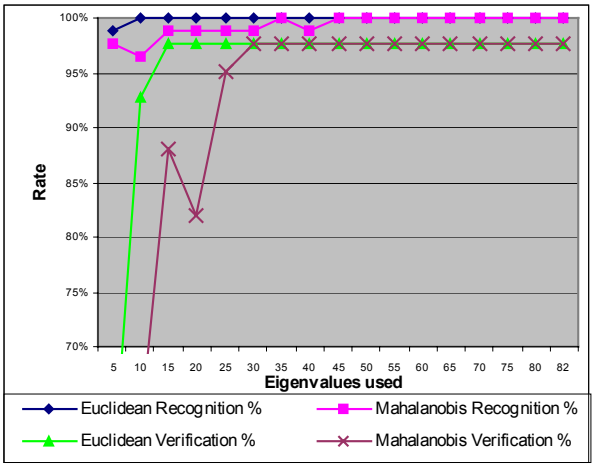


Fig. 17. The Recognition and Verification rates using shape only

The same testing criteria are used for recognition and verification using texture only. Figure 18 shows the recognition results of experiments using just the texture of the datasets and the Euclidean and Mahalanobis distance to measure facial similarity as well as the percentage of Correct Rejections given that there are no False Acceptances using the same measurements.

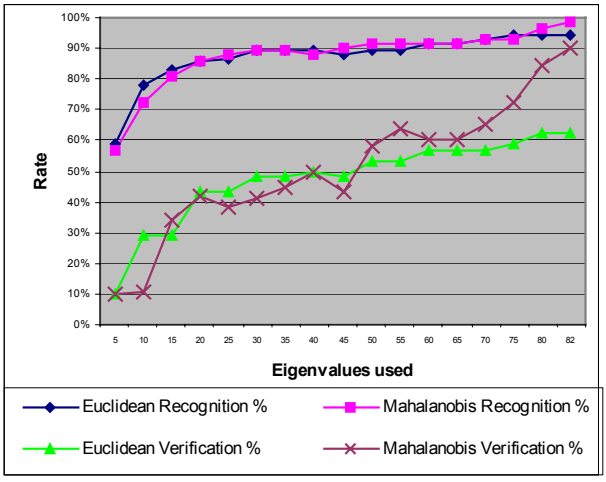


Fig. 18. The recognition and verification rates using texture only

Using just 30 eigenvalues the texture recognition rate goes up to almost 90% with both Euclidean and Mahalanobis distance. The verification rates are significantly poorer with 90% reached using all eigenvalues with Mahalanobis distance.

7 Discussion

The results clearly demonstrate that there is enough discriminatory power in 3D shape information to be used as a classifier. Texture has also proven to be quite powerful although not as powerful as the shape. Not only does the verification rate for texture go up to 90% compare to almost 98% for shape but also the shape rate reach above 90% levels just by using ten eigenvalues to describe a dataset while texture verification with the same number of eigenvalues is at a low 11%. The texture as a predictor gets at acceptably high levels only when all the eigenvalues are used. This is expected as we can see in the variance graphs in the previous sections. The variance graph for shape is "pushed" much closer to the upper left corner than the variance graph for texture. In order words we need more eigenvalues in order to describe a certain amount of variability in texture than we would need in shape.

It is important to note that combining the two models did improve the verification rates when less eigenvalues were used but not when all were used. This is because the two cases that we fail to correctly verify the identity of a person in the shape experiments are the same cases that fail in the texture ones. Therefore, combining the two would not improve the recognition results. Upon close inspection of those two examples we note that they are slightly misregistered to each other. Misregistrations lead to bad correspondence between points. The two faces that fail in verification in the shape experiment are the closest matches in the recognition test but the distance from each other is too great for them to pass the verification test. A possible improvement would come if one uses landmarks for registration between the faces as suggested in [22].

Something that could be done in order to improve the texture results and decrease the variability in the dataset, thus leading to better models is to pre-process the 2D data of the faces with some filters trying to get rid of some of the illumination effects as they do in [23] where they normalize the texture.

One of the reasons we extract the faces from the sphere rather than leaving the sphere is in order to get rid of some of the noise where the face meets the sphere. In experiments that we have run on the spheres rather than the extracted areas the results have not been as good. Figure 19 shows the rates. The explanation is two-fold. The area where the face merges with the sphere is an area that is contains a lot of noise and it does influence the results. Secondly, since the faces are cleaned manually, it is natural that different faces contain more information than others. By extracting the same area from the spheres and making sure that it is smaller than the average face, we include the number of actual "facial" points in the datasets. Figure 19 shows the recognition and verification results using the sphered faces rather than the extracted areas.

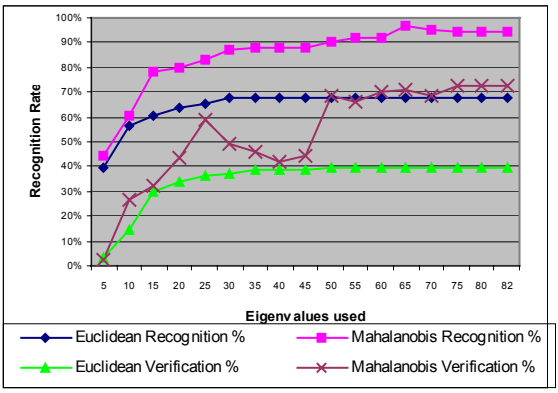


Fig. 19. The recognition and verification rates using sphered faces without extraction

References

[1] G.C. Baylis, E.T. Rollis, C.M. Leonard, Selectivity between faces in the response of a population of neurons in the cortex in the superior suculus of the monkey, Brain Res. 342 (1985) 91-102

[2] J.V. Haxby, E.A. Hoffman, M.I. Gobbini, The distributed human neural system for face perception, Trends Cognitive Science 4 (2000) 223-233

[3] R. Chellapa, C.L. Wilson and S. Sirohey, Human and Machine recognition of faces: A survey, Proceedings of the IEEE, Vol. 83, No. 5, (1995)

[4] A.J. O' Toole, T. Price, T. Vetter, J.C. Bartlett, V. Blanz, 3D shape and 2D surface textures of human faces: the role of averages in attractiveness and age, Image and Vision Computing 18 (1999) 9-19

[5] B. Hwang, S. Lee, Reconstruction of Partially Damaged Face Images Based on a Morphable Face Model, IEEE Transcactions on pattern analysis and machine intelligence,Vol 24, No. 3, (2003), 365-372

[6] H. Hill, A. Johnston, Categorizing sex and identity from the biological motion of faces, Current Biology, (2001), 11:880-885

[7] M.A. Turk, A.P. Pentland, Face Recognition Using Eigenfaces, Journal of Cognitive Neuroscience, 3(1), (1991).

[8] V. Blanz, T. Vetter, A Morphable Model for the synthesis of 3D faces, Computer Graphics Proceedings, Siggraph (1999), Los Angeles, 187-194

[9] C. Beumier, M. Acheroy, Automatic 3D face authentication, Image and Vision Computing 18 (2000) 315-321

[10] A.M. Bronstein, M.M.Bronstein, and R. Kimmel, Expression-Invariant 3D Face Recognition, Proceedings of 4th International Conference, AVBPA (2003), Guildford, UK, 62-69

[11] C. Beumier, M. Acheroy, Face verification from 3D and grey level clues, Pattern Recognition Letters 22, (2001), 1321-1329

[12] F. Tsalakanidou, D. Tzocaras, M. Strintzis, Use of depth and colour eigenfaces for face recognition, Pattern Recognition Letters, (2003), 24:1427-1435.

[13] K. Chang, K. Bowyer, P. Flynn, Face recognition using 2D and 3D facial data, 2003 Multimodal User Authentication Workshop, (2003), 25-32.

[14] A. Bronstein, M. Bronstein and R. Kimmel, Expression-invariant 3D face recognition, Proc. Audio & Video-based Biometric Person Authentication (AVBPA), Lecture Notes in Comp. Science 2688, Springer, (2003), 62-69

[15] G. Zigelman, R. Kimmel, N. Kiryati, Texture mapping using surface flattening via multi-dimensional scaling, IEEE Transactions in Visualization and Computer Graphics, vol. 8, (2002), 198-207.
[16] R. Grossman, N. Kiryati, R. Kimmel, Computational surface flattening: a voxel-based approach, IEEE Transactions PAMI, vol. 24, (2002), 433-441.
[17] N. Smith, I. Meir, G. Hale, R. Howe, L. Johnson, P. Edwards, D. Hawkes, M. Bidmead, D. Landau, Real-Time 3D Surface Imaging for Patient Positioning in Radiotherapy, awaiting publication in the ASTRO 2003 proceedings
[18] P.J. Besl, N.D. McKay, A Method for Registration of 3-D Shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 14, No 2, (1992), 239-256
[19] S. Lee, G. Wolberg, S.Y. Shin, Scattered data interpolation with multilevel b-splines, IEEE Transactions on Visualization and Computer Graphics, vol. 3, no. 3, (1997).
[20] V. Blanz, T. Vetter, Face recognition based on fitting a 3D Morphable Model, IEEE transactions on pattern analysis and machine intelligence, vol. 25, no.9, (2003)
[21] R. Szeliski, S. Lavallee, ``Matching 3D Anatomical Surfaces with Non-Rigid Deformations using Octree-Splines", IEEE Workshop on Biomed. Image Anal., June, (1994), 144-153
[22] K. Chang, K. Bowyer, P. Flynn, Face recognition using 2D and 3D facial data, 2003 Multimodal User Authentication Workshop, (2003), 25-32.
[23] N.P. Costen, T.F. Cootes, G.J. Edwards, C.J. Taylor, Automatic extraction of the face identity-subspace, Image and Vision Computing, vol. 20, (2002), 319-329.

Dynamic Approach for Face Recognition Using Digital Image Skin Correlation

Satprem Pamudurthy[1], E Guan[2], Klaus Mueller[1], and Miriam Rafailovich[2]

[1] Department of Computer Science, State University of New York at Stony Brook
{satprem,mueller}@cs.sunysb.edu
http://www.cs.sunysb.edu/~vislab
[2] Department of Material Science and Engineering
State University of New York at Stony Brook
eguan@ic.sunysb.edu, mrafailovich@notes.cc.sunysb.edu

Abstract. With the recent emphasis on homeland security, there is an increased interest in accurate and non-invasive techniques for face recognition. Most of the current techniques perform a structural analysis of facial features from still images. Recently, video-based techniques have also been developed but they suffer from low image-quality. In this paper, we propose a new method for face recognition, called Digital Image Skin Correlation (DISC), which is based on dynamic instead of static facial features. DISC tracks the motion of skin pores on the face during a facial expression and obtains a vector field that characterizes the deformation of the face. Since it is almost impossible to imitate another person's facial expressions these deformation fields are bound to be unique to an individual. To test the performance of our method in face recognition scenarios, we have conducted experiments where we presented individuals wearing heavy make-up as disguise to our DISC matching framework. The results show superior face recognition performance when compared to the popular PCA+ LDA method, which is based on still images.

1 Introduction

In this paper, we propose a face recognition method that is based on a feature tracking method we call *Digital Image Skin Correlation* (DISC). Unlike other feature tracking methods that require a set of sparse markers to be attached to the actor's face or have a sparse set of feature points, DISC uses the natural texture of the skin for tracking the facial motions. The distribution of skin pores on the face provides a highly dense texture, which we exploit for capturing fine-scale facial movements. By tracking an individual's skin pores during a facial expression, such as a subtle smile, we can generate a vector field that describes the individual's facial movement. We refer to these vector fields as facial deformation fields. Since facial movements are comprised of a complex sequence of muscle actions, which are unique to an individual, it is almost impossible to imitate an individual's facial expressions. This suggests that we can use the facial deformation fields for face recognition.

Face recognition is the most popular and successful application of image analysis and understanding. It has wide-ranging applications in a number of areas such as security, surveillance, virtual reality and human-computer interaction [16]. Face recognition has received renewed interest in recent times, which can be attributed to increased concerns about security, and the rapid developments in enabling technologies. Most of the existing methods for face recognition are based on still images and

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1010–1018, 2005.
© Springer-Verlag Berlin Heidelberg 2005

video. Image-based techniques are mainly interested in the shape, size and position of features such as eyes, nose and mouth. But the face of an individual is unique in other respects as well, such as in the way it moves during facial expressions. Video-based methods use both the temporal and spatial information for face recognition. But they suffer from low image quality and limited resolution. On the other hand, while high-resolution images are available to still-image-based methods, they only contain spatial information. Our method leverages the strengths of both still-image and video-based methods, using high-resolution images to extract temporal information.

An important feature of our method is that it is completely based on affordable mainstream technologies. The only hardware we require is a high-resolution digital camera. In our experiments, we have observed that a digital camera with resolution greater than 4 mega pixels is sufficient to capture images that enable us to accurately track the skin pores on an individual's face [2] (see Figure 1 below). Nowadays, cameras with such resolutions are commonplace. Some newer cell phone cameras have resolutions as high as 5 mega pixels. The rapid explosion in digital camera technology means that images can now be acquired at hitherto unheard of detail. Our method exploits this detail to accurately track an individual's facial motions.

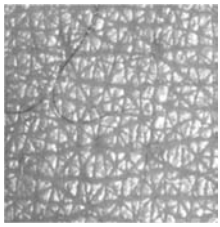


Fig. 1. Skin texture 8x8mm. The *skin pores* are clearly visible

Face and facial expressions are the most visually distinguishing features of an individual and they cannot be easily, if at all, imitated. As such, an accurate and efficient feature tracking method such as DISC has great potential to be successfully applied to face recognition, and this paper presents some of the results that we obtained for face recognition with our method and compares them with still-image-based recognition.

The rest of the paper is organized as follows. Section 3 describes DISC. Section 4 describes our approach. Section 5 presents the results for our method and compares them with some standard algorithms. Section 6 concludes the paper.

2 Previous Work

Most of the early still-image-based methods were based on PCA. Eigenfaces [3] was one of the first successful face recognition methods. It can handle variations and blur in the images quite well, but it requires a large number of training images. Moghaddam et. al. [5] improved on the eigenface method by using a Bayesian probabilistic measure of similarity instead of Euclidean distance. PCA+LDA [4] is another popular method. It attempts to produce linear transformations that emphasize the difference between classes while reducing the difference within classes. Elastic Bunch Graph Matching [6] locates landmarks on faces and extracts Gabor jets from each landmark. These jets are used to form a face graph, which is used to compare for similarity. In

general, appearance-based methods use multiple images per subject to deal with pose and illumination variations. Zhang et. al. [7] proposed a face recognition method using Harmonic Image Exemplars that is robust to lighting changes and requires only one training image per subject. They later extended this method to be pose-invariant by using Morphable Models to capture the 3D shape variations [8]. Some researchers have proposed 3D face matching techniques. Jain et. al. [9] proposed a face surface matching method that takes into account both the rigid and non-rigid variations. A number of other still-image methods based on neural networks, support vector machines and genetic algorithms have also been proposed for face recognition [16].

Earlier video-based face recognition methods used still-image methods for recognition after detecting and segmenting the face from the video. Wechsler et. al. [10] employed RBF networks for face recognition. McKenna et. al. [11] developed a method using PCA. Instead of using a single frame for recognition, they implemented a voting scheme based on the results from each frame in the video. A disadvantage is that that the voting scheme results in increased computational cost. Choudhury et. al. [12] use both face and voice for person recognition. They use Shape from Motion to compute the 3D information for the individual's head so as to differentiate between a real person and an image of that person. Li and Chellappa's method [13] is perhaps the closest to our approach. They use Gabor attributes to define a set of feature points on a regular 2D grid. These feature points are tracked to obtain trajectories of the individual in the video, which are then employed to identify that individual. In their experiments, this method performed better than frame-to-frame and voting-based methods, even when there were large lighting changes. This method can track features in low-resolution videos, while our method can more accurately track features in high-resolution images. As high-resolution video capture devices become available in the future, our method may be further extended to be video-based.

3 DISC

Digital Image Skin Correlation (DISC) is based on a material science technique called Digital Image Speckle Correlation. DISC analyzes two images taken before and after the skin has deformed and yields a vector field corresponding to the deformation of the skin. It derives the displacement vectors by tracking skin pores in the digital images before and after deformation. As shown in Figure 1, human skin provides abundant features i.e. skin pores, that can serve as tracking markers.

As shown in Figure 2, two images of a face are taken, one before and one after deformation, producing a reference image and a deformed image. DISC divides these two images into smaller windows and compares each window in the reference image with it's neighboring windows in the deformed image. The windows are matched by computing the normalized cross-correlation coefficient between them, and the window with the highest coefficient is considered to be the corresponding window. The correlation is computed as,

$$S\left(x, y, u, v, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y}\right) = 1 - \frac{\sum I(x,y) * I^*(x^*, y^*)}{\sqrt{\sum I(x,y)^2 * \sum I^*(x^*, y^*)^2}}, \quad (1)$$

where,

$$x^* = x + u,$$
$$y^* = y + v.$$

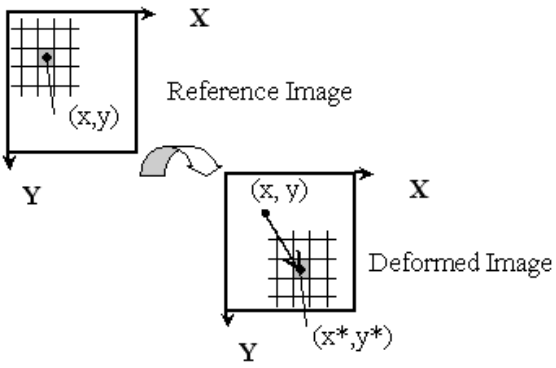


Fig. 2. Schematic of DISC

and $I(x,y)$ and $I^*(x^*,y^*)$ are the intensities within the window. The normalized cross-correlation function is computed over the entire window. The coordinate difference between the matched window pair gives the average displacement vector at the center of the reference window. More details on the tracking are available in [1,2].

DISC shows that the distribution of skin pores on the face provides a natural and highly dense texture for tracking. It eliminates the need for cumbersome markers to be attached to an individual's face, which will not provide tracking information at the skin's high pore-resolution in any case.

4 Our Approach

In order to recognize an individual, we capture two high-resolution images of the subject's face – one with a neutral expression and the other with a subtle smile. Following, we use DISC to compute the deformation field that represents the way the subject's face deforms. Then, instead of directly using the deformation field as the feature for recognition, we generate two scalar fields from each vector field, which are then used in the recognition process. We call these scalar fields *projection images* since they are obtained by projecting each vector in the deformation field onto the X- and Y-axes, respectively. Alternatively, we could also separate the deformation field into magnitude and orientation fields, and possibly use their gradients as features. We have employed the projection images for the experiments presented in this paper since they have yielded satisfactory results.

We obtain two projection images from each vector field, one corresponding to projections onto the X-axis and the other onto the Y-axis. We could use either of them for recognition or use both and weigh the results. Figure 3 below shows some typical projection images, in this case images obtained from identical twins. We observe strong differences in the projection images (right), while the still images (left) are very similar.

In the recognition procedure, we compare the projection image of the candidate subject with the projection images stored in our database. However, before we can compare the projection images of two individuals, we need to deal with the variations in geometry. Even if the images are of the same view, we might have to scale, rotate or shift them in order to align the images. The first step would be to align the eyes and

then adjust the aspect ration so as to align the mouth. But then other facial features may still be misaligned. The solution to dealing with different facial geometries is to warp both projection images to an average geometry. Before we can align or warp the projection images, we need to find the fiducial points for registration. Though we use the information derived from the deformation fields for classification, we do not disregard the still images entirely. Instead, we use them to identify the fiducial points such as the corners of the eye, mouth and nose bridge. We then use this information to warp the projection images.

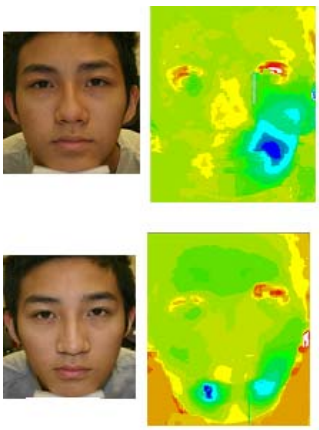


Fig. 3. Two *projection images* (right) of *identical twins* (left)

We use a nearest-neighbor classifier for recognizing the subject's face. For each projection image in the database, we align it with the subject's projection image and compute the similarity between them. The measure of similarity we use is the normalized cross-correlation coefficient. Normalized cross-correlation computes the similarity between two functions, which in our case are the projection images. If *I(x,y)* and *I*(x,y)* are the two projection images under consideration, where *(x,y)* are the image coordinates, the normalized cross-correlation between these projection images is given as:

$$S = 1 - \frac{\sum I(x,y) * I^{*}(x\ ,y\ )}{\sqrt{\sum I(x,y)^{2} * \sum I^{*}(x\ ,y\ )^{2}}}, \tag{2}$$

The distance between the projection images is computed as

$$d = 1 - s\,. \tag{3}$$

The subject is recognized by finding the projection image in the database that is closest to that of the subject.

5 Experiments

In order to validate the effectiveness of our framework and test its robustness, we compared the ability of our approach with that of still-image-based approaches in the task of recognizing people with heavy make-up. For this purpose, we have created

two small databases, since the standard face recognition databases do not contain the kind of images and tracking data we require. The first of these databases consists of ten subjects, with at least three still images per subject. The second database consists of the same ten subjects, but here we store only two images, one with a neutral expression and the other with a subtle smile, as well as the corresponding projection images. The first database is used for the still-image methods, while the second database is employed for our approach. In our experiment, we used the projection images corresponding to the X-axis. We used a Canon EOS 300D/Digital Rebel camera to capture the still-images. This camera can capture images with a maximum resolution of 6.3 mega pixels. Figure 4 shows some of the subjects in our database and the corresponding X-projection images.

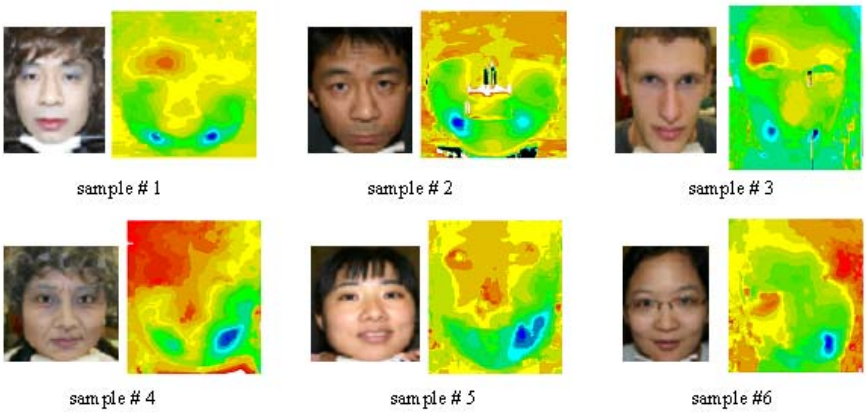


Fig. 4. Some of the subjects in our database. The *projection images* have been color-mapped so as to emphasize the variations

To compare our approach to still-image-based recognition, we used the PCA+LDA algorithm from the CSU Face Identification Evaluation System (CSU FaceId Eval) [15] as our baseline algorithm. CSU FaceId Eval is a collection of standard face recognition algorithms and statistical systems for comparing face recognition algorithms used in the FERET [14] tests. It includes four distinct algorithms – Principle Component Analysis (PCA) [3], a combined Principle Component Analysis and Linear Discriminant Analysis (PCA+LDA) [4], Bayesian Intrapersonal/Extrapersonal Classifier (BIC) [5] and Elastic Bunch Graph Matching (EBGM) [6]. For both of the approaches, we employed normalized cross-correlation as the similarity measure. The distance is computed as

$$d = 1 - s \tag{4}$$

For our experiment, we had a professional make-up artist disguise two of our subjects. We then applied the two approaches in an attempt to recognize them. Table 1 (on the next page) shows the distances we obtained with the PCA+LDA algorithm. The table also shows the query image (the image of the subject wearing make-up, that is presented to the algorithm), the target image (the image of the same subject stored in the database) and the top matches reported by PCA+LDA. From the results in Table 1, we observe that PCA+LDA failed to recognize the individual. For the first

Table 1. Distances obtained with the LDA algorithm

Query Image	Target Image	Best Match	2nd Best Match	3rd Best Match
				
Distance	0.83	0.32	0.45	0.73
Query Image	Target Image	Best Match	2nd Best Match	3rd Best Match
				
Distance	0.56	0.47	0.56	0.925

individual, the target image is not even among the top three matches and for the second individual the target image is only the second best match. Table 2 shows the distances we obtained with our method. We observe that in both the cases, our method clearly recognized the individual, while this is not easy to do even by visual inspection. Note, for our approach, we used the corresponding projection images for recognition and not the images shown in the first column of Table 2.

Table 2. Distances obtained with our DISC algorithm. Note, comparisons were done on the projection images and not the images shown here

Query person	Target person	Best Match	2nd Best Match	3rd best Match
				
Distances	0.077	0.077	0.709	0.797
Query person	Target person	Best Match	2nd Best Match	3rd Best Match
				
Distances	0.088	0.088	0.709	0.789

6 Conclusions

In this paper, we propose a novel method for face recognition that is based on tracking the dense texture of skin pores naturally present on the face instead of placing

cumbersome and less dense markers on the person's face. Our approach combines the strengths of both the still-image methods and video-based methods. It requires two subsequent high-resolution images of the individual performing a subtle facial motion, such as a faint smile. These images may be acquired by any higher-end consumer-grade camera. A central part of our approach is the deformation field that is computed for each individual using our Digital Image Skin Correlation (DISC) method. During the recognition process, projection images are generated from these deformation fields and a distance metric to projection images of individuals stored in the database is computed. The initial results presented in this paper verify the potential of our method to recognize faces accurately, even with heavy make-up, where the tested existing popular still-image method had difficulties.

This paper shows that our skin pore tracking-based face recognition technique is a promising way to accurately recognize faces. Our future work will seek to compute 3-dimensional deformation fields, to make our approach pose-invariant. We are also exploring the use of higher-level analysis methods. To this end, we are characterizing the deformation fields using vector derivatives and flow analysis techniques. Also, to more effectively evaluate the performance of our method in a variety of scenarios, we are planning to considerably expand our database. Another goal of ours is to optimize the DISC algorithm and its implementation, which will reduce the computation time down to seconds and also allow us to compute the deformable fields faster.

Acknowledgements

We would like to thank the NYSTAR "Center for Maritime and Port Security" for support, and we would also like to thank the subjects in the database for their participation. Work performed under IRB# FWA #00000125, project ID 20045536.

References

1. H.A. Bruck, S.R. McNeil, M.A. Sutton and W.H. Peters, "Digital Image Correlation Using Newton-Raphson Method of Partial Differential Correction," *Experimental Mechanics*, pp. 261-267, September 1989.
2. E. Guan, S. Smilow, M. Rafailovich and J. Sokolov, "Determining the Mechanical Properties of Rat Skin with Digital Image Speckle Correlation," *Dermatolog*, vol. 208, no. 2, pp. 112-119, 2004.
3. M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces," In Proceedings, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586-591, June 1991.
4. W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition," In Proceedings, *International Conference on Face and Gesture Recognition*, pp. 336, 1998.
5. B. Moghaddam, C. Nastar, and A. Pentland, "A Bayesian similarity measure for direct image matching," In Proceedings, In Proceedings, *International Conference on Pattern Recognition*, pp. 350–358, 1996.
6. L. Wiskott, J. Fellous, N. Kruger and C. Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Tranactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775-779, 1997.
7. L. Zhang and D. Samaras, "Face Recognition Under Variable Lighting using Harmonic Image Exemplars," In Proceedings, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19-25, 2003.

8. L. Zhang and D. Samaras, "Pose Invariant Face Recognition under Arbitrary Unknown Lighting using Spherical Harmonics," In Proceedings, *Biometric Authentication Workshop (in conjunction with ECCV 2004)*, 2004
9. X. Lu, A.K. Jain, "Deformation Analysis for 3D Face Matching," To appear in, *IEEE Workshop on Applications of Computer Vision*, 2005
10. H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen, "Automatic Video-based Person Authentication using the RBF Network," In Proceedings, *International Conference on Audio and Video-Based Person Authentication*, pp. 85-92, 1997.
11. S. McKenna and S. Gong, "Recognizing Moving Faces," In H.Wechsler, P.J. Phillips,V. Bruce, F.F. Soulie, and T.S. Huang, editors, *Face Recognition: From Theory to Applications*, NATO ASISeries F, vol. 163, 1998.
12. T. Choudhury, B. Clarkson, T. Jebara and A. Pentland, "Multimodal Person Recognition using Unconstrained Audio and Video," In Proceedings, *International Conference on Audio- and Video-Based Person Authentication*, pp. 176-181, 1999.
13. B. Li and R. Chellappa, "Face Verification through Tracking Facial Features", *Journal Optical Society of America,* vol. 18, no. 12, pp. 2969-81, 2001.
14. P.J. Phillips, H.J. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face Recognition Algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.
15. R. Beveridge, "Evaluation of Face Recognition Algorithms", *http://cs.colostate.edu/evalfacerec.*
16. W. Zhao, Chellappa R, Philips PJ, Rosenfeld A, "Face Recognition: A Literature Survey", *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, 2003.

Rank-Based Decision Fusion for 3D Shape-Based Face Recognition

Berk Gökberk, Albert Ali Salah, and Lale Akarun

Boğaziçi University
Computer Engineering Department
Turkey
{gokberk,salah,akarun}@boun.edu.tr

Abstract. In 3D face recognition systems, 3D facial shape information plays an important role. Various shape representations have been proposed in the literature. The most popular techniques are based on *point clouds*, *surface normals*, *facial profiles*, and statistical analysis of *depth images*. The contribution of the presented work can be divided into two parts: In the first part, we have developed face classifiers which use these popular techniques. A comprehensive comparison of these representation methods are given using 3D RMA dataset. Experimental results show that the *linear discriminant analysis-based* representation of depth images and *point cloud* representation perform best. In the second part of the paper, two different multiple-classifier architectures are developed to fuse individual shape-based face recognizers in parallel and hierarchical fashions at the decision level. It is shown that a significant performance improvement is possible when using rank-based decision fusion in ensemble methods.

1 Introduction

Despite two decades of intensive study, the challenges of face recognition remain: changes in the illumination and in-depth pose problems make this a difficult problem. Recently, 3D approaches to face recognition have shown promise to overcome these problems [1]. 3D face data essentially contains multi-modal information: *shape* and *texture*. Initial attempts in 3D research have mainly focused on *shape* information, and combined systems have emerged which fuse shape and texture information.

Surface normal-based approaches use facial surface normals to align and match faces. A popular method is to use the EGI representation [2, 3]. *Curvature-based* approaches generally segment the facial surface into patches and use curvatures or shape-index values to represent faces [4]. *Iterative Closest Point-based (ICP)* approaches perform the registration of faces using the popular ICP algorithm [5], and then define a similarity according to the quality of the fitness computed by the ICP algorithm [6–8]. *Principal Component Analysis-based (PCA)* methods first project the 3D face data into a 2D intensity image where the intensities are determined by the depth function. Projected 2D depth images can later be processed as standard intensity images [9–11]. *Profile-based* or *contour-based*

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1019–1028, 2005.
© Springer-Verlag Berlin Heidelberg 2005

approaches try to extract salient 2D/3D curves from face data, and match these curves to find the identity of a person [12, 13]. *Point signature-based* methods encode the facial points using the relative depths according to their neighbor points [14, 15].

In addition to the pure shape-based approaches, 2D texture information has been combined with 3D shape information. These multi-modal techniques generally use PCA of intensity images [16, 17], facial profile intensities [13], ICP [18, 19], and Gabor wavelets [14]. These studies indicate that combining shape and texture information reduces the misclassification rate of a face recognizer.

One aim in this paper is to evaluate the usefulness of state-of-the-art shape-based representations and to compare their performance on a standard database. For this purpose, we have developed five different 3D shape-based face recognizers. They use: *ICP-based point cloud* representation, *surface normal-based* representation, *profile-based* representation, and two *depth image-based* representations: PCA and Linear Discriminant Analysis (LDA), respectively. Our second aim is to analyze whether combining these distinct 3D shape representation approaches can improve the classification performance of a face recognizer. To accomplish the fusion, we have designed two fusion schemes, *parallel* and *hierarchical*, at the sensor decision level. Although it has been shown in the literature that fusion of texture and shape information can increase the performance of the system, the fusion of different 3D shape-based classifiers has remained as an open problem. In this work, we show that the integration of distinct shape-based classifiers by using a rank-based decision scheme can greatly improve the overall performance of a 3D face recognition system.

2 3D Shape-Based Face Recognizers

2.1 Registration

Registration of facial data involves two steps: a preprocessing step and a transformation step. In the preprocessing step, a surface is fitted to the raw 3D facial point data. Surface fitting is carried out to sample the facial data regularly. After surface fitting, central facial region is cropped and only the points inside the cropped ellipsoid are retained. In order to determine the central cropping region, nose tip coordinates are used. Figure 1 shows a sample of the original facial data, and the cropped region. Cropped faces are translated so that the nose tip locations are at the same coordinates. In the rest of the paper, we refer to the cropped region as the facial data.

After preprocessing of faces, a transformation step is used to align them. In the alignment step, our aim is to rotate and translate faces such that later on we can define acceptable similarity measures between different faces. For this purpose, we define a *template face model* in a specific position in the 3D coordinate system. Template face is defined as the average of the training faces. Each face is rigidly rotated and translated to fit the template. Iterative Closest Point (ICP) algorithm is used to find rotation and translation parameters. The correspondences found between the template face and any two faces F_i and F_j by the ICP algorithm are then used to establish point-to-point dense correspondences.

2.2 3D Facial Shape Representations

Several 3D features can be extracted from registered faces. The simplest feature consists of the 3D coordinates of each point in the registered facial data (*point cloud representation*). Another representation, *surface normal representation*, is based on surface normals calculated at each 3D facial point. Both point cloud and surface normal-based approaches are related to whole facial surfaces. Besides surface-based features, facial profiles are also found to be important for discriminating 3D faces. In this work, we have extracted seven equally spaced vertical profiles, one central and three from either side of the profile (*profile set representation*). See Figure 1.b for the extracted profiles.

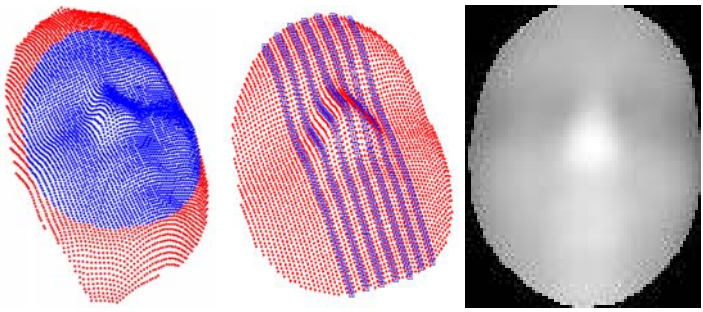


Fig. 1. (a) Cropped region, (b) extracted facial profiles, and (c) depth image

Facial profile can be found by using the 3D symmetry property of faces. However, in this work, we have used the nose region to find the central profile. We use the (x, y) coordinates of the topmost points over the nose. These points form an approximately ellipsoid cluster on the xy-plane. The vertical line passing through the center of nose can then be easily found by calculating the principal direction. To find the principal direction, we have performed PCA on the x and y coordinates of the topmost k nose points. Since all faces are registered to the template face, we can speed up the profile extraction process by simply finding the first principal direction in the face template once, and searching for closest points in a given registered 3D face image. This approach performs better since average template face is more robust to irregular nose shapes.

Registration of profile contours is performed by translating profile curves in such a way that nose tips of profiles are always at the same xy- coordinates. After aligning profile curves, a spline is fitted to the profile curve, and it is regularly sampled in order to be able to compute Euclidean distances between two profiles.

In the PCA and LDA techniques, the 3D face points are projected to a 2D image where the intensity of a pixel denotes the depth of a 3D point. Figure 1.c shows a sample depth image. Statistical feature extraction methods can be used to extract features from depth images. In this work, we have employed PCA (*Depth-PCA*) and LDA (*Depth-LDA*) to extract features from depth images.

2.3 Similarity Measures and Classifiers

In our system, we have used k-nearest neighbor algorithm (k-NN) as a pattern classifier which is intensively used in face recognition systems due to its high recognition accuracy. In order to use k-NN, we have to define a similarity measure for each representation used. Let Φ_i be a 3D face. We can represent Φ_i in *point cloud representation* as $\Phi_i^P = \{p_1^i, p_2^i, ...p_N^i\}$, where N is the number of points in the face and p^is are 3D coordinates. We define the distance between two faces Φ_i and Φ_j as:

$$D(\Phi_i^P, \Phi_j^P) = \sum_{k=1}^{n} ||p_k^i - p_k^j|| \tag{1}$$

where $||.||$ denotes Euclidean norm. Similarly, in *surface normal representation*, face Φ_i is represented by $\Phi_i^N = \{n_1^i, n_2^i, ...n_N^i\}$, where n^is are surface normals calculated at points p^is. Distance between two faces Φ_i and Φ_j in surface normal representations can be defined as in the above formula, but by replacing Φ^Ps with Φ^Ns.

In *profile set representation*, we have seven equally spaced vertical profiles, $C_k, (k = 1..7)$. Each profile curve C_k is a vector and contains n_k depth coordinates: $C_k = \{z_1, z_2, ..., z_{n_k}\}$. Therefore, we represent a face in profile set representation as $\Phi_i^R = \bigcup C_k$. The distance between two corresponding k^{th} profile curves of face i and face j can be determined by $d(C_k^i, C_k^j) = \sum_{m=1}^{n_k} ||z_m^i - z_m^j||$. Then, the distance between faces Φ_i and Φ_j is defined as the sum of the distances between each corresponding profile curve. In depth image-based face representations, the distance between two faces is calculated as the Euclidean distance between extracted feature vectors.

3 Combination of Shape-Based Face Recognizers

When working on different representations, classifiers can be made more accurate through combination. Classifier combination has caught the attention of many researchers due to its potential for improving the performance in many applications [20–22]. In classifier fusion, the outputs of individual classifiers (*pattern classifiers*) are fused by a second classifier (*combination classifier*) according to a combination rule. In order to produce a successful ensemble classifier, individual pattern classifiers should be highly tuned, diverse, and should not be redundant. In this work, the diversity of the pattern classifiers is provided by letting them use a different face representation. In our system, the outputs of individual pattern classifiers are the ranked class labels and their associated similarity scores. However, we only use the rank information because of the variability of the score functions produced by different representations.

In our fusion schemes, a *combination set* is formed by selecting the most similar k classes for each pattern classifier and by feeding these into the combining classifier. As *combination rules* for rank-output classifiers, we have used *consensus voting*, *rank-based combination* and *highest-rank majority* methods [23]. In *consensus voting*, the class labels from the combination set of each pattern classifier are pooled, and the most frequent class label is selected as the output. In

rank-based combination, the sum of the rankings of each class in all combination sets are used to compute a final ranking (*rank-sum* method). A generalization of the rank-sum method is to transform ranks by a function f which maps ranks $\{1, 2, 3, ..., K\}$ to $\{f(1), f(2), f(3), ..., f(K)\}$. f may be any nonlinear monotonically increasing function. The motivation to use such a function f is to penalize the classes at the bottom of a ranked list. In this work, $f(x) = x^n$ is used as a mapping function. In *highest-rank majority*, a consensus voting is performed among the rank-1 results of each pattern classifier.

3.1 Parallel Fusion of Face Classifiers

We have designed a parallel ensemble classifier which fuses the rank-outputs of different face pattern classifiers. Profile set, Depth-LDA, point cloud and surface normal-based face representations are chosen in these pattern classifiers. Combination set is formed by selecting the most similar N classes in the rank outputs of each classifier. As a combination rule, four different types of rules are used: consensus voting, rank-sum, nonlinear rank-sum and highest-rank majority rule. In nonlinear rank-sum method, $f(x) = x^n$ function is used. If $n = 1$, nonlinear rank-sum method is identical to the standard rank-sum method. As a generalization of the rank outputs of individual classifiers, we have also used the ranking of each training instance whereas in standard rank-output classifiers, classes are assigned a single rank. In the rest of the paper, we will refer to the generalized method as *instance-based ranking*, and the standard method as *class-based ranking*. See Figure 2 for a schematic diagram of the parallel fusion scheme.

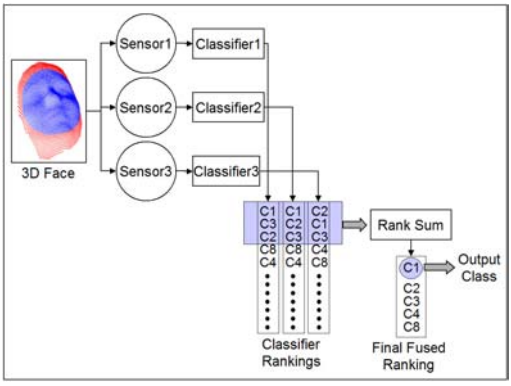


Fig. 2. A schematic diagram of the parallel combination scheme

3.2 Hierarchical Fusion of Face Classifiers

In addition to the parallel fusion scheme, we have also designed a hierarchical fusion methodology. The main motivation of the hierarchical architecture is to filter out the most similar K classes using a simple classifier, and then to feed

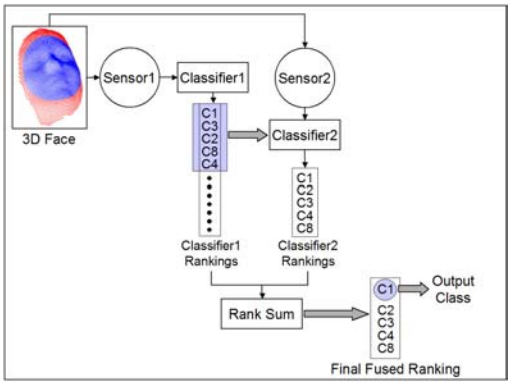


Fig. 3. A schematic diagram of the hierarchical combination scheme

these K classes into a more complex and powerful second classifier. For this purpose, we have used the point cloud-based nearest neighbor classifier as the first classifier C_1, and depth map-based LDA classifier as the second classifier C_2. The use of LDA as a second classifier is based on the idea that it can boost the differences between similar classes in the transformed feature space. See Figure 3 for a schematic diagram of the hierarchical fusion.

As in the previous section, C_1 produces an instance-based ranking R_1, and then class labels of the top K instances are passed to C_2. C_2 then performs a linear discriminant analysis on the depth images of the training examples of these classes, and forms a feature space. Nearest neighbor classifier is used in this feature space to produce a new instance-based ranking R_2. If only the rank-1 class output of R_2 is used, the information in C_1 is discarded. We use a nonlinear rank-sum method to fuse R_1 and R_2 which is superior to using R_2 alone.

4 Experimental Results

In our experiments, we have used the 3D RMA dataset [13]. Specifically, a subset of the automatically prepared faces were used in experiments, which consists of 106 subjects each having five or six shots. The data is obtained with a stereo vision assisted structured light system. On the average, faces contain about 4000 3D points, and they cover different portions of the faces and the entire data is subject to expression and rotation changes. To be able to statistically compare the algorithms, we have designed five experimental sessions.

Table 1 shows which shots of a subject are placed into the training and test sets for each session. At each session, there are exactly 193 test shots in total.

4.1 Performance of Different Shape Features

Table 2 summarizes the classification accuracies of each representation method. Best performance is obtained using *Depth-LDA* which has an average recognition

Table 1. Training and test set configurations

Session	Training Set Shots	Test Set Shots
S_1	{1, 2, 3, 4}	{5, 6}
S_2	{1, 2, 3, 5}	{4, 6}
S_3	{1, 2, 4, 5}	{3, 6}
S_4	{1, 3, 4, 5}	{2, 6}
S_5	{2, 3, 4, 5}	{1, 6}

Table 2. Classification accuracies of each classifier for each experimental session. d denotes the feature dimensionality of the representations

Session	Point Cloud ($d = 3,389 \times 3$)	Surface N. ($d = 3,389 \times 3$)	Depth-PCA ($d = 300$)	Depth-LDA ($d = 30$)	Profile Set ($d = 1,557$)
S_1	93.26	93.26	49.74	95.34	94.30
S_2	94.82	97.93	52.33	97.41	92.75
S_3	96.89	93.26	49.74	95.34	92.75
S_4	97.41	96.89	51.30	96.37	95.86
S_5	97.41	96.37	50.78	96.89	95.86
Mean	95.96	95.54	50.78	96.27	94.30
STD	1.85	2.16	1.10	0.93	1.55

accuracy of 96.27 per cent. The dimensionality of the reduced feature vector of the Depth-LDA method is 30. Point cloud and surface normal representations 95.96 and 95.54 per cent correct recognition rate on the test set, respectively. In each of these representation schemes, feature vector size is $3,389 \times 3 = 10167$, since there are 3,389 points in each representation method and each point is a 3D vector. Profile set representation has a recognition accuracy of 94.30 per cent. The feature dimensionality of the profile set representation is the sum of the number of sampled points for each individual profile curve. In our representation, this dimensionality is 1,557. Depth-PCA method performed worst with a 50.78 per cent recognition accuracy, using 300 dimensional feature vectors.

4.2 Performance of Parallel and Hierarchical Decision Fusion Schemes

In our experiments on the parallel fusion scheme, we have tested all possible combinations of *point-cloud*, *surface normal*, *profile-set*, and *Depth-LDA* based classifiers. We have also analyzed the effect of the combination set size (N) in the fusion process. Average recognition accuracies of different ensemble architectures are shown in Table 3. The best classification accuracy is obtained by a nonlinear rank-sum combination rule where the pattern classifiers are *profile set*, *Depth-LDA* and *surface normal*-based representations. In this architecture, combination set size is $N = 6$, and the nonlinear function used is $f(x) = x^3$. It is seen that instance-based ranking outperforms class-based ranking except for the

Table 3. Mean classification accuracies of hierarchical fusion methods. S denotes the selected individual classifiers in the ensemble, where S = { 1: Profile set, 2: Depth-LDA, 3: Point cloud, 4: Surface Normals}

	Instance-based Ranking	Class-based Ranking
Consensus Voting	98.76 (N=2) S={2,3,4}	98.34 (N=1) S={1,2,3,4}
Nonlinear Rank-Sum	99.07 (N=6) S={1,2,4}	98.86 (N=1) S={1,2,3,4}
Highest Rank Majority	98.13 (N=1), S={1,2,3,4}	98.34 (N=1) S={1,2,3,4}

highest rank majority rule. As a combination rule, nonlinear rank-sum method consistently outperforms its alternatives. We observe that parallel combination of different pattern classifiers which rely on distinct feature sets significantly improves the recognition accuracies in all cases. We confirm this finding with paired t-test on five-fold experiments.

In hierarchical fusion experiments, point-cloud-based first classifier C_1 produces an *instance-based* rank list. On the average, first rank-80 instances provide 100 per cent recognition accuracy in C_1. We have seen that 80 training instances in the combination set corresponds to approximately 25 classes. Therefore, our Depth-LDA based second classifier C_2 dynamically constructs a feature space using these 25 classes. Finally, the ranks produced by C_1 and C_2 are integrated using nonlinear rank-sum technique where $f(x) = x^3$. The average performance of the hierarchically combined classifiers is found to be 98.13 per cent, which is statistically significantly different from all individual classifier's accuracies. As in the parallel case, hierarchical fusion is found to be beneficial when compared to individual classifier accuracies. The accuracy of the parallel fusion of point cloud and Depth-LDA using nonlinear rank-sum is 98.45 per cent and is better than hierarchical fusion.

5 Conclusion

In this work, we have compared some of the state-of-the-art 3D shape-based face representation techniques frequently used in 3D face recognition systems. They include ICP-based point cloud representations, surface normal-based representations, PCA and LDA-based depth map techniques and facial profile-based approaches. It has been shown that among these methods, Depth-LDA method performs best, and point cloud and surface normal-based classifiers have a comparable recognition accuracy. Our results on Depth-PCA confirmed the sensitivity of PCA to alignment procedure. To obtain better results, facial landmarks need to be correctly aligned, possibly by warping of faces. In our work, we choose not to warp facial surfaces since it is known that such a warping process suppresses discriminative features [8].

We have also developed parallel and hierarchical combination schemes to fuse the outputs of individual shape-based classifiers. In the parallel architecture, a subset of the rank outputs of surface-normal, Depth-LDA, and profile-based classifiers are fused using nonlinear rank-sum method, and the recognition accuracy

improved to 99.07 per cent from 96.27 per cent which is the best individual classifier's (Depth-LDA) accuracy. In the hierarchical fusion scheme, we transfer the most probable classes found by our first point-cloud based classifier to a Depth-LDA based second classifier, where LDA makes use of the differences between similar classes in the transformed feature space. The hierarchical architecture reaches a 98.13 per cent recognition accuracy which is statistically superior to all individual performances according to paired t-tests. As a conclusion, we observe that the combination of separate shape-based face classifiers improves the classification accuracy of the whole system, when compared to using individual classifiers alone. As a future work, we plan to investigate the fusion of shape-based ensemble classifiers with texture-based ensemble methods.

References

1. Bowyer, K.W., Chang, K., Flynn, P.J.: A survey of 3D and multi-modal 3D+2D face recognition. In: International Conference on Pattern Recognition. (2004)
2. Lee, J.C., Milios, E.: Matching range images of human faces. In: International Conference on Computer Vision. (1990) 722–726
3. Tanaka, H.T., Ikeda, M., Chiaki, H.: Curvature-based face surface recognition using spherical correlation principal directions for curved object recognition. In: International Conference on Automated Face and Gesture Recognition. (1998) 372–377
4. Moreno, A.B., Sanchez, A., Velez, J.F., Diaz, F.J.: Face recognition using 3D surface-extracted descriptors. In: Irish Machine Vision and Image Processing Conference. (2003)
5. Besl, P., McKay, N.: A method for registration of 3D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14** (1992) 239–256
6. Medioni, G., Waupotitsch, R.: Face recognition and modeling in 3D. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures. (2003) 232–233
7. Lu, X., Colbry, D., Jain, A.: Matching 2.5d scans for face recognition. In: International Conference on Pattern Recognition. (2004) 30–36
8. Irfanoglu, M.O., Gokberk, B., Akarun, L.: 3D shape based face recognition using automatically registered facial surfaces. In: International Conference on Pattern Recognition. (2004) 183–186
9. Hesher, C., Srivastava, A., Erlebacher, G.: A novel technique for face recognition using range imaging. In: International Symposium on Signal Processing and Its Applications. (2003) 201–204
10. Pan, G., Wu, Z., Pan., Y.: Automatic 3d face verification from range data. In: International Conference on Acoustics, Speech, and Signal Processing. Volume 3. (2003) 193–196
11. Xu, C., Wang, Y., Tan, T., Quan, L.: Automatic 3D face recognition combining global geometric features with local shape variation information. In: International Conference on Automated Face and Gesture Recognition. (2004) 308–313
12. Lee, Y., Park, K., Shim, J., Yi, T.: 3D face recognition using statistical multiple features for the local depth information. In: International Conference on Vision Interface. (2003)
13. Beumier, C., Acheroy, M.: Face verification from 3D and grey level cues. Pattern Recognition Letters **22** (2001) 1321–1329

14. Y.Wang, Chua, C., Ho, Y.: Facial feature detection and face recognition from 2D and 3D images. Pattern Recognition Letters **23** (2002) 1191–1202
15. Chua, C.S., Han, F., Ho, Y.K.: 3D human face recognition using point signature. In: Proceedings of Int. Conf. on Automatic Face and Gesture Recognition. (2000) 233–237
16. Tsalakanidou, F., Tzocaras, D., Strintzis, M.: Use of depth and colour eigenfaces for face recognition. Pattern Recognition Letters **24** (2003) 1427–1435
17. Chang, K., Bowyer, K., Flynn, P.: Face recognition using 2D and 3D facial data. In: Multimodal User Authentication Workshop. (2003) 25–32
18. Papatheodorou, T., Reuckert, D.: Evaluation of automatic 4d face recognition using surface and texture registration. In: International Conference on Automated Face and Gesture Recognition. (2004) 321–326
19. Lu, X., Jain, A.K.: Integrating range and texture information for 3D face recognition. In: IEEE Workshop on Applications of Computer Vision. (2005) To appear
20. Toygar, O., Acan, A.: Multiple classifier implementation of a divide-and-conquer approach using appearance-based statistical methods for face recognition. Pattern Recognition Letters **25** (2004) 1421–1430
21. Khuwaja, G.A.: An adaptive combined classifier system for invariant face recognition. Digital Signal Processing **12** (2002) 21–46
22. Jing, X., Zhang, D.: Face recognition basedon linear classifiers combination. Neurocomputing **50** (2003) 485–488
23. Melnik, O., Vardi, Y., Zhang, C.H.: Mixed group ranks: Preference and confidence in classifier combination. IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004) 973–981

Robust Active Shape Model Construction and Fitting for Facial Feature Localization

Zhenghui Gui and Chao Zhang

National Laboratory on Machine Perception, Peking University,
Beijing, 100871, P.R. China
{agui,chzhang}@cis.pku.edu.cn

Abstract. Active Shape Model (ASM) has proved to be a powerful tool for interpreting face images. However, it may fail in the presence of non-Gaussian noise, or outliers. In this paper, we present a framework for both automatic model construction and efficient model fitting with outliers. In model construction, the training face samples are automatically labeled by local image search using Gabor wavelet features. Then robust principle component analysis (RPCA) is applied to capture the statistics of shape variations. In model fitting, an error function is introduced to deal with the outlier problem, which provides a connection to robust M-estimation. Gauss-Newton algorithm is adopted to efficiently optimize the robust energy function. Extensive experiments demonstrate the efficiency and robustness of our approach over previous methods.

1 Introduction

Active Shape Model (ASM) [1] is a generative parametric model commonly used to locate facial features. Accurate facial feature localization plays an important role in many areas such as face recognition [2], head pose estimation [3] and face expression analysis [4].

The traditional ASM is built from a training set of annotated images, in which canonical feature points have been manually marked. One implicit but crucial assumption is that all these feature points should be visible in these images. However, in many scenarios, there is the opportunity for undesirable artifacts due to ambiguous local image evidence arousing from strong illumination, occlusion or image noise, etc. These artifacts may lead to inaccurate annotation and are viewed as outliers in statistics. Since image ambiguity is very common in face images, in order to automate ASM training process, it is important to be able to construct the model from training set containing outliers.

ASM search also suffers from the outlier problem since it is based on a least squares approach, as is pointed out by M. Rogers and J. Graham [5]. They introduced robust methods, e, g. using M-estimation and random sampling, to avoid the bias in ASM shape parameter estimation. Unfortunately, straightforward applications of these techniques are time consuming.

In this paper, we present an efficient and robust approach, called "R-ASM", to deal with the outlier problem. The first part of this work is the automatic construction of the statistic model. The training face samples are automatically labeled by Gabor wavelet based local image search. Robust principal analysis with missing data and outliers (RPCA) [6], which has shown to be successful for performing face tracking

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1029–1038, 2005.
© Springer-Verlag Berlin Heidelberg 2005

in the presence of occlusion [7], is then carried out on this training set containing outliers to capture shape variations. In the model fitting part, an error function is introduced to deal with the outlier problem, which is much like M-estimation. The optimization of robust energy function is achieved by Gauss-Newton scheme, which is commonly used for non-linear least squares fitting. Besides, in our approach, the error function's scale parameter is learned from training samples, this imposes a priori knowledge of shape variations on image searching, and is much reliable for outlier rejection.

The rest of this paper is arranged as follows: in section 2, a brief review of ASM is given. In section 3, we describe how to construct the robust model automatically. We show how to efficiently fit an ASM with outliers in section 4. Experimental results are presented in Section 5. Finally, the concluding remarks are given in Section 6.

2 A Review of Active Shape Model

In this section we briefly sketch the ASM framework, a more detailed description of ASM can be found in [1].

In ASM, an object shape is represented by a set of landmark points. Several instances of the same object class are included in a training set. In order to model the variations, the landmark points from each image are represented as a vector **x** after alignment to the model co-ordinate frame. Then PCA is performed to capture variation statistics. Any shape can therefore be approximated by: $\boldsymbol{x} = \bar{\boldsymbol{x}} + \boldsymbol{P}\boldsymbol{b}$, where $\bar{\boldsymbol{x}}$ is the mean shape, $\boldsymbol{P}$ is a set of modes of variation and $\boldsymbol{b}$ is the controlling parameters.

For a given shape $\boldsymbol{x}$, iterative matching is performed to approximate it as closely as possible. In each step of iteration, we wish to find $\Delta\boldsymbol{b}$ so as to minimize the error term:

$$\boldsymbol{E} = \| \boldsymbol{x} - \boldsymbol{x}_0 \|_2 \tag{1}$$

where $\boldsymbol{x}_0$ is current reconstruction of the model and is initialized to the mean shape at the beginning of iteration. The minimization can be shown to have a solution of the form: $\Delta\boldsymbol{b} = (\boldsymbol{P}^T * \boldsymbol{P})^{-1}\boldsymbol{P}^T\Delta\boldsymbol{x}$, where $\Delta\boldsymbol{x}$ is the shape update vector. As $\boldsymbol{P}$ is orthonormal, this simplifies to $\Delta\boldsymbol{b} = \boldsymbol{P}^T\Delta\boldsymbol{x}$. In ASM search, $\Delta\boldsymbol{x}$ is obtained by searching local image region for strong edges around each landmark point.

It can be seen that the parameter estimation of standard ASM aims to minimize the sum of squares of residuals between the model and the data. As is wildly known, PCA regression is suboptimal when the noise is not Gaussian. While in local image searching, outliers often occur due to the ambiguous local image evidence. In this situation, ASM may diverge from the desired result.

3 Automatic Model Construction with Outliers

We first describe automatic labeling of face samples, and then introduce robust PCA for model construction with outliers.

It has been wildly recognized that Gabor wavelets offer the best localization ability in both frequency and image space. Wiskott *et al.* [8] used Gabor wavelets to gener-

ate a data structure named the Elastic Bunch Graph to locate facial features. In Elastic Bunch Graph, local image structure is modeled by sets of Gabor wavelet components (jets). Two similarity functions are defined between two jets: one is phase-insensitive, which measures how "similar" the two jets are, the other is phase-sensitive, which can be used to estimate displacement between the two jets taken from object image locations sufficiently close.

To automate the training process, we start with a number of selected representative face images. They are manually labeled and a face bunch graph is built up. After aligning the shapes to a mean shape **x**, the following search procedure is performed to increase training set:

Input: Unlabeled face images, face bunch graph, mean shape **x**, inlier similarity threshold s*im_thresh*, inlier proportion threshold *pro_thresh*

1) Use the face detection algorithm to get a rough estimation of eye regions.
2) Use eye bunch to locate both eye positions in the eye regions
3) Rotate, scale and translate mean shape **x** to match the found eyes in image co-ordinate frame as initial searching shape.
4) For each remaining landmarks:
   a) Use the phase-sensitive similarity function to get new position candidates and use phase-insensitive similarity function to calculate the similarity between the new jets and the face bunches.
   b) If the new position is within the neighborhood of 8 pixels of the initial search position and the highest similarity value is above *sim_thresh*, the new position is accepted as an "inlier", otherwise is recognized as an "outlier".
5) If detected inlier proportion is above *pro_thresh*, add current face image to the training set. Turn to next image.

As is a common practice, the multi-resolution strategy is adopted for large images.

In traditional ASM, all landmarks are used for alignment. In our case, only inliers are used. Observing that the distributions of inliers' landmark residuals can be well approximated by Gaussian density, we compute the variance of each landmark points in the training set as: $\sigma_{t,j} = \frac{1}{T}\sum_{i=1}^{T} e_i^2(j)$, where T is the number of inliers for the jth landmark, and e_i is residual. Since in Gaussian distribution, about 95% of the computed residuals' absolute value belong to the interval $[0, 2\sigma_{t,j}]$, it is a common practice to treat points with residuals' absolute value without this interval as outliers. Fig. 1 shows the landmark variances overlaid on the mean shape.

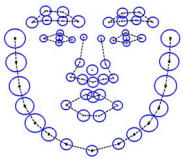


Fig. 1. Landmark point variances of training set overlaid on the mean shape. The circles show $2\sigma_{t,j}$ variation of corresponding landmarks

The training shapes are then refined and re-aligned. The landmark coordinates are initially binary weighted and then robust principle component analysis is performed on them to construct a robust shape model. Though there are many RPCA algorithms

in the literatures [10][13][14], we adopt Torre and Black's method [10], which is motivated by dealing with intra-sample noise in training images and best suits our case.

4 Model Fitting with Outliers

We now describe how to search a face shape with outliers both robustly and efficiently. We first introduce our robust energy function, and then describe previous robust method (non-efficient). After that, we'll show how it can be improved to efficiently fit to images.

In R-ASM fitting, to robustly estimate the coefficients b, we replace the quadratic error norm in Eq. (1) with a robust error norm ρ, and we aim to minimize the following robust energy function:

$$E_{rpca}(P,b,\bar{x},\sigma) = \rho(e,\sigma) = \sum_{i=1}^{d} \rho(x_i - \bar{x}_i - \sum_{j=1}^{m} P_{ij} b_{ji}, \sigma) \tag{2}$$

where d is the number of landmarks, $\bar{x}$ is the mean shape, $\boldsymbol{e}$ is the residual, σ is a scale parameter controlling the shape of the error function.

In Rogers' approach, the M-estimation based estimation of parameter $\boldsymbol{b}$ is the solution to the following m equations:

$$\sum_{i=1}^{d} \psi(e_i) \frac{\partial e_i}{\partial b_j} = 0 , \quad \text{for } j=1,\ldots,m, \tag{3}$$

where the deviation $\psi(x) = d\rho(x)/dx$ is called the influence function, representing the influence of a datum on the parameter estimation. Eq. (3) was then formed as a weighted least squares problem and solved by a closed form solution, see [5] for further information. The algorithm is summarized in Fig.2. Note that we only consider shapes in model coordinate frame for simplicity.

The Weighted Least Squares Approach For Parameter Estimation

Input: mean shape $\bar{x}$, current shape parameter **b**, searched shape y

1. Generate the model point positions using: $x = \bar{x} + Pb$
2. Calculate absolute value of residuals: $e = |x - y|$
3. Estimate scale parameters of error function from the median of **e**, known as Median Absolute Deviation (MAD): $\sigma = 1.4826(1+5/(d-m))median(e_i)$, use this to weight each residual according to Huber influence function.
4. Calculate parameter update in the closed form: $\Delta b = ((P^T W^T W P)^{-1} P^T W^T W)^T (x-y)$
5. Update **b**. Repeat until convergence

Fig. 2. The closed form solution of weighted least squares approach for parameter estimation

$W \in R^{d \times d} = diag(w_i)$ is a diagonal matrix containing the positive weighting coefficients given by the influence function for the data elements.

The most computationally expensive part of the algorithm involves computing the closed form solution (step 4) and MAD estimation of scale parameter σ (step 3).

It is worth noting that the scale parameter is the same for every datum element and need to be estimated in each iteration step. This is very un-natural as can be seen from Fig.1, landmark coordinate variations differ with one another. In the R-ASM, the inlier point variations have been computed in the training stage, it is nature to make use of these variations to decide different scale parameters for different landmarks. The energy function can then be reformulated as:

$$\begin{aligned} E_{rpca}(P,b,\bar{x},\sigma) &= \rho(e,\sigma) \\ &= \sum_{i=1}^{d} \rho(x_i - \bar{x}_i - \sum_{j=1}^{m} P_{ij} b_{ji}, \sigma_i) \end{aligned} \tag{4}$$

We use Geman-McClure error function here since it is twice differentiable: $\rho(x,\sigma) = \frac{x^2}{x^2+\sigma^2}$.

Considering that points with residuals $|e_i| > 2\sigma_{t,i}$ are outliers and that for non-convex robust functions ρ, the influence of the outliers begins to decrease at their inflexion point, the scale parameters for Geman-McClure error function can be obtained as $\sigma_i = 2\sqrt{3}\sigma_{t,i}$.

There are several possible ways to update the parameters more efficiently, rather than a closed form solution [10]. Here we adopt Gauss-Newton scheme to minimize Eq (4). Gauss-Newton algorithm is an iterative gradient descent method wildly used for non-linear least squares problems.

In each minimization iteration step, we update b according to: $b_i^{(n+1)} = b_i^{(n)} - \Delta b$, with Δb given by normalized partial deviation of Eq (4). We drop the sample index n for a single test shape:

$$\Delta b_i = \frac{1}{w(b_i)} \frac{\partial}{\partial b_i} E_{rpca}(P, b^{(n)}, \bar{x}, \sigma) = -\frac{1}{w(b_i)} \sum_{j=1}^{d} P_i(j) \psi(e_j, \sigma_j) \tag{5}$$

where P_i is the i th mode, e_j is the reconstruction error of the j th element and σ_j the corresponding scale parameter. $\psi(e_j,\sigma_j) = \frac{2e_i{\sigma_j}^2}{({\sigma_j}^2 + {e_j}^2)^2}$ is the first deviation of the error function ρ.

The step size is determined by a local quadratic approximation. The norm term $w(b_i)$ is an upper bound of the second deviation respect to parameter **b**:

$$w(b_i) = \sum_{j=1}^{d} (P \frac{\partial}{\partial b_i} b)^2 \max \psi_i' = \sum_{j=1}^{d} P_i^2 \max \psi_i' \tag{6}$$

in which P_i^2 is the square of P_i at element j, and $\max \psi_i' = \max_x \frac{\partial^2 \rho(x,\sigma_i)}{\partial x^2} = \frac{2}{{\sigma_i}^2}$

Our approach of one iteration of parameter estimation can be summarized as:

The Proposed Gauss-Newton Method For Shape Parameter Estimation

Input: pre-computed σ , mean shape $\bar{x}$, current shape parameter **b**, searched shape y

1. Generate the model point positions using: $x = \bar{x} + Pb$
2. Calculate residuals: $e = x - y$
3. Calculate parameter update Δb according to Eq. (5)
4. Update **b**. Repeat until convergence

Fig. 3. The proposed Gauss-Newton method for parameter estimation

As can be seen from above, the computational cost for one iteration of closed form solution is $O(m^2 d) + O(d \log d)$, while using Gauss-Newton method and pre-computed scale parameter, this is reduced to $O(md)$.

In RPCA training, Geman-McClure error function is used to get a unique solution since it is convex. However, its corresponding influence functions is monotone, the influence of outliers is thus bounded, but not null. From this point of view, in model fitting, we further trisected the output of Geman-McClure influence function to weight the residuals experimentally, which is much similar to Huber's strategy [11]:

$$w_i = \begin{cases} 1 & r_i < \sigma_i / 2\sqrt{3} \\ \dfrac{2 r_i \sigma_i^2}{(r_i^2 + \sigma_i^2)^2} & \sigma_i / 2\sqrt{3} \le r_i < \sqrt{3}\sigma_i \\ 0 & r_i \ge \sqrt{3}\sigma_i \end{cases}$$

where r_i is the *i* th residual element, and w_i is the corresponding weight.

5 Experiment Results

The experiments have been conducted on a data set consisted of 505 frontal face images, in which each face area is about 150*150 pixels. See Fig. 4 for a fully labeled face example with 54 landmarks.

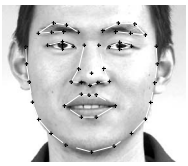


Fig. 4. A labeled face image sample

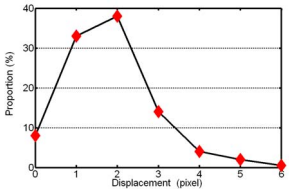


Fig. 5. Point displacement between training set inliers and their ground truth positions

Model Construction: 100 representative face images from training set were chosen and manually labeled for model construction, then automatic search was performed on another randomly selected 350 images, as described in section 3.2. Rejecting samples containing more than 12% outliers, our training set was finally consisted of 400 images. The remaining 105 images are used for testing.

Displacement between inlying training landmark locations and their corresponding ground truth positions is shown in Fig. 5. The displacement (*x-coordinate*) is plotted against corresponding percentage (*y-coordinate*). We can see that the automatic search gives very accurate results.

After alignment, RPCA was applied to the noisy set to construct a PDM, referred as robust PDM. The shape modes were set to be the orthonormalized eigenvectors with the largest eigenvalues. As a common practice, we retain enough shape modes to explain 95% of the observed variation in the training set. A conventional PDM was also built using the same 400 samples, but with manually labeled image samples.

Fig. 6 shows the variations of the first 3 main modes overlaid on mean shapes for both PDMs. The Valid Shape Ranges (VSR) were both set to $\pm 3\sqrt{\lambda_i}$. We found that the mean shapes are very similar. In fact, the average point-point distance of mean shapes is below 0.001 pixels. The modes of the two PDMs also exhibit similar variations, except that the robust PDM is relatively compact than the conventional one.

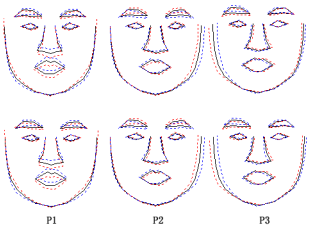
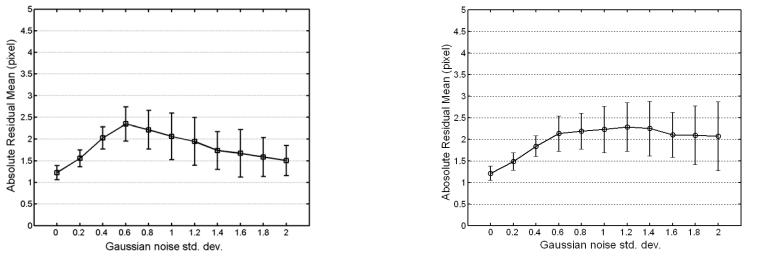


Fig. 6. The variations of first three modes for PDM of ASM (*top*) and R-ASM (*bottom*). Note that the resulting shape modes are very similar

Table 1. Efficiency comparison for the three methods. Mean (std. dev.) time of one re-weighting estimation iteration is given

Algorithm	Time (ms)
Gauss-Newton Algorithm	1.7 (0.6)
M-Estimation	8.6 (0.7)
Random Sampling	1983 (300)

R-ASM Robustness and Stability: To evaluate the robustness and stability of R-ASM, mean value of the point-point distance is compared respect to different initial outlier distance. We start with the 105 fully labeled test images. 25% of the landmark points in each labeled sample were chosen at random and their coordinates were perturbed by Gaussian noise. The size of the standard deviation σ used to create outliers was varied between 0 and 2 times the corresponding landmark's standard deviation in training set. We systematically compared the two methods: robust PDM with Gauss-Newton algorithm for parameter estimation and conventional PDM with parameters calculated in the closed form, referred as MAD-ASM. Fig. 7 shows absolute residual means and standard deviations against each initialization outlier σ for both methods. The result validates the reliability of the pre-computed scale parameter. It is worth noting that the R-ASM has better distance tolerant ability than MAD-ASM and it gets more accurate result as outlier distance increases.

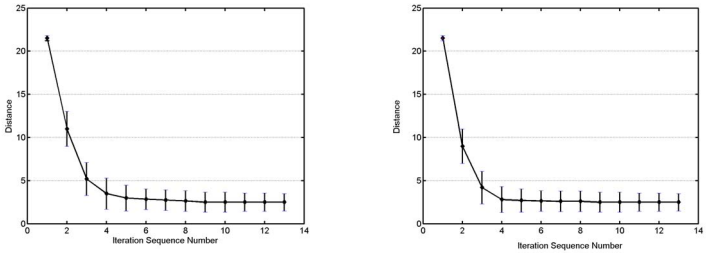


(a) R-ASM (b) MAD-ASM

Fig. 7. Robustness comparison. Absolute residual mean is plotted against initialization outlier σ as a ratio of corresponding landmark's std. dev in training set

R-ASM Efficiency: A main disadvantage for many robust estimators is inefficiency. While using our approach, computation cost can be considerably cut down, as can be seen from table 1. The experiments were conducted on the 105 testing images and 15% landmarks were made outliers. The Ransac resample subset size was set to 35% of the total landmarks. Only single re-weighting iteration of parameter estimation was recorded for comparison. As discussed above, it outperforms much better than previous methods in efficiency. All three algorithms were implemented in matlab on a P4 1.8G computer with 512M memories.

In fact, efficiency can also be obtained from robustness. We exploited the combination of R-ASM and Gabor feature based local image search, which is much like W-ASM [12]. Taking advantage of the accurate localization ability of Gabor wavelet features, the algorithm converges much quickly. In Fig. 8, the statistics of the average point-point distance of each step of search are compared between W-ASM and the proposed combination approach. The statistics were obtained from the 105 testing images in a 3 level image pyramid (resolution was reduced 1/2 level by level) with initial displacements about 20 pixels from the ground truth.



(a) W-ASM (b) R-ASM + Gabor wavelet search

Fig. 8. Mean error of point-point distance and standard deviation between searched result and ground truth in each iteration step

The result indicates that R-ASM converges very fast. It takes only 5 iterations to converge in average, while W-ASM scheme needs 8 iterations.

6 Conclusion

In this work, we present a novel approach to construct and fit ASM with outliers. The training set was assembled by automatic search using Gabor wavelet features. Then RPCA was applied on this training set containing outliers to obtain a robust PDM. Experiment shows that the outliers have little affect on the training process and the resulting PDM is quite similar to the convention one. In model fitting stage, Gauss-Newton scheme is adopted to efficiently optimize the robust energy function. The computational cost for one iteration of model parameter estimation is linear, while it is quadratic when using previous robust method. Experiments convinced the validity and efficiency of the optimization. In comparison to previous method that based on MAD estimation of scale parameter in each iteration stage, we make explicit use of fixed scale parameter learned from training set. This improves both efficiency and accuracy of model fitting, especially for large distance outliers.

Furthermore, taking advantage of the robustness of the R-ASM, image local search range can be extended and the algorithm converges rapidly.

Acknowledgement

This work is supported by the National Key Basic Research Project of China under Grant No. 2004CB318005.

References

1. T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham: Active Shape Models- their training and application. Computer Vision and Image Understanding, vol. 61. (1995) 38-59.
2. A. Martinez, Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. IEEE Trans. Patt. Anal. Mach. Intell, vol. 24. No. 6 (2002) 748-763.
3. S.Y. Ho and H. Huang. An Analytic Solution for the Pose Determination of Human Faces from a Monocular Image, Pattern Recognition Letters, Vol.19 (1998), 1045-1054.
4. J. Cohn, A. Zlochower, J.J. Lien, and T. Kanade: Feature-point tracking by optical flow discriminates subtle differences in facial expression. In: Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, (1998) 396-401.
5. M. Rogers and J. Gram: Robust Active Shape Search. In: Proceedings of the European Conference on Computer Vision (ECCV), (2002) 517-530.
6. H. Shum, K. Ikeuchi, and R. Reddy: Principal component analysis with missing data and its application to polyhedral object modeling. IEEE Trans. Patt Anal. Mach. Intell, vol. 17. No. 9 (1995) 854-867.
7. R. Gross, I. Matthews, and S. Baker: Constructing and Fitting Active Appearance Models with Occlusion. In: Proceedings of the IEEE Workshop on Face Processing in Video, (2004).
8. L. Wiskott, J.M. Fellous, N. Krüger, Cvd. Malsburg: Face Recognition by Elastic Graph Matching. Intelligent Biometric Techniques in Fingerprint and Face Recognition, eds. L.C. Jain et al., publ. CRC Press, Chapter 11, (1999) 355-396.
9. Black, M.J. and Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Computer Vision and Image Understanding, vol. 63 (1996) 75–104.

10. Fernando De la Torre and Michael J. Black: A Framework for Robust Subspace Learning. International Journal of Computer Vision. Vol. 54 (2003) 117-142
11. P. J. Huber: Robust Statistics. Wiley, New York, (1981).
12. Feng Jiao, Stan Li, Heung-Yeung Shum and Dale Schuurmans: Face Alignment Using Statistical Models and Wavelet Features. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1 (2003) 321-327.
13. L. Xu and A. Yuille: Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Transactions on Neural Networks, Vo. 6 No. 1 (1995) 131-143
14. T. N. Yang and S. D. Wang: Robust algorithms for principal component analysis. Pattern Recognition Letters, Vol. 20 No. 9 (1999) 927-933.

Comparative Assessment of Content-Based Face Image Retrieval in Different Color Spaces

Peichung Shih and Chengjun Liu

New Jersey Institute of Technology, Newark NJ 07102, USA
ps9@oak.njit.edu, liu@cs.njit.edu

Abstract. Content-based face image retrieval is concerned with computer retrieval of face images (of a given subject) based on the geometric or statistical features automatically derived from these images. It is well known that color spaces provide powerful information for image indexing and retrieval by means of color invariants, color histogram, color texture, etc.. This paper assesses comparatively the performance of content-based face image retrieval in different color spaces using a standard algorithm, the Principal Component Analysis (PCA), which has become a popular algorithm in the face recognition community. In particular, we comparatively assess 12 color spaces (RGB, HSV, YUV, $YCbCr$, XYZ, YIQ, $L^*a^*b^*$, $U^*V^*W^*$, $L^*u^*v^*$, $I_1I_2I_3$, HSI, and rgb) by evaluating 7 color configurations for every single color space. A color configuration is defined by an individual or a combination of color component images. Take the RGB color space as an example, possible color configurations are R, G, B, RG, RB, GB, and RGB. Experimental results using 1,800 FERET R, G, B images corresponding to 200 subjects show that some color configurations, such as R in the RGB color space and V in the HSV color space, help improve face retrieval performance.

1 Introduction

Content-based face image retrieval is concerned with computer retrieval of face images (of a given subject) based on the geometric or statistical features automatically derived from these images [1], [2]. Efficient retrieval requires a robust feature extraction method that has the ability to learn meaningful low-dimensional patterns in spaces of very high dimensionality. Low-dimensional representations are also important when one considers the intrinsic computational aspect. The Principal Component Analysis (PCA) [3] has been widely used to perform dimensionality reduction for face indexing and retrieval [4], [5], [6], [7]. In particular, PCA is the method behind the Eigenfaces coding scheme [8] whose primary goal is to project the similarity judgment for face recognition into a low-dimensional space. This space defines a feature space, or a "face space", which drastically reduces the dimensionality of the original space, and face detection and identification are carried out in this reduced face space.

It is well known that color spaces provide powerful information for image indexing and retrieval by means of color invariants, color histogram, color texture, etc.. Different color spaces, which are defined by means of transformations from the original RGB (red, green, blue) color space, display different color properties. The HSV (hue, saturation, value) color space and its variants, such as the HSI (hue, saturation, intensity)

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1039–1048, 2005.
© Springer-Verlag Berlin Heidelberg 2005

color space and the HLS (hue, lightness, saturation) color space, are often applied in locating and extracting facial features [9]. The $YCbCr$ (luminance, Chrominance-blue, Chrominance-red) color space, the YIQ (luminance, in-phase, quadrature) color space, and the YUV color space have wide applications in color clustering and quantization for skin color regions [9], [10]. The perceptually uniform color spaces, such as the CIE-$U^*V^*W^*$ color space, the CIE-$L^*u^*v^*$ color space, and the CIE-$L^*a^*b^*$ color space have general and ubiquitous applications [11], [12].

In this paper, we assess the performance of content-based face image retrieval in different color spaces using a standard algorithm, PCA [3]. Specifically, we assess comparatively 12 color spaces (RGB, HSV, YUV, $YCbCr$, XYZ, YIQ, $L^*a^*b^*$, $U^*V^*W^*$, $L^*u^*v^*$, $I_1I_2I_3$, HSI, and rgb) by evaluating 7 color configurations for every single color space. A color configuration is defined by an individual or a combination of color component images. Take the RGB color space as an example, possible color configurations are R, G, B, RG, RB, GB, and RGB.

2 Color Spaces

This section details the 12 color spaces assessed in this paper. The rgb color space is defined by projecting the R, G, B values onto the $R = G = B = max\{R, G, B\}$ plane, such that $r = R/(R + G + B)$, $g = G/(R + G + B)$, and $b = B/(R + G + B)$. The $I_1I_2I_3$ color space proposed by Ohta et al. [13] applies a Karhunen-Loeve transformation to decorrelate the RGB components. The linear transformation based on Ohta's experimental model is defined as: $I_1 = (R + G + B)/3$, $I_2 = (R - B)/2$, and $I_3 = (2G - R - B)/2$ [13].

The HSV and the HSI color spaces are motivated by the human vision system in the sense that human describes color by means of hue, saturation, and brightness. Let $MAX = max(R, G, B)$, $MIN = min(R, G, B)$, and $\delta = MAX - MIN$, the HSV color space is defined as follows [14]:

$$V = MAX; \quad S = \delta/MAX; \quad H = \begin{cases} 60(G - B)/\delta & \text{if } MAX = R \\ 60(B - R + 2\delta)/\delta & \text{if } MAX = G \\ 60(R - G + 4\delta)/\delta & \text{if } MAX = B \end{cases} \tag{1}$$

The HSI color space is specified as follows [15]:

$$I = (R + G + B)/3; \quad S = 1 - I * MIN; \quad H = \begin{cases} \theta & \text{if } B \leq G \\ 360 - \theta & \text{otherwise} \end{cases} \tag{2}$$

where $\theta = cos^{-1}\left\{\frac{1}{2}[(R - G) + (R - B)]/[(R - G)^2 + (R - B)(G - B)]^{\frac{1}{2}}\right\}$. Note that in both Eq. 1 and Eq. 2, the R, G, B values are scaled to [0,1].

The YUV and the YIQ color spaces are commonly used in video for transmission efficiency. The YIQ color space is adopted by the NTSC(National Television System Committee) video standard in reference to $RGB\ NTSC$, while the YUV color space is used by the PAL (Phase Alternation by Line) and the SECAM (System Electronique Couleur Avec Memoire). The YUV color space and the YIQ color space are specified as follows:

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ -0.1471 & -0.2888 & 0.4359 \\ 0.6148 & -0.5148 & -0.1000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$
$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5957 & -0.2745 & -0.3213 \\ 0.2115 & -0.5226 & 0.3111 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3}$$

The $YCbCr$ color space is a scaled and offset version of the YUV color space. The Y component has 220 levels ranging from 16 to 235, while the Cb, Cr components have 225 levels ranging from 16 to 240:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.4810 & 128.5530 & 24.9660 \\ -37.7745 & -74.1592 & 111.9337 \\ 111.9581 & -93.7509 & -18.2072 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{4}$$

where the R, G, B values are scaled to [0,1].

The CIE (Commission Internationale de l'Éclairage) perceptually uniform color spaces, such as the $U^*V^*W^*$, the $L^*u^*v^*$, and the $L^*a^*b^*$ color spaces, are defined based on the XYZ tristimulus:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.000 & 0.066 & 1.116 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5}$$

Note that the Y component defined here is consistent with the luminance defined in Eq. 3 or Eq. 4. In addition, a chromaticity diagram can be derived via the chromaticity coordinates x, y, which are specified by the X, Y, Z tristimulus. This CIE chromaticity diagram, however, is not perceptually uniform [16]. To overcome such a shortcoming, the CIE uv chromaticity diagram was proposed [16]:

$$\begin{aligned} u &= 4x/(-2x+12y+3) \quad or \quad 4X/(X+15Y+3Z) \\ v &= 6y/(-2x+12y+3) \quad or \quad 6Y/(X+15Y+3Z) \end{aligned} \tag{6}$$

Based on this uniform chromaticity scale (UCS), a CIE uniform color space $U^*V^*W^*$ was proposed. The W^* component corresponds to luminance, while the U^*, V^* components correspond to chrominance [16]:

$$W^* = \begin{cases} 116(\frac{Y}{Y_o})^{\frac{1}{3}} - 16 & \text{if } \frac{Y}{Y_o} > 0.008856 \\ 903.3(\frac{Y}{Y_o}) & \text{otherwise} \end{cases}; \; U^* = 13W^*(u - u_o); \; V^* = 13W^*(v - v_o) \tag{7}$$

where the u_o and v_o are derived from the reference white stimulus.

Although the CIE-uv diagram is perceptually uniform, it has its own deficiency in representing yellow-red colors as the area of yellow-red in the diagram is relatively small [17]. To improve this deficiency, a new $u'v'$ coordinate system is defined: $u' = u$, $v' = (3/2)v$. Based on this new coordinate system, two CIE uniform color spaces were defined, namely the CIE-$L^*u^*v^*$ color space and the CIE-$L^*a^*b^*$ color space [16]. The CIE-$L^*u^*v^*$ color space is proposed to obsolete the $U^*V^*W^*$ color space by

substituting W^*, U^*, V^*, u, v in Eq. 7 for L^*, u^*, v^*, u', v', respectively. The $L^*a^*b^*$ color space is modeled based on human vision system and is defined as follows:

$$L^* = 116f(\frac{Y}{Y_o}) - 16; \quad a^* = 500[f(\frac{X}{X_o}) - f(\frac{Y}{Y_o})]; \quad b^* = 200[f(\frac{Y}{Y_o}) - f(\frac{Z}{Z_o})] \tag{8}$$

where $f(x) = x^{\frac{1}{3}}$ if $x > 0.008856$; $f(x) = 7.787x + \frac{16}{116}$ otherwise.

3 Principal Component Analysis and Classification Rule

PCA is a standard decorrelation technique and following its application one derives an orthogonal projection basis that directly leads to dimensionality reduction and feature extraction. Let $X \in \mathbb{R}^N$ be a random vector representing an image, and $\Sigma_X \in \mathbb{R}^{N \times N}$ be the covariance matrix of X. The PCA procedure factorizes the Σ_X into the form: $\Sigma_X = \Phi \Lambda \Phi^t$, where Φ is an orthogonal eigenvector matrix and Λ a diagonal eigenvalue matrix with diagonal elements in decreasing order.

An important property of PCA is its optimal signal reconstruction in the sense of minimum Mean Square Error (MSE) when only a subset of principal components is used to represent the original signal. Following this property, an immediate application of PCA is the dimensionality reduction by projecting a random vector X onto the eigenvectors: $Y = P^t X$, where $P \in \mathbb{R}^{N \times m}$ is a subset of eigenvector matrix Φ and $m < N$. The lower dimensional vector $Y \in \mathbb{R}^m$ captures the most expressive features of the original data X.

After dimensionality reduction, feature vectors are compared and classified by the nearest neighbor (to the mean) rule using a similarity (distance) measure δ: $\delta(\mathcal{Y}, \mathcal{M}_k) = \min_j \delta(\mathcal{Y}, \mathcal{M}_j) \longrightarrow \mathcal{Y} \in \omega_k$, where $\mathcal{Y}$ is a testing feature vector and $\mathcal{M}_k^0, k = 1, 2, \ldots, L$ is the mean of the training samples for class ω_k. The testing feature vector, $\mathcal{Y}$, is classified as belonging to the class of the closest mean, $\mathcal{M}_k$, using the similarity measure δ.

The similarity measures used in our experiments to evaluate the efficiency of different representation and recognition methods include the L_1 distance measure, δ_{L_1}, the L_2 distance measure, δ_{L_2}, the Mahalanobis distance measure, δ_{Md}, and the cosine similarity measure, δ_{cos}, which are defined as follows:

$$\begin{aligned} \delta_{L_1}(\mathcal{X}, \mathcal{Y}) &= \textstyle\sum_i |\mathcal{X}_i - \mathcal{Y}_i| \\ \delta_{L_2}(\mathcal{X}, \mathcal{Y}) &= (\mathcal{X} - \mathcal{Y})^t (\mathcal{X} - \mathcal{Y}) \\ \delta_{Md}(\mathcal{X}, \mathcal{Y}) &= (\mathcal{X} - \mathcal{Y})^t \Sigma^{-1} (\mathcal{X} - \mathcal{Y}) \\ \delta_{cos}(\mathcal{X}, \mathcal{Y}) &= \frac{-\mathcal{X}^t \mathcal{Y}}{\|\mathcal{X}\| \, \|\mathcal{Y}\|} \end{aligned} \tag{9}$$

where Σ is the covariance matrix, and $\| \cdot \|$ denotes the norm operator. Note that the cosine similarity measure includes a minus sign because the nearest neighbor (to the mean) rule applies minimum (distance) measure rather than maximum similarity measure [18].

4 Experiments

This section assesses the performance of content-based face image retrieval in the 12 color spaces defined in Sect. 2. The 1,800 images from the FERET database [19] are

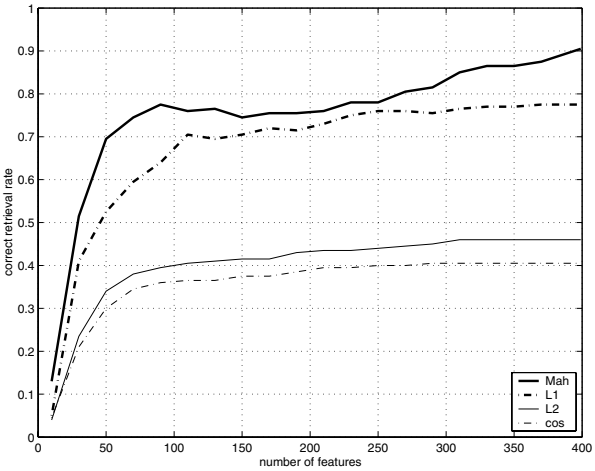


Fig. 1. The performance of content-based face image retrieval using the PCA method on the intensity images derived by averaging the R, G, B color components. The similarity measures applied are the Mahalanobis distance measure (Mah.), the L_1 distance measure (L1), the L_2 distance measure (L2), and the cosine similarity measure (cos)

used for our experiments. The images correspond to 200 subjects such that each subject has 9 images (3 sets of R, G, B images). As there are three sets of images for each subject, two sets are randomly chosen for training, while the remaining set (unseen during training) is used for testing. Note that all images are normalized to the size of 128×128 to extract facial regions that contain only faces, so that the performance of face retrieval is not affected by the factors not related to face, such as hair style.

To provide a baseline performance for comparison, our first set of experiments applies different similarity measures as defined in Sect. 3 on the intensity images derived by averaging the R, G, B color components. Fig. 1 shows the performance of content-based face image retrieval using the PCA method as detailed in Sect. 3. The horizontal axis indicates the number of features used, and the vertical axis represents the correct retrieval rate, which is the accuracy rate for the top response being correct. Fig. 1 shows that the Mahalanobis distance measure performs the best, followed in order by the L_1 distance measure, the L_2 distance measure, and the cosine similarity measure. The experimental results provide a baseline face retrieval performance based on the intensity images, and suggest that one should use the Mahalanobis distance measure for the comparative assessment in different color spaces.

We now assess comparatively content-based face image retrieval in the 12 different color spaces as defined in Sect. 2. For each color space, we define 7 color configurations by means of an individual or a combination of color component images. Take the RGB color space as an example, possible color configurations are R, G, B, RG, RB, GB, and RGB. Note that when two or three color component images are used to define a color configuration, each color component image is first normalized to zero mean and unit variance, and then the normalized color component images are concatenated to form an augmented vector representing the color configuration.

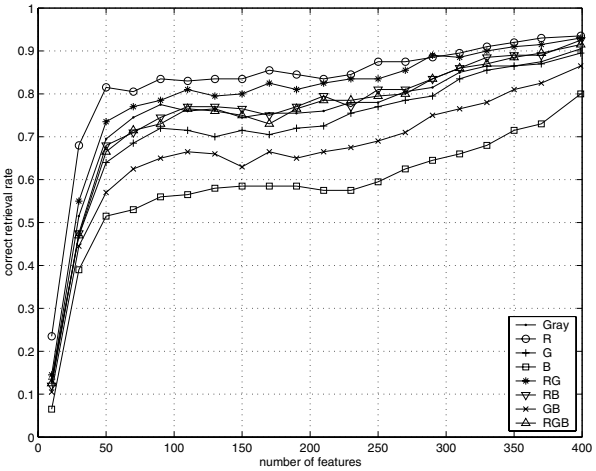


Fig. 2. Content-based face image retrieval performance of the 7 color configurations in the RGB color space. Note that the performance curve of the intensity images (Gray) is also included for comparison (same in the following figures)

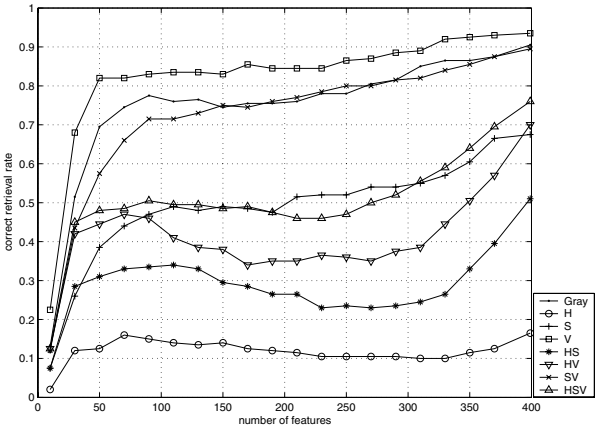


Fig. 3. Content-based face image retrieval performance of the 7 color configurations in the HSV color space

Our next set of experiments assesses the following color spaces: RGB, HSV, YUV, $YCbCr$, XYZ, YIQ, $L^*a^*b^*$, $U^*V^*W^*$, and $L^*u^*v^*$. The face retrieval performance of the 7 color configurations in these color spaces is shown in Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively. Note that the YUV and the $YCbCr$ color spaces (as well as the $U^*V^*W^*$ and the $L^*u^*v^*$ color spaces) have identical face retrieval performance due to their definitions (see Sect. 2). In particular, Fig. 2 shows that the R and the RG color configurations perform better than the intensity images. Fig. 3 shows that the V color configuration outperforms the intensity images. Fig. 4 shows that the Y and YV color configurations in the YUV color space or the Y and YCr color configurations in the $YCbCr$ color space have better face retrieval perfor-

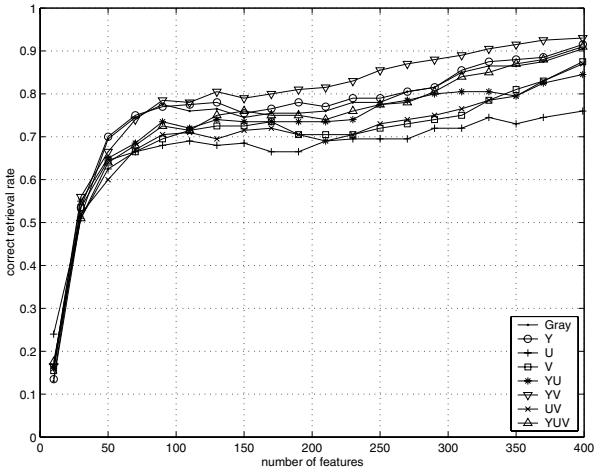


Fig. 4. Content-based face image retrieval performance of the 7 color configurations in the YUV color space. Note that the face retrieval results in the $YCbCr$ color space are the same when the Y, U, and V color components are replaced by their counterparts Y, Cb, and Cr, respectively

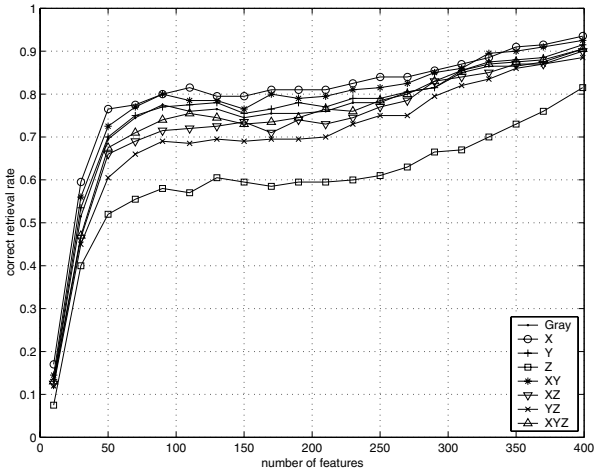


Fig. 5. Content-based face image retrieval performance of the 7 color configurations in the XYZ color space

mance than the intensity images. Fig. 5 shows the X, Y, and XY color configurations perform better than the intensity images. Fig. 6 shows that the Y and YI color configurations outperform the intensity images. Fig. 7 shows that the L^*, L^*a^*, L^*b^*, and $L^*a^*b^*$ color configurations are better than the intensity images for face retrieval. Fig. 8 shows that the W^* and U^*W^* color configurations in the $U^*V^*W^*$ color space or the L^* and L^*u^* color configurations in the $L^*u^*v^*$ color space perform better than the intensity images.

The color configurations, which perform better than the intensity images for face retrieval, are summarized in Table 1. Note that those color configurations with better face

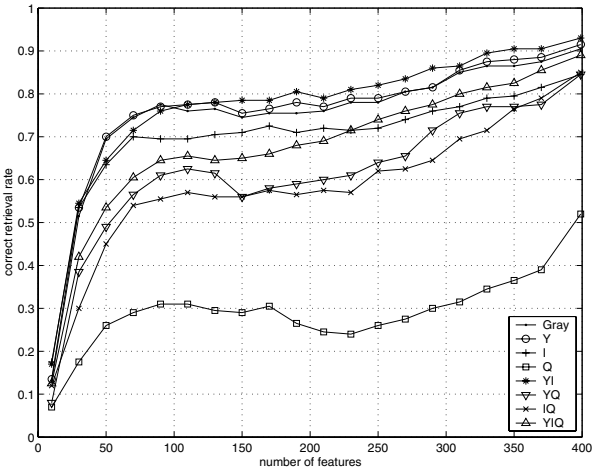


Fig. 6. Content-based face image retrieval performance of the 7 color configurations in the YIQ color space

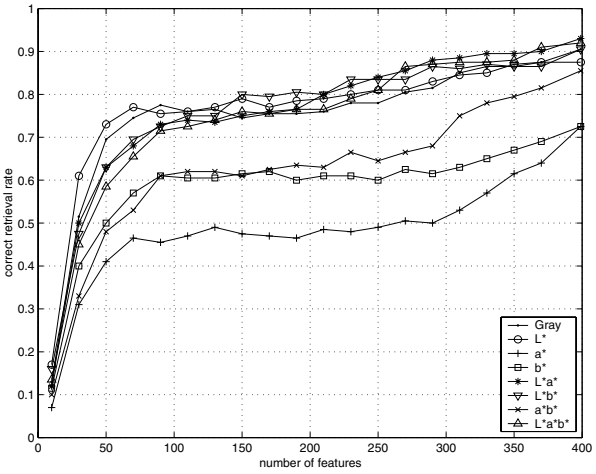


Fig. 7. Content-based face image retrieval performance of the 7 color configurations in the $L^*a^*b^*$ color space

retrieval performance shown in Table 1 all share one common characteristic: they contain both chromatic and achromatic (intensity) components. The pure chromatic color configurations, however, all display worse (than the intensity images) face retrieval performance. Specifically, these pure chromatic color configurations include the HS in Fig. 3, the UV and the $CbCr$ in Fig. 4, the IQ in Fig. 6, the a^*b^* in Fig. 7, and the U^*V^* and the u^*v^* in Fig. 8. Note also that simply applying all the color components does not necessarily achieve the best face retrieval performance. The reason for this finding is that some color configurations, such as the B in the RGB color space (Fig. 2), the Z in the XYZ color space (Fig. 5), and the pure chromatic color configurations discussed above, all perform worse than the intensity images for content-based

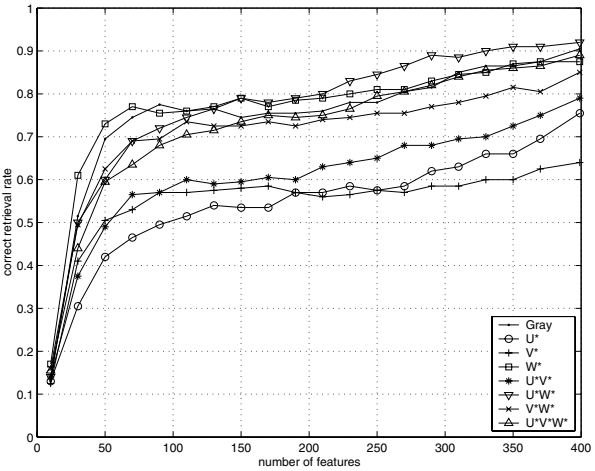


Fig. 8. Content-based face image retrieval performance of the 7 color configurations in the $U^*V^*W^*$ color space. Note that the face retrieval results in the $L^*u^*v^*$ color space are the same when the U^*, V^*, and W^* color components are replaced by their counterparts u^*, v^*, and L^*, respectively

Table 1. The color configurations, which perform better than the intensity images for face retrieval, in the color spaces: RGB, HSV, YUV, $YCbCr$, XYZ, YIQ, $L^*a^*b^*$, $U^*V^*W^*$, and $L^*u^*v^*$

color space	color configurations with better face retrieval performance
RGB	R, RG
HSV	V
YUV / $YCbCr$	Y, YV / Y, YCr
XYZ	X, Y, XY
YIQ	Y, YI
$L^*a^*b^*$	$L^*, L^*a^*, L^*b^*, L^*a^*b^*$
$U^*V^*W^*$ / $L^*u^*v^*$	W^*, U^*W^* / L^*, L^*u^*

face image retrieval. We have experimented with the $I_1I_2I_3$, HSI, and rgb color spaces as well, but the experimental results show that the color configurations in these color spaces do not improve face retrieval performance.

5 Conclusion

We have assessed comparatively the performance of content-based face image retrieval in 12 color spaces using a standard algorithm, the PCA, which has become a popular algorithm in the face recognition community. In particular, we have comparatively assessed the RGB, HSV, YUV, $YCbCr$, XYZ, YIQ, $L^*a^*b^*$, $U^*V^*W^*$, $L^*u^*v^*$, $I_1I_2I_3$, HSI, and rgb color spaces by evaluating 7 color configurations for every single

color space. Experimental results using 1,800 FERET R, G, B images corresponding to 200 subjects show that some color configurations, such as R in the RGB color space and V in the HSV color space, help improve face retrieval performance.

Acknowledgments

This work was partially supported by the TSWG R&D Contract N41756-03-C-4026.

References

1. Zhao, W., Chellappa, R., Phillips, J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys **35** (2003) 399–458
2. Stiefelhagen, R., Yang, J., Waibel, A.: Modeling focus of attention for meeting index based on multiple cues. IEEE Trans. Neural Networks **13** (2002) 928–938
3. Fukunaga, K.: Introduction to Statistical Pattern Recognition. second edn. Academic Press (1990)
4. Liu, C.: Gabor-based kernel PCA with fractional power polynomial m odels for face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence **26** (2004) 572–581
5. Liu, C., Wechsler, H.: Robust coding schemes for indexing and retrieval from large face databases. IEEE Trans. on Image Processing **9** (2000) 132–137
6. Zhao, W., Chellappa, R.: Symmetric shape-from-shading using self-ratio image. Int. Journal Computer Vision **45** (2001) 55–75
7. Yu, H., Yang, J.: A direct lda algorithm for high-dimensional data - with application to face recognition. Pattern Recognition **34** (2001) 2067–V2070
8. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience **13** (1991) 71–86
9. Garcia, C., Tziritas, G.: Face detection using quantized skin color regions merging and wavelet packet analysis. IEEE Trans. Multimedia **1** (1999) 264–277
10. Habili, N., Lim, C.: Hand and face segmentation using motion and color cues in digital image sequences. In: Proc. IEEE International Conference on Multimedia and Expo 2001, Tokyo, Japan (2001)
11. Wu, H., Chen, Q., Yachida, M.: Face detection from color images using a fuzzy pattern matching method. IEEE Trans. Pattern Analysis and Machine Intelligence **21** (1999) 557–563
12. Terrillon, J., Shirazi, M., Fukamachi, H., Akamatsu, S.: Comparative performance of different skin chrominance models and chrominance space for the automatic detection of human faces in color images. In: Proc. The Fourth International Conference on Face and Gesture Recognition, Grenoble, France (2000)
13. Ohta, Y.: Knowledge-Based Interpretation of Outdoor Natural Color Scenes. Pitman Publishing, London (1985)
14. Smith, A.: Color gamut transform pairs. Computer Graphics **12** (1978) 12–19
15. Gonzalez, R., Woods, R.: Digital Image Processing. Prentice Hall (2001)
16. Judd, D., Wyszecki, G.: Color in Business, Science and Industry. John Wiley & Sons, Inc. (1975)
17. Chamberlin, G., Chamberlin, D.: Colour: Its Measurement, Computation and Application. Heyden & Son, London (1980)
18. Moon, H., Phillips, P.: Analysis of pca-based face recognition algorithms. In Bowyer, K.W., Phillips, P.J., eds.: Empirical Evaluation Techniques in Computer Vision, Wiley-IEEE Computer Society (1998)
19. Phillips, P., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing **16** (1998) 295–306

A Principled Approach to Score Level Fusion in Multimodal Biometric Systems

Sarat C. Dass[1], Karthik Nandakumar[2], and Anil K. Jain[2]

[1] Department of Statistics and Probability
Michigan State University, MI - 48824, USA
sdass@stt.msu.edu
[2] Department of Computer Science and Engineering
Michigan State University, MI - 48824, USA
{nandakum,jain}@cse.msu.edu

Abstract. A multimodal biometric system integrates information from multiple biometric sources to compensate for the limitations in performance of each individual biometric system. We propose an optimal framework for combining the matching scores from multiple modalities using the likelihood ratio statistic computed using the generalized densities estimated from the genuine and impostor matching scores. The motivation for using generalized densities is that some parts of the score distributions can be discrete in nature; thus, estimating the distribution using continuous densities may be inappropriate. We present two approaches for combining evidence based on generalized densities: (i) the product rule, which assumes independence between the individual modalities, and (ii) copula models, which consider the dependence between the matching scores of multiple modalities. Experiments on the MSU and NIST multimodal databases show that both fusion rules achieve consistently high performance without adjusting for optimal weights for fusion and score normalization on a case-by-case basis.

Keywords: Biometric recognition, multimodal biometric systems, fusion, Gaussian copula models, Generalized densities, Neyman-Pearson theorem.

1 Introduction

Biometrics refers to the automatic identification of an individual based on his/her physiological traits [1]. Biometric systems based on a single source of information (unimodal systems) suffer from limitations like the lack of uniqueness, non-universality and noisy data [2] and hence, may not be able to achieve the desired performance requirements of real-world applications. In contrast, multimodal biometric systems combine information from its component modalities to arrive at a decision [3]. Several studies [4–8] have demonstrated that by consolidating information from multiple sources, better performance can be achieved compared to the individual unimodal systems. In a multimodal biometric system, integration can be done at (i) feature level, (ii) matching score level, or (iii) decision level. Matching score level fusion is commonly preferred because matching scores are easily available and contain sufficient information to distinguish

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1049–1058, 2005.
© Springer-Verlag Berlin Heidelberg 2005

between a genuine and an impostor case. Given a number of biometric systems, one can generate matching scores for a pre-specified number of users even without knowing the underlying feature extraction and matching algorithms of each biometric system. Thus, combining information contained in the matching scores seems both feasible and practical.

We propose a framework for optimally combining the matching scores from multiple modalities based on generalized densities estimated from the genuine and impostor matching scores. The motivation for using generalized densities is that some parts of the score distributions can be discrete in nature. As a result, estimating the densities using continuous density functions can be inappropriate. We present two approaches for combining evidence based on generalized densities: (i) the product rule, which assumes independence between the individual modalities, and (ii) copula models, which parametrically model the dependence between the matching scores of multiple modalities. Our proposed method bypasses the need for score normalization and selection of optimal weights for the score combination on a case-by-case basis [3, 9, 10], and therefore, is a more principled approach with performance comparable to the commonly used fusion methods. Experiments have shown that our method achieves consistently high performance over the MSU and NIST multimodal databases.

2 Generalized Densities

2.1 Estimation of Marginal Distributions

Let X be a generic matching score with distribution function F, i.e., $P(X \leq x) = F(x)$. We denote the genuine (impostor) matching score by X_{gen} (X_{imp}) and the corresponding distribution function by F_{gen} (F_{imp}). Assuming that $F_{gen}(x)$ and $F_{imp}(x)$ have densities $f_{gen}(x)$ and $f_{imp}(x)$, respectively, the Neyman-Pearson theorem states that the *optimal* ROC curve is the one corresponding to the likelihood ratio statistic $NP(x) = f_{gen}(x)/f_{imp}(x)$ [11]. The ROC curve corresponding to $NP(x)$ has the highest genuine accept rate (GAR) for every given value of the false accept rate (FAR) compared to any other statistic $U(x) \neq NP(x)$ (this is true even for the original matching scores corresponding to $U(x) = x$).

However, when $f_{gen}(x)$ and $f_{imp}(x)$ are unknown (which is typically the case) and are estimated from the observed matching scores, the ROC corresponding to $NP(x)$ may turn out to be suboptimal. This is mainly due to the large errors in the estimation of $f_{gen}(x)$ and $f_{imp}(x)$. Thus, for a set of genuine and impostor matching scores, it is important to be able to estimate $f_{gen}(x)$ and $f_{imp}(x)$ reliably and accurately. Previous studies by Griffin [11] and Prabhakar et al. [12] assume that the distribution function F has a continuous density with no discrete components. In reality, most matching algorithms apply thresholds at various stages in the matching process. When the required threshold conditions are not met, specific matching scores are output by the matcher (e.g., some fingerprint matchers produce a score of zero if the number of extracted minutiae is less than a threshold). This leads to discrete components in the matching score distribution that cannot be modeled accurately using a continuous density

function. A score value x_0 is said to be discrete if $P(X = x_0) = p > 0$. It is easy to see that F cannot be represented by a density function in a neighborhood of x_0 (since this would imply that $P(X = x_0) = 0$). Thus, discrete components need to be detected and modeled separately to avoid large errors in estimating $f_{gen}(x)$ and $f_{imp}(x)$. Our approach consists of detecting discrete components in the genuine and impostor matching score distributions, and then modeling the observed distribution of matching scores as a mixture of discrete and continuous components. Hence, this approach generalizes the work of [11, 12].

The following methodology can model a distribution based on a generic set of observed scores. For a fixed threshold T, the discrete values are identified as those values x_0 with $P(X = x_0) > T$, where $0 \leq T \leq 1$. Since the underlying matching score distribution is unknown, we estimate the probability $P(X = x_0)$ by $\frac{N(x_0)}{N}$, where $N(x_0)$ is the number of observations in the data set that equals x_0, and N is the total number of observations. The collection of all discrete components for a matching score distribution will be denoted by

$$\mathcal{D} \equiv \{\, x_0 \,:\, \frac{N(x_0)}{N} > T\}. \tag{1}$$

The discrete components constitute a proportion $p_D \equiv \sum_{x_0 \in \mathcal{D}} \frac{N(x_0)}{N}$ of the total observations. We obtain the collection $\mathcal{C}$ by removing all discrete components from the entire data set. The scores in $\mathcal{C}$ constitute a proportion $p_C \equiv 1 - p_D$ of the entire data set, and they are used to estimate the continuous component of the distribution ($F_C(x)$) and the corresponding density ($f_c(x)$). A non-parametric kernel density estimate of $f_c(x)$ is obtained from $\mathcal{C}$ as follows. The empirical distribution function for the observations in $\mathcal{C}$ is computed as

$$\hat{F}_C(x) = \frac{1}{N_C} \sum_{s \in \mathcal{C}} I\{\, s \leq x \,\}, \tag{2}$$

where $I\{s \leq x\} = 1$ if $s \leq x$, and $= 0$, otherwise; also, $N_C \equiv N\, p_C$. Note that $\hat{F}_C(x) = 0 \;\forall\; x < s_{min}$ and $\hat{F}_C(x) = 1 \;\forall\; x \geq s_{max}$, where s_{min} and s_{max}, respectively, are the minimum and maximum of the observations in $\mathcal{C}$. For values of $x, s_{min} < x < s_{max}$, not contained in $\mathcal{C}$, $\hat{F}_C(x)$, is obtained by linear interpolation. Next, B samples are simulated from $\hat{F}_C(x)$, and the density estimate of $f_C(x)$, $\hat{f}_C(x)$, is obtained from the simulated samples using a Gaussian kernel density estimator. The optimal bandwidth, h, is obtained using the "solve-the-equation" bandwidth estimator [13], which is an automatic bandwidth selector that prevents oversmoothing and preserves important features of the distribution of matching scores (see Figure 1). The generalized density is defined as

$$l(x) = p_C\, \hat{f}_C(x) + \sum_{x_0 \in \mathcal{D}} \frac{N(x_0)}{N} \cdot I\{x = x_0\}, \tag{3}$$

where $I\{x = x_0\} = 1$ if $x = x_0$, and $= 0$, otherwise. The distribution function corresponding to the generalized density is defined as

$$L(x) = p_C \int_{-\infty}^{x} \hat{f}_C(u)\, du + \sum_{x_0 \in \mathcal{D},\, x_0 \leq x} \frac{N(x_0)}{N}. \tag{4}$$

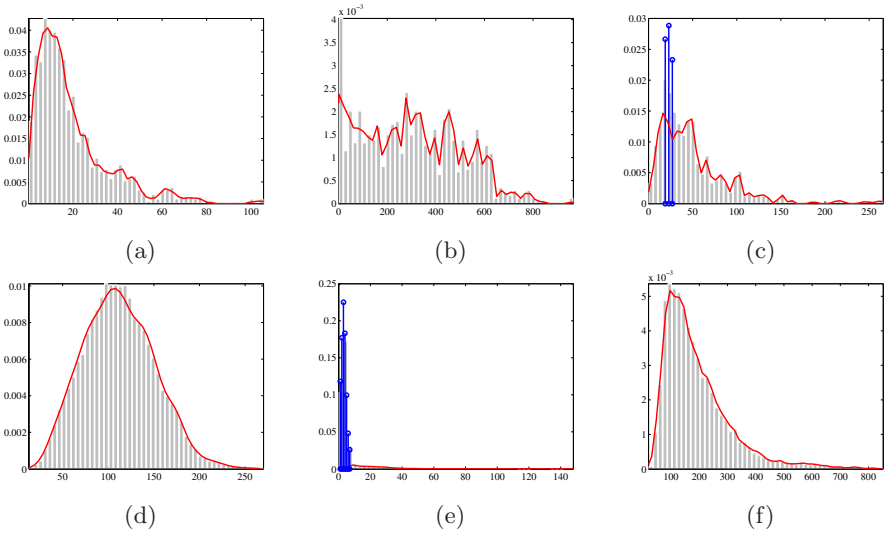


Fig. 1. Histograms of matching scores and corresponding generalized density estimates. Row 1: Histograms of genuine scores for face (a), finger (b), and hand-geometry (c). Row 2: Histograms of impostor scores for face (d), finger (e), and hand-geometry (f). The solid line is the estimated density using the kernel density estimator, and the spikes in (c) and (e) correspond to detected discrete components. Note that no pre-processing of the matching score data (including the conversion of distance measures into similarity scores) was performed before density estimation

For a multimodal system with K modalities, the generalized densities and distributions estimated for the genuine (impostor) scores for the k^{th} modality will be denoted by $l_{gen,k}(x)$ and $L_{gen,k}(x)$ ($l_{imp,k}(x)$ and $L_{imp,k}(x)$), respectively, for $k = 1, 2, \ldots, K$. Figures 1 (a)-(f) give the plots of $l_{gen,k}(x)$ and $l_{imp,k}(x)$ for the distribution of observed genuine and impostor matching scores for $K = 3$ modalities of the MSU-Multimodal database (see Section 4). Figures 1 (a)-(f) also give the histograms of the genuine and impostor matching scores for the three modalities. The "spikes" (see Figure 1 (c) and (e)) represent the detected discrete components and have a height greater than the threshold $T = 0.02$. Note that the individual "spikes" cannot be represented by a continuous density function. Forcing a continuous density estimate for these values will result in gross estimation errors and yield suboptimal ROC curves.

2.2 Joint Density Estimation Using Copula Models

The methodology described in Section 2.1 only estimates the marginal score distributions of each of the K modalities without estimating the joint distribution. One way to estimate the joint distribution of matching scores is by using copula models [14]. Let $H_1, H_2, \ldots, H_K$ be K continuous distribution functions on the real line and H be a K-dimensional distribution function with the k^{th} marginal given by H_k for $k = 1, 2, \ldots, K$. According to Sklar's Theorem [14], there exists a unique function $C(u_1, u_2, \ldots, u_K)$ from $[0, 1]^K$ to $[0, 1]$ satisfying

$$H(s_1, s_2, \ldots, s_K) = C(H_1(s_1), H_2(s_2), \ldots, H_K(s_K)), \tag{5}$$

where $s_1, s_2, \ldots, s_K$ are K real numbers. The function C is known as a K-copula function that "couples" the one-dimensional distributions functions $H_1, H_2, \ldots, H_K$ to obtain the K-variate function H. Equation (5) can also be used to construct K-dimensional distribution functions H whose marginals are the distributions $H_1, H_2, \ldots, H_K$: choose a copula function C and define H as in (5).

Copula functions are effective in modeling the joint distribution when the marginal distributions are non-normal and do not have a parametric form (as is usually the case for biometric data, see Figure 1). The family of copulas considered in this paper is the K-dimensional multivariate Gaussian copulas [15]. These functions can represent a variety of dependence structures using a $K \times K$ correlation matrix R. The K-dimensional Gaussian copula function with correlation matrix R is given by

$$C_R^K(u_1, u_2, \ldots, u_K) = \Phi_R^K(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_K)) \tag{6}$$

where each $u_k \in [0, 1]$ for $k = 1, 2, \ldots, K$, $\Phi(\cdot)$ is the distribution function of the standard normal, $\Phi^{-1}(\cdot)$ is its inverse, and Φ_R^K is the K-dimensional distribution function of a random vector $\mathcal{Z} = (Z_1, Z_2, \ldots, Z_K)^T$ with component means and variances given by 0 and 1, respectively. The (m, n)-th entry of R, ρ_{mn}, measures the degree of correlation between the m-th and n-th components for $m, n = 1, 2, \ldots, K$. In practice, ρ_{mn} will be unknown and hence, will be estimated using the product moment correlation of normal quantiles corresponding to the observed scores from the K modalities.

We denote the density of C_R^K by

$$\begin{aligned} c_R^K(u_1, u_2, \ldots, u_K) &\equiv \frac{\partial C_R^K(u_1, u_2, \ldots, u_K)}{\partial u_1 \partial u_2 \ldots \partial u_K} \\ &= \frac{\phi_R^K(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_K))}{\prod_{k=1}^K \phi(\Phi^{-1}(u_k))}, \end{aligned} \tag{7}$$

where $\phi_R^K(x_1, x_2, \ldots, x_K)$ is the joint probability density function of the K-variate normal distribution with mean 0 and covariance matrix R, and $\phi(x)$ is the standard normal density function. We will assume that the joint distribution function of genuine (impostor) matching scores for K modalities, F_{gen}^K (F_{imp}^K), is of the form (5) for some correlation matrix R_0 (R_1). For the genuine (impostor) case, H_k will be estimated by $L_{gen,k}(x)$ ($L_{imp,k}(x)$) for $k = 1, 2, \ldots, K$.

3 Fusion Based on Generalized Densities

Two methods of fusion have been considered in this paper. The first method assumes independence between the K biometric modalities and combines the estimated marginal densities using the product rule. For the matching score set $\mathcal{S} = (S_1, S_2, \ldots, S_K)$, the product fusion score of $\mathcal{S}$, $PFS(\mathcal{S})$, is given by

$$PFS(\mathcal{S}) = \prod_{k=1}^{K} \frac{l_{gen,k}(S_k)}{l_{imp,k}(S_k)}, \tag{8}$$

where $l_{gen,k}(\cdot)$ and $l_{imp,k}(\cdot)$ are the estimates of generalized densities of the genuine and impostor scores of the k^{th} biometric modality.

The copula fusion rule combines the individual modalities using the estimated Gaussian copula functions for the score distributions. The copula fusion score of a matching score set $\mathcal{S} = (S_1, S_2, \ldots, S_K)$, $CFS(\mathcal{S})$, is given by $CFS(\mathcal{S}) =$

$$PFS(\mathcal{S}) \cdot \frac{c^K_{R_0}(\Phi^{-1}(L_{gen,1}(S_1)), \Phi^{-1}(L_{gen,2}(S_2)), \ldots, \Phi^{-1}(L_{gen,K}(S_K)))}{c^K_{R_1}(\Phi^{-1}(L_{imp,1}(S_1)), \Phi^{-1}(L_{imp,2}(S_2)), \ldots, \Phi^{-1}(L_{imp,K}(S_K)))}, \quad (9)$$

where $L_{gen,k}(S_k)$ and $L_{imp,k}(S_k)$ are, respectively, the estimates of generalized distribution functions for the k^{th} biometric modality, and c^K_R is the density of C^K_R as defined in (7). This fusion rule assumes that the Gaussian copula functions can adequately model the dependence between the K biometric modalities.

4 Experimental Results

Experiments on fusion of matching scores using rules (8) and (9) were carried out on two different multimodal databases. For each experiment, 70% of the genuine and impostor matching scores were randomly selected to be the training set for the estimation of the generalized densities and the correlation matrices. The remaining 30% of the genuine and impostor scores were used to generate the ROC curves. This training-testing partition was repeated 20 times and the performance results reported for each value of FAR are the median GAR values.

4.1 Databases

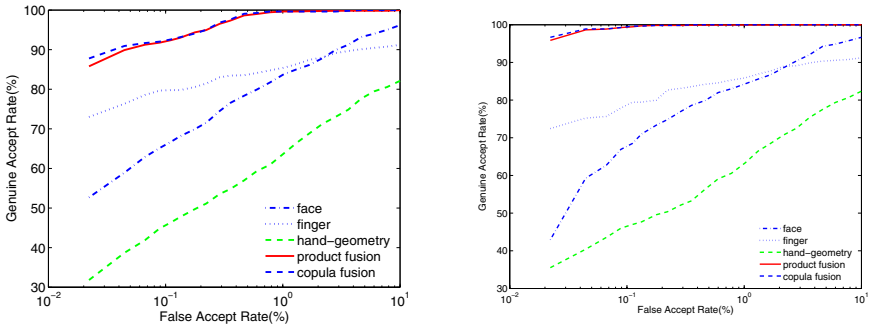
Table 1 summarizes the multimodal databases used in our experiments. The first database (referred to as the MSU-Multimodal database) consisted of 100 "virtual" subjects each providing five samples of face, fingerprint (left-index) and hand-geometry modalities. Face images were represented as eigenfaces [16] and the Euclidean distance between the eigen coefficients of the template-query pair was used as the distance metric. Minutia points were extracted from fingerprint images and the elastic string matching technique [17] was used for computing the similarity between two minutia point patterns. Fourteen features describing the geometry of the hand shape [18] were extracted from the hand images and Euclidean distance was computed for each template-query pair.

Table 1. Summary of Multimodal Databases Used

Database	Modalities	K	No. of Users
MSU-Multimodal	Fingerprint, Face, Hand-geometry	3	100
NIST-Multimodal	Fingerprint (Two fingers), Face (Two matchers)	4	517

Experiments were also conducted on the first partition of the Biometric Scores Set - Release I (BSSR1) released by NIST [19]. The NIST-Multimodal

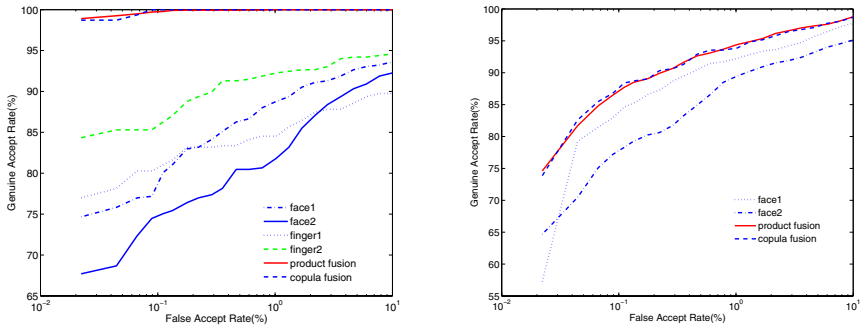
database consists of 517 users and is "truly multimodal" in the sense that the fingerprint and face images used for genuine matching score computation came from the same individual. One fingerprint score was obtained by comparing a pair of impressions of the left index finger and another score was obtained by comparing impressions of the right index finger. Two different face matchers were applied to compute the similarity between frontal face images. Even though the number of subjects in the NIST database is relatively large, there are only two samples per subject. So the number of genuine scores is still rather small.



(a) (b)

Fig. 2. Performance of product and copula fusion on the MSU-Multimodal database based on (a) continuous and (b) generalized density estimates

Figure 2 gives the ROC curves for the two fusion rules and the ROC curves based on the matching scores of individual modalities for the MSU-Multimodal database. Figure 2(a) shows the recognition performance when the genuine and impostor score distributions of the three modalities are modeled purely by continuous densities. The performance improvement obtained by modeling the matching score distributions as a mixture of discrete and continuous components (generalized densities) can be observed by comparing Figures 2(a) and 2(b). The ROC curves for the two fusion rules on the NIST-Multimodal database are shown in Figure 3(a). We see that both fusion rules give significantly better matching performance compared to the best single modality in each database. We also observe that the best single modality in both the databases is uncorrelated to the other modalities. For the MSU-Multimodal database, the estimates of the correlation of the best single modality (fingerprint) with the other two modalities (face and hand-geometry) are -0.01 and -0.11 for the genuine scores, and -0.05 and -0.04 for the impostor scores. For the NIST-Multimodal database (the best single modality is finger 2), the correlation estimates (with face1, face2, and finger1 modalities, respectively) are -0.02, -0.06, and 0.43 for the genuine cases and 0.04, 0.02, and 0.14 for the impostor cases. Since the fusion is driven mostly by the best modality, the fact that this modality is approximately independent of the others means that the product and copula fusion rules should be comparable to each other as reflected by the ROC curves in Figures 2 and 3(a).



(a) (b)

Fig. 3. ROC curves for the NIST-Multimodal database; (a) all four modalities (b) only face1 and face2 modalities

In order to study the usefulness of the copula fusion rule, we analyzed the fusion results of the face1 and face2 modalities of the NIST-Multimodal database (see Figure 3(b)). This pair had the highest degree of correlation among all pairs in the two databases (0.75 and 0.29 for the genuine and impostor scores, respectively). We observed that even in this case, the performance difference between the product and copula fusion rules is not significant. This may be due to the fact that although incorporating the correlation between the multiple matching scores into the fusion rule should result in better performance than fusion based on the independence assumption, the difference will be significant only in a few cases. The following simulations illustrate this fact. Let the matching scores of two biometric modalities follow the bivariate normal distribution with the following parameters (these values were chosen so as to closely model the matching scores of face1 and face2 modalities in the NIST-Multimodal database):

$$S_{gen} \sim N\left(\mu_{gen} = \begin{bmatrix} 0.72 \\ 76.78 \end{bmatrix}, \Sigma_{gen} = \begin{bmatrix} 0.006 & 0.15 \\ 0.15 & 8.31 \end{bmatrix}\right), \tag{10}$$

$$S_{imp} \sim N\left(\mu_{imp} = \begin{bmatrix} 0.53 \\ 66.87 \end{bmatrix}, \Sigma_{imp} = \begin{bmatrix} 0.0015 & 0.03 \\ 0.03 & 9.45 \end{bmatrix}\right). \tag{11}$$

We generated $100,000$ genuine and $100,000$ impostor scores from the above distributions. In the first experiment, we assume that the parameters in equations (10) and (11) are known. The likelihood ratios were computed by utilizing the full Σ, and under the independence assumption (non-diagonal elements of Σ matrix are set to zero). The ROC curves for these two cases are plotted in Figure 4(a) which show that for this specific parameter set, utilizing the correlation information does not substantially improve the performance. On the other hand, if the Σ_{gen} matrix is changed to $\begin{bmatrix} 0.006 & 0.20 \\ 0.20 & 8.31 \end{bmatrix}$ (corresponds to increasing the correlation between the genuine matching scores of the two modalities (ρ_{gen}) from 0.75 to 0.90), we observe that fusion accounting for the correlation provides

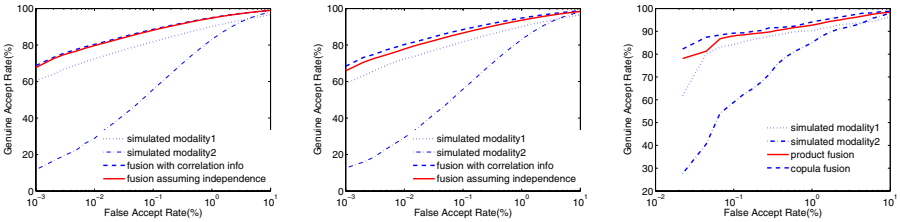


Fig. 4. ROC curves for the simulated data; (a) fusion with true parameters when $\rho_{gen} = 0.75$, (b) fusion with true parameters when $\rho_{gen} = 0.90$ and (c) fusion using estimated parameters when $\rho_{gen} = 0.90$

substantial improvement over the independence case (see Figure 4(b)). Now, if we estimate the parameters in equations (10) and (11) using the simulated data, the copula fusion rule outperforms the product rule as shown in Figure 4(c). These experiments illustrate that improvement in the recognition performance by using copula fusion rule depends on the underlying distribution of the matching scores. In the general case, the copula rule will perform at least as good as the product rule, provided there is sufficient amount of training data to estimate the correlation matrices accurately.

5 Summary

Based on the generalized density estimates of the genuine and impostor matching scores, two methods of fusion that follow the Neyman-Pearson rule are described. The first fusion rule computes the product of the likelihood ratios for each component modality of a multimodal system and is optimal when the modalities are independent of each other. The second fusion rule assumes that the generalized joint density of matching scores can be modeled using a Gaussian copula function and is a generalization of the product rule when the component modalities are not independent. Experimental results indicate that the two fusion rules achieve better performance compared to the single best modality in both the databases. The proposed method bypasses the need to perform score normalization and choosing optimal combination weights for each modality on a case-by-case basis. In this sense, the proposed solution is a principled and general approach that is optimal when the genuine and impostor matching score distributions are either known or can be estimated with high accuracy.

Acknowledgments

This research is partially supported by the NSF ITR grant 0312646 and the Center for Identification Technology Research (CITeR).

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics **14** (2004) 4–20
2. Jain, A.K., Ross, A.: Multibiometric Systems. Communications of the ACM, Special Issue on Multimodal Interfaces **47** (2004) 34–40
3. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recognition Letters, Special Issue on Multimodal Biometrics **24** (2003) 2115–2125
4. Bigun, E.S., Bigun, J., Duc, B., Fischer, S.: Expert Conciliation for Multimodal Person Authentication Systems using Bayesian Statistics. In: Proceedings of First International Conference on AVBPA, Crans-Montana, Switzerland (1997) 291–300
5. Kittler, J., Hatef, M., Duin, R.P., Matas, J.G.: On Combining Classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 226–239
6. Lam, L., Suen, C.Y.: Optimal Combination of Pattern Classifiers. Pattern Recognition Letters **16** (1995) 945–954
7. Wang, Y., Tan, T., Jain, A.K.: Combining Face and Iris Biometrics for Identity Verification. In: Proceedings of Fourth International Conference on AVBPA, Guildford, U.K. (2003) 805–813
8. Toh, K.A., Jiang, X., Yau, W.Y.: Exploiting Global and Local Decisions for Multimodal Biometrics Verification. IEEE Transactions on Signal Processing **52** (2004) 3059–3072
9. Snelick, R., Uludag, U., Mink, A., Indovina, M., Jain, A.K.: Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 450–455
10. Jain, A.K., Nandakumar, K., Ross, A.: Score Normalization in Multimodal Biometric Systems. To appear in Pattern Recognition (2005)
11. Griffin, P.: Optimal Biometric Fusion for Identity Verification. Identix Corporate Research Center Preprint RDNJ-03-0064 (2004)
12. Prabhakar, S., Jain, A.K.: Decision-level Fusion in Fingerprint Verification. Pattern Recognition **35** (2002) 861–874
13. Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman & Hall, CRC Press (1995)
14. Nelsen, R.B.: An Introduction to Copulas. Springer (1999)
15. Cherubini, U., Luciano, E., Vecchiato, W.: Copula Methods in Finance. Wiley (2004)
16. Turk, M., Pentland, A.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience **3** (1991) 71–86
17. Jain, A.K., Hong, L., Bolle, R.: On-line Fingerprint Verification. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 302–314
18. Jain, A.K., Ross, A., Pankanti, S.: A Prototype Hand geometry-based Verification System. In: Proceedings of Second International Conference on AVBPA, Washington D.C., USA (1999) 166–171
19. National Institute of Standards and Technology: NIST Biometric Scores Set. Available at `http://http://www.itl.nist.gov/iad/894.03/biometricscores` (2004)

A Score-Level Fusion Benchmark Database for Biometric Authentication

Norman Poh and Samy Bengio

IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
{norman,bengio}@idiap.ch

Abstract. Fusing the scores of several biometric systems is a very promising approach to improve the overall system's accuracy. Despite many works in the literature, it is surprising that there is no coordinated effort in making a benchmark database available. It should be noted that fusion in this context consists not only of multimodal fusion, but also intramodal fusion, i.e., fusing systems using the same biometric modality but different features, or same features but using different classifiers. Building baseline systems from scratch often prevents researchers from putting more efforts in understanding the fusion problem. This paper describes a database of scores taken from experiments carried out on the XM2VTS face and speaker verification database. It then proposes several fusion protocols and provides some state-of-the-art tools to evaluate the fusion performance.

1 Motivation

Biometric authentication (BA) is a process of verifying an identity claim using a person's behavioral and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. Examples of biometric modalities are fingerprint, face, voice, hand-geometry and retina scans [1]. However, today, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. Biometric data is often noisy because of deformable nature of biometric traits, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

One very promising approach to improve the overall system's accuracy is to fuse the scores of several biometric systems [2]. Despite many works in the literature, e.g. [3, 4], it is surprising that there is no coordinated effort in making a benchmark database available for such task. This work is one step towards better sharing of scores to *focus* on better understanding of the fusion mechanism.

In the literature, there are several approaches towards studying fusion. One practice is to use virtual identities whereby a biometric modality from one person is paired with the biometric modality of another person. From the experiment point of view, these biometric modalities belong to the same person. While this practice is somewhat accepted in the literature, it was questioned whether this was a right thing to do or not

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1059–1070, 2005.
© Springer-Verlag Berlin Heidelberg 2005

during the 2003 Workshop on Multimodal User Authentication [5]. The fundamental issue here is the independence assumption that two or more biometric traits of a single person are independent from each other[1]. Another practice is more reasonable: use off-the-shelf biometric systems [6] and quickly acquire scores. While this is definitely a better solution, committing to acquire the systems and to collect the data is admittedly a very time-consuming process. None of the mentioned approaches prevails over the others in understanding the problem of fusion. There are currently on-going but independent projects in the biometric community to acquire multimodal biometric databases, e.g., the BANCA [7], XM2VTS [8], BIOMET [9], MYCT [10] and University of Notre Dame Biometrics multimodal databases[2]. BANCA and XM2VTS contain face and speech modalities; BIOMET contains face, speech, fingerprint, hand and signature modalities; MYCT contains ten-print fingerprint and signature modalities and University of Notre Dame Biometrics Database contains face, ear profile and hand modalities acquired using visible, Infrared-Red and range sensors at different angles. Taking multimodal biometrics in a wider context, i.e., in the sense that it involves different sensors, the FRGC[3] database can also be considered as "multimodal". It contains face modality captured using camera (at different angles) and range sensors in different (controlled or uncontrolled) settings.

As a matter of fact, most reported works in the literature about fusion often concentrates on treatment of the baseline systems. While baseline systems are definitely important, the subject of fusion is unfortunately downplayed. Hence, we propose here not only to publish scores resulted from biometric authentication experiments, but also to provide a clear documentation of the baseline systems and well-defined *fusion protocols* so that experimental results can be compared. To the best of our knowledge, this is first ever published score data set. It is intended for comparison of different fusion classifiers *on a common setting*. We further provide a set of evaluation tools such as the DET [11] curve and the recent Expected Performance Curve (EPC) [12], visualisation of False Acceptance and False Rejection Rates versus threshold, distribution of client and impostor scores, and the HTER significance test [13], among others.

The scores are taken from the publicly available XM2VTS face and speech database[4]. It should be mentioned here that there exists another software tool that analyses biometric error rate called PRESS[14]. However, it does not include the DET curve. The tools proposed here, together with the database, provide a new plot called Expected Performance Curve (EPC) [12] and a significance test specially designed to test the Half Total Error Rate (HTER) [13].

Section 2 explains the XM2VTS database, the Lausanne Protocols and the proposed Fusion Protocols. Section 3 documents the 8 baseline systems that can be used for fusion. Section 4 presents the evaluation criteria, i.e., how experiments should be reported and compared. This is followed by conclusions in Section 5.

[1] To the best of our knowledge, there is no work in the literature that approves or disapproves such assumption

[2] Accessible from http://www.nd.edu/∼cvrl/UNDBiometricsDatabase.html

[3] Accessible from http://www.frvt.org/FRGC/

[4] The database of scores as well as the tools mentioned are freely available for download at http://www.idiap.ch/∼norman/fusion

2 Database and Protocols

2.1 The XM2VTS Database and the Lausanne Protocols

The XM2VTS database [15] contains synchronised video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence. The database is divided into three sets: a training set, an evaluation set and a test set. The training set (LP Train) was used to build client models, while the evaluation set (LP Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (LP Test) was used to estimate the performance.

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. In both configurations, the test set remains the same. Their difference is that there are three training shots per client for LP1 and four training shots per client for LP2. Table 1 is the summary of the data. The last column of Table 1 is explained in Section 2.2. Note that LP Eval's of LP1 and LP2 are used to calculate the optimal thresholds that will be used in LP Test. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). More details can be found in [8].

2.2 The Fusion Protocols

The fusion protocols are built upon the Lausanne Protocols. Before the discussion, it is important to distinguish two categories of approaches: *client-independent* and *client-dependent* fusion approaches. The former approach has only a global fusion function that is common to *all* identities in the database. The latter approach has a different fusion function for a different identity. It has been reported that client-dependent fusion is better than client-independent fusion, given that there are "enough" client-dependent

Table 1. The Lausanne Protocols of XM2VTS database. The last column shows the terms used in the fusion protocols presented in Section 2.2. LP Eval corresponds to the Fusion protocols' development set while LP Test corresponds to the Fusion Protocols' evaluation set

Data sets	Lausanne Protocols		Fusion Protocols
	LP1	LP2	
LP Train client accesses	3	4	NIL
LP Eval client accesses	600 (3×200)	400 (2×200)	Fusion dev
LP Eval impostor accesses	40,000 ($25 \times 8 \times 200$)		Fusion dev
LP Test client accesses	400 (2×200)		Fusion eva
LP Test impostor accesses	112,000† ($70 \times 8 \times 200$)		Fusion eva

†: Due to one corrupted speech file of one of the 70 impostors in this set, this file was deleted, resulting in 200 less of impostor scores, or a total of 111,800 impostor scores.

score data. Examples of client-dependent fusion approach are client-dependent threshold [16], client-dependent score normalisation [17] or different weighing of expert opinions using linear [18] or non-linear combination [19]. The fusion protocols that are described here can be client-dependent or client-independent.

It should be noted that one can fuse any of the 8 baseline experiments in LP1 and 5 baseline experiments in LP2 (to be detailed in Section 3). We propose a full combination of all these systems. This protocol is called **FP-full**. Hence, there are altogether $2^8 - 8 - 1 = 248$ possible combinations for LP1 and $2^5 - 5 - 1 = 26$ for LP2. The reasons for minus one and minus the number of experts are that using zero expert and using a single expert are not valid options. However, some constraints are useful. For instance, in some situations, one is constrained to using a single biometric modality. In this case, we propose an intramodal fusion (**FP-intramodal**) . When no constraint is imposed, we propose a full combination (**FP-multimodal**). FP-intramodal contains $2^5 - 5 - 1 = 26$ face-expert fusion experiments for LP1, $2^3 - 3 - 1 = 4$ speech-expert fusion experiments for LP1, 1 face-expert fusion experiment for LP2 and $2^3 - 3 - 1 = 4$ speech expert-fusion experiments for LP2. Hence, FP-intramodal contains 35 fusion experiments. The second protocol contains $\sum_{m=1}^{5}\sum_{n=1}^{3}({}^5C_m\,{}^3C_n) = 217$ combinations, where ${}^nC_k$ is "n choose k". As can be seen, the first three fusion protocols contain an exponential number of combinations. For some specific study, it is also useful to introduce a smaller set of combinations, each time using only two baseline experts, according to the nature of the base-expert. This protocol is called **FP-2**. Three categories of fusion types have been identified under FP-2, namely multimodal fusion (using different biometric traits), intramodal fusion with *different* feature sets and intramodal fusion with the *same* feature set but *different* classifiers. There are altogether 32 such combinations (not listed here; see [20] for details).

Note that there are 8 biometric samples in the XM2VTS database on a per client basis. They are used in the following decomposition: 3 samples are used to train the baseline experts in LP1 (and 4 in LP2) on LP Train. There are remaining 3 samples in the in LP1 Eval (and only 2 in LP2 Eval). Finally, for both protocols, 2 client accesses for testing in the *test set*. Because fusion classifiers cannot be trained using scores from the *training set*, or they are simply not available in the current settings, we are effectively using the LP Eval to train the fusion classifiers and then LP Test to test the fusion classifiers' performance on the LP Test. To avoid confusion in terminology used, we call LP Eval as the *fusion development set* and LP Test as the *fusion evaluation set*.

3 Baseline System Description

There are altogether 8 baseline systems[5] All the 8 baseline systems were used in LP1. On the other hand, 5 out of 8 were used in LP2. This results in 13 baseline experiments (for LP1 and LP2). The following explanation describes these systems in terms of their features, classifiers, and the complete system which is made up of the pair (feature type, classifier).

[5] Public contribution of score files is welcome. More will be released in the future as they become available

3.1 Face and Speech Features

The face baseline experts are based on the following features:

1. **FH**: normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**) [21].
2. **DCTs**: DCTmod2 features [22] extracted from face images with a size of 40×32 (rows $\times$ columns) pixels. The Discrete Cosine Transform (DCT) coefficients are calculated from an 8×8 window with horizontal and vertical overlaps of 50%, i.e., 4 pixels in each direction. Neighbouring windows are used to calculate the "delta" features. The result is a set of 35 feature vectors, each having a dimensionality of 18. (**s** indicates the use of this small image compared to the bigger size image with the abbreviation **b**.)
3. **DCTb**: Similar to DCTs except that the input face image has 80×64 pixels. The result is a set of 221 feature vectors, each having a dimensionality of 18.

The speech baseline experts are based on the following features:

1. **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) [23] speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. The first temporal derivatives are added to the feature set.
2. **PAC**: The Phase Auto-Correlation Mel Filter-bank Cepstral Coefficient (PAC-MFCC) features [24] are derived with a window length of 20 milliseconds and each window moves at a rate of 10 milliseconds. 20 DCT coefficients are computed to decorrelate the 30 coefficients obtained from the Mel-scale filter-bank. The first temporal derivatives are added to the feature set.
3. **SSC**: Spectral Subband Centroid (SSC) features, originally proposed for speech recognition [25], were used for speaker authentication in [26]. It was found that mean-subtraction could improve these features significantly. The mean-subtracted SSCs are obtained from 16 coefficients. The γ parameter, which is a parameter that raises the power spectrum and controls how much influence the centroid, is set to 0.7 [27]. Also, the first temporal derivatives are added to the feature set.

3.2 Classifiers

Two different types of classifiers were used for these experiments: Multi-Layer Perceptrons (MLPs) and a Bayes Classifier using Gaussian Mixture Models (GMMs) [28]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice MLPs are better at matching feature vectors of fixed-size while GMMs are better at matching sequences (feature vectors of unequal size). Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the feature vectors associated to the client are treated as positive patterns while all other feature vectors *not* associated to the client are treated as negative patterns. All MLPs reported here were trained using the stochastic version of the error-back-propagation training algorithm [28].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [28]. The world model is then adapted for each client to the corresponding client data using the Maximum-A-Posteriori adaptation [29] algorithm.

3.3 Baseline Systems

The baseline experiments based on DCTmod2 feature extraction were reported in [30] while those based on normalised face images and RGB histograms (FH features) were reported in [21]. Details of the experiments, coded in the pair **(feature, classifier)**, for the face experts, are as follows:

1. **(FH, MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [21].
2. **(DCTs, GMM)** The face features are the DCTmod2 features calculated from an input face image of 40×32 pixels, hence, resulting in a sequence of 35 feature vectors each having 18 dimensions. There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [30].
3. **(DCTb, GMM)** Similar to (DCTs,GMM), except that the features used are DCTmod2 features calculated from an input face image of 80×64 pixels. This produces in a sequence of 221 feature vectors each having 18 dimensions. The corresponding GMM has 512 Gaussian components [30].
4. **(DCTs, MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [30]. Note that in this case a training example consists of a *big single* feature vector with a dimensionality of 35×18. This is done by simply concatenating 35 feature vectors each having 18 dimensions[6].
5. **(DCTb, MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used. Note that in this case the MLP in trained on a *single* feature vector with a dimensionality of 221×18 [30].

and for the speech experts:

1. **(LFCC, GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 200 Gaussian components, with the minimum relative variance of each Gaussian fixed to 0.5,

[6] This may explain why MLP, an inherently discriminative classifier, has worse performance compared to GMM, a generative classifier. With high dimensionality yet having only a few training examples, the MLP cannot be trained optimally. This may affect its generalisation on unseen examples. By treating the features as a sequence, GMM was able to generalise better and hence is more adapted to this feature set

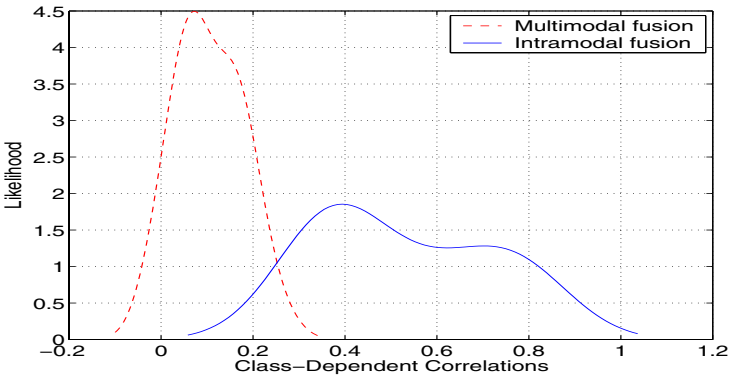


Fig. 1. Smoothed distribution (measured as unnormalised likelihood using Parzen window technique) of class-dependent correlations on the 32 fusion data sets of the FP-2 protocol according to the two categories of fusion: multimodal or intramodal. Since each data set has two classes (client and impostor), there are altogether $2 \times 32 = 64$ correlation values

and the MAP adaptation weight equals 0.1. This is the best known model currently available [31] under clean conditions.

2. **(PAC, GMM)** The same GMM configuration as in LFCC is used. Note that in general, 200-300 Gaussian components would give about 1% of difference of HTER [31]. This system is particularly robust to very noisy conditions (less than 6 dBs, as tested on the NIST2001 one-speaker detection task).
3. **(SSC, GMM)** The same GMM configuration as in LFCC is used [27]. This system is known to provide an optimal performance under moderately noisy conditions (18-12 dBs, as tested on NIST2001 one-speaker detection task).

3.4 Preliminary Correlation Analysis

A preliminary analysis was carried out on the FP-2 protocol. There are 32 fusion data sets here and each data set contains scores of two experts. Each data set contains two classes: client or impostor scores. For each class of each data set, we computed the correlation between scores of two experts in the linear space. The GMM and SVM scores are used as they are. Since correlation measures the linear relationship among variables, it fails to measure the MLP scores which are trained using a tanh or a sigmoid *nonlinear* activation function. An inverse of tanh or sigmoid function is applied to the scores prior to computing the correlation values. With the absence of this *corrective* procedure, the strong correlation is *systematically* under-estimated for the intramodal fusion datasets. The resultant distribution of these correlation values, categorised into intramodal and multimodal fusion datasets, are shown in Fig. 1. As can be observed, the multimodal fusion datasets have correlation around zero whereas the intramodal fusion datasets have relatively high correlation values. This is a strong indication that the gain from fusion using the intramodal data sets will be less than that from using the multimodal data sets.

4 Performance Evaluation

There are three important concepts about evaluation of a biometric system: (1) types of errors in biometric authentication, namely false acceptance, false rejection and their combined error called Weighted Error Rate (WER), (2) threshold criterion and (3) evaluation criterion. A *threshold criterion* refers to a strategy to choose a threshold to be applied on *an evaluation (test) set*. It is necessarily tuned on *a development (training) set*. An *evaluation criterion* is used to measure the final generalisation performance and is necessarily calculated on *an evaluation set*. A fully operational biometric system makes a decision using the following *decision function*:

$$F(\mathbf{x}) = \begin{cases} accept & \text{if } y(\mathbf{x}) > \Delta \\ reject & \text{otherwise,} \end{cases} \tag{1}$$

where $y(\mathbf{x})$ is the output of the underlying expert supporting the hypothesis that the biometric sample received $\mathbf{x}$ belongs to a client. The variables that follow will be derived from $y(\mathbf{x})$. For simplicity, we write y instead of $y(\mathbf{x})$. The same convention applies to variables that follow. Because of the accept-reject outcomes, the system may make two types of errors, i.e., false acceptance (FA) and false rejection (FR). Normalised versions of FA and FR are often used and called false acceptance rate (FAR) and false rejection rate (FRR), respectively. They are defined as:

$$\text{FAR}(\Delta) = \frac{\text{FA}(\Delta)}{NI}\ , \tag{2}$$

$$\text{FRR}(\Delta) = \frac{\text{FR}(\Delta)}{NC}\ . \tag{3}$$

where FA and FR count the number of FA and FR accesses, respectively; and NI and NC are the total number of impostor and client accesses, respectively.

To choose an "optimal threshold" Δ, it is necessary to define a threshold criterion. This has to be done on a development set. Two commonly used criteria are the Weighted Error Rate (WER) and Equal Error Rate (EER). WER is defined as:

$$\text{WER}(\alpha, \Delta) = \alpha\text{FAR}(\Delta) + (1 - \alpha)\ \text{FRR}(\Delta), \tag{4}$$

where $\alpha \in [0, 1]$ balances between FAR and FRR. A special case of WER is EER, which assumes that the costs of FA and FR are equal. It further assumes that the class prior distributions of client and impostor accesses are equal. As a result $\alpha = 0.5$. Let Δ^*_α be the optimal threshold that *minimises* WER on a *development set*. It can be calculated as follows:

$$\Delta^* = \arg\min_{\Delta} \text{WER}(\alpha, \Delta). \tag{5}$$

Note that the EER criterion can be calculated similarly by fixing $\alpha = 0.5$.

Having chosen an optimal threshold using the WER threshold criterion discussed previously, the final performance is measured using Half Total Error Rate (HTER). Note that the threshold is found with respect to a given α. It is defined as:

$$\text{HTER}(\Delta^*_\alpha) = \frac{\text{FAR}(\Delta^*_\alpha) + \text{FRR}(\Delta^*_\alpha)}{2}. \tag{6}$$

It is important to note that the FAR and FRR do not have the same *resolution*. Because there are more simulated impostor accesses than the client accesses, FRR changes more drastically when falsely rejecting a client access whereas FAR changes less drastically when falsely accepting an impostor access. Hence, when comparing the performance using HTER(Δ^*_α) from two systems (at the *same* Δ^*_α), the question of whether the HTER difference is significant or not has to take into account the imbalanced numbers of client and impostor accesses. This issue was studied in [13], and as a result, the HTER significance test was proposed. Finally, it is important to note that HTER in Eqn. (6) is identical to EER (WER with $\alpha = 0$) except that HTER is a *performance measure* (calculated on an *evaluation set* whereas EER is a *threshold criterion* optimised on a *development set*. Because of their usage in different context, EER should not be interpreted as a performance measure (in place of HTER) to compare the performance of different systems. Such practice, to our opinion, leads to an *unrealistic* comparison. The argument is that in an actual operating system, the threshold has to be fixed *a priori*. To distinguish these two concepts, when discussing HTER calculated on a development set using a threshold criterion also calculated on the same set, the HTER should be called *a posteriori* HTER. When discussing HTER calculated on an evaluation set with a threshold optimised on a development set, the HTER should be called *a priori* HTER.

The most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [11] and Receiver's Operating Characteristic (ROC) curve[7]. A DET curve is a ROC curve plotted in normal probability co-ordinate scales in its X- and Y-axes. It has been pointed out [12] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [12] that such threshold should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [12] was proposed. This curve is constructed as follows: for various values of α between 0 and 1, select the optimal threshold Δ on a development (training) set, apply it on the evaluation (test) set and compute the HTER on the evaluation set. This HTER is then plotted with respect to α. The EPC curve can be interpreted similarly to the DET curve, i.e., the lower the curve, the better the generalisation performance. Although EPC is *recommended*, due to the popularity of ROC and DET curves, it is reasonable to report experimental results with these curves as well alongside with EPC. In this case, the pair of FAR and FRR values that constitute a point in ROC can be derived from the FAR and FRR terms in Eqn. (6), i.e., with the threshold Δ^*_α derived from the development (training) set.

5 Conclusions

In this study, we presented a score-level fusion database, several fusion protocols in different scenarios and some evaluation tools to encourage researchers to focus on the problem of biometric authentication score-level fusion. To the best of our knowledge, there has been no work in the literature that provides a benchmark database for score-level fusion. Hence, the efficiency of fusion classifiers can now be compared on equal

[7] A good introduction can be found in "http://www.anaesthetist.com/mnm/stats/roc/"

platforms. We also further encourage contribution of scores following the *same* Lausanne Protocols to enrich this corpus. An extended version of this report, which includes a greater level of details on the evaluation tools, can be found in [20]. Finally, some baseline results on this data set using the fusion protocol with two experts (FP-2) can be found in [32].

Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. The authors thank Julian Fierrez-Aguilar and the anonymous reviewers for giving suggestions and constructive comments, and Fabien Cardinaux and Sébastien Marcel for providing the data sets. This publication only reflects the authors' view.

References

1. A.K. Jain, R. Bolle, and S. Pankanti, *Biometrics: Person Identification in a Networked Society*, Kluwer Publications, 1999.
2. J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, "Combining Evidence in Personal Identity Verification Systems," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, 1997.
3. J. Kittler, K. Messer, and J. Czyz, "Fusion of Intramodal and Multimodal Experts in Personal Identity Authentication Systems," in *Proc. Cost 275 Workshop*, Rome, 2002, pp. 17–24.
4. J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez, "A Comparative Evaluation of Fusion Strategies for Multimodal Biometric Verification," in *Springer LNCS-2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2003)*, Guildford, 2003, pp. 830–837.
5. J-L. Dugelay, J-C. Junqua, K. Rose, and M. Turk (Organizers), *Workshop on Multimodal User Authentication (MMUA 2003)*, no publisher, Santa Barbara, CA, 11–12 December, 2003.
6. M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 99–106.
7. E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, "The BANCA Database and Evaluation Protocol," in *Springer LNCS-2688, 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication, AVBPA'03*. 2003, Springer-Verlag.
8. J. Lüttin, "Evaluation Protocol for the XM2FDB Database (Lausanne Protocol)," Communication 98-05, IDIAP, Martigny, Switzerland, 1998.
9. S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrétaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities," in *Springer LNCS-2688, 4th Int'l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, Guildford, 2003, pp. 845–853.
10. J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro, "Biometric on the Internet MCYT Baseline Corpus: a Bimodal Biometric Database," *IEE Proc. Visual Image Signal Processing*, vol. 150, no. 6, pp. 395–401, December 2003.

11. A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.
12. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.
13. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.
14. M. E. Schuckers and C. J. Knickerbocker, *Documentation for Program for Rate Estimation and Statistical Summaries PRESS*, Department of Mathematics, Computer Science and Statistics St Lawrence University, Canton, NY 13617 and Center for Identification Technology Research, West Virginia University.
15. J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.
16. J.R. Saeta and J. Hernando, "On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 215–218.
17. J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target Dependent Score Normalisation Techniques and Their Application to Signature Verification," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 498–504.
18. A. Jain and A. Ross, "Learning User-Specific Parameters in Multibiometric System," in *Proc. Int'l Conf. of Image Processing (ICIP 2002)*, New York, 2002, pp. 57–70.
19. A. Kumar and D. Zhang, "Integrating Palmprint with Face for User Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 107–112.
20. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Research Report 04-44, IDIAP, Martigny, Switzerland, 2004.
21. S. Marcel and S. Bengio, "Improving Face Verification Using Skin Color Information," in *Proc. 16th Int. Conf. on Pattern Recognition*, Quebec, 2002, p. unknown.
22. C. Sanderson and K.K. Paliwal, "Fast Features for Face Authentication Under Illumination Direction Changes," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.
23. L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Oxford University Press, 1993.
24. S. Ikbal, H. Misra, and H. Bourlard, "Phase Auto-Correlation (PAC) derived Robust Speech Features," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003, pp. 133–136.
25. K. K. Paliwal, "Spectral Subband Centroids Features for Speech Recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, 1998, vol. 2, pp. 617–620.
26. N. Poh, C. Sanderson, and S. Bengio, "An Investigation of Spectral Subband Centroids For Speaker Authentication," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 631–639.
27. N. Poh, C. Sanderson, and S. Bengio, "An Investigation of Spectral Subband Centroids For Speaker Authentication," Research Report 03-62, IDIAP, Martigny, Switzerland, 2003.
28. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.
29. J.L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains," *IEEE Tran. Speech Audio Processing*, vol. 2, pp. 290–298, 1994.

30. F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS," in *Springer LNCS-2688, 4th Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, Guildford, 2003, pp. 911–920.
31. N. Poh and S. Bengio, "Noise-Robust Multi-Stream Fusion for Text-Independent Speaker Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 199–206.
32. N. Poh and S. Bengio, "Non-Linear Variance Reduction Techniques in Biometric Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 123–130.

Fusion for Multimodal Biometric Identification

Yongjin Lee[1], Kyunghee Lee[2], Hyungkeun Jee[1], Younhee Gil[1], Wooyong Choi[1], Dosung Ahn[1], and Sungbum Pan[3]

[1] Biometrics Technology Research Team
Electronics and Telecommunications Research Institute
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, Korea
{solarone,hkjee,yhgil,wychoi4,dosung}@etri.re.kr
[2] Department of Electrical Engineering
The University of Suwon, Korea
khlee@suwon.ac.kr
[3] Division of Information and Control Measurement Engineering
Chosun University, Korea
sbpan@chosun.ac.kr

Abstract. In this paper, we investigate fusion methods for multimodal identification using several unimodal identification results. One fingerprint identification system and two face identification systems are used as fusion sources. We discuss rank level and score level fusion methods. Whereas the latter combines similarity scores, the other one combines the orders of the magnitudes of the similarity scores. For rank level methods, Borda Count and Bayes Fuse are considered and, for score level methods, Sum Rule and Binary Classification Approach are considered. Especially, we take a more detailed look at Binary Classification Approach, which simplifies a multiple class problem into a binary class problem. Finally, we compare experimental results using the fusion methods in different combinations of the sources.

1 Introduction

As biometrics attracts more and more attention from many areas, the demand of high accuracy and reliability is also increasing. Although various biometric systems have been developed and improved, there are still limitations which have to be overcome to meet stringent performance requirements from many applications. Most systems are unimodal, i. e., they rely on the single source of information for establishing identity, and most limitations are imposed by their unimodality. The problems which those systems have to contend with are such as noise in sensed data, intra-class variation, inter-class similarity, non-universality, and spoof attack [1]. When a large number of users are comprised in a system, inter-class similarity may be a major cause for performance degradation. Golfarelli et al. state that the number of distinguishable patterns in two of the most commonly used representations of hand geometry and face are only of the order of 10^5 and 10^3, respectively [2][1]. Multimodal biometrics may be, therefore, the only way to construct a robust identification system since identification itself intrinsically involves a large number of users.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1071–1079, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Multimodal biometric systems can be implemented by fusing more than two biometric systems. The three possible levels of fusion are: fusion at the feature extraction level, fusion at the matching score level, and fusion at the decision level [3]. In feature level fusion, feature vectors, which are extracted from several biometric traits, are concatenated into a singe higher dimensional vector. Although feature level fusion can contain the richest information among the three fusion methods, it causes to compare the heterogeneous data in higher dimension. Since an identification system has to compare all templates in the database, comparison of high dimensional data is a serious burden to an identification system. By integrating individual unimodal systems, distribution process can be achieved naturally and is more efficient for management. Thus, only score and decision level fusions are reasonable approaches to an identification system. The rank level fusion is a more proper name than decision level fusion for identification due to its multiplicity.

In this paper, we investigate four fusion methods, Borda Count, Bayes Fusion, Sum Rule, and Binary Classification Approach. The former two are based on rank and the latter two are based on score. Especially, we take a more detailed look at Binary Classification Approach, which simplifies a multiple class problem into a binary class problem. One fingerprint and two face identification systems are used as fusion sources. The rest of the paper is organized as follows. The three unimodal identification systems and fusion methods will be described in Sec. 2 and Sec. 3, respectively. And experimental results for fusion are shown in Sec. 4. Finally, we present our conclusion in Sec. 5.

2 Unimodal Identification

In this section, we briefly describe fingerprint and face identification systems, which provide sources for fusion. Given a probe, each unimodal identification system return N(the number of candidates in a database) similarity or dissimilarity scores, which are results from comparison of one probe with N galleries in the database.

2.1 Fingerprint Identification

It is widely known that a professional fingerprint examiner relies on details of ridge structures to make fingerprint identifications. It implies that fingerprint authentication can be based on the matching of structural patterns. Generally, structural features used in fingerprint identification are composed of the point where ridge ends and that ridge bifurcates, which are called minutiae. Our representation is minutiae based, and each minutia is described by its position in x, y coordinates, the direction it flows and the type i.e., ridge ending or bifurcation. After refinement and alignment of fingerprint images, two minutiae from a probe and a gallery set are compared based on their position, direction, and type. Then, a matching score is computed. A detailed method can be found in [4].

2.2 Face Identification

For face recognition, feature extraction is required to represent high dimensional image data into low dimensional feature vectors. Among various methods, we use PCA(Principal Component Analysis), which is known as *eigenface* in face recognition field [5]. The bases, which are used to extract face features, are eigenvectors of the covariance matrix of the face images and thought of as face models, called *eigenfaces*. By projecting a face image onto the eigenfaces, the linear combination weights for eigenfaces are calculated. These weights are used as representations of the face. In this paper, comparisons between feature vectors of a probe and a galley set are performed using Euclidean distance and SVM(Support Vector Machine) [6], which models each person in the database. While Euclidean distance gives distance values as dissimilarity scores, SVM gives classification results as similarity scores.

3 Fusion Methods

For consolidating more than two identification results, score and rank level fusions are possible. Whereas the former uses similarity (or dissimilarity) scores, the other uses the orders of the magnitudes of similarity (or dissimilarity) scores. Fusion at score level is more flexible and contains more information than rank level fusion, but it requires transforming scores of multiple sources into common domain. Fusion at rank level is convenient to use but rather rigid. In this section, we describe four fusion methods at rank level and score level.

3.1 Borda Count

Borda Count is a voting algorithm from social choice theory and was used for a metasearch engine in information retrieval [7]. Metasearch engine combines lists of documents returned by multiple search engines to obtain better results. The general ideas of a metasearch engine and a multimodal biometric system are the same. The difference is that information retrieval systems can have multiple targets (*relevant set*), but biometric systems can have only one target (*genuine*).

Borda Count is simple, fast, and unsupervised rank level method. It works as follows. Suppose that there are M unimodal systems and N candidates. For each systems, the top ranked candidate is given by N points, the second ranked candidate is given $N-1$ points, and so on. The fused score for candidate i can be written as follows.

$$f_i = \sum_{m=1}^{M} (N - r_{i,m} + 1) \tag{1}$$

where $r_{i,m}$ represents rank of candidate i given by unimodal system m. Candidate list is reranked using the fused score.

3.2 Bayes Fuse

Bayes Fuse was developed in information retrieval field, too [7][8]. It is supervised and rank level method, which is based on Bayesian inference. The training and

test is very fast and simple. Originally it used *relevant* and *irrelevant* terms for document sets, but they can be changed into *genuine* and *impostor* for human candidates and can be applied to a human identification system. The fused score for candidate i is given as

$$f_i = \sum_{m=1}^{M} \log \frac{Pr[r_{i,m}|genuine]}{Pr[r_{i,m}|impostor]} \tag{2}$$

We need training data to estimate the likelihood probabilities $Pr[r_m|genuine]$ and $Pr[r_m|impostor]$. $Pr[r_m|genuine]$ is the probability that a genuine would be ranked at level r_m by system m. Similarly, $Pr[r_m|impostor]$ is the probability that an impostor would be ranked at level r_m by system m.

Eq.(2) is derived easily from the odds ratio of two posterior probabilities,

$$P_G = Pr[genuine|r_1, r_2, \ldots, r_M],$$
$$P_I = Pr[impostor|r_1, r_2, \ldots, r_M]$$

with the assumption of independence between systems. Detailed derivation can be found in [7] and [8].

3.3 Sum Rule

This is a simple and unsupervised method at score level. After transforming all scores into a similarity measure, it sums all the scores. The fused score for candidate i is given as

$$f_i = \sum_{m=1}^{M} s_{i,m} \tag{3}$$

where $s_{i,m}$ is a similarity score for candidate i given by system m.

3.4 Binary Classification Approach

Given a probe to an identification system, N comparisons are performed against a gallery set in each unimodal system. Through the N comparisons, one *genuine score* and $(N-1)$ *impostor scores* are generated. By concatenating the matching scores from M unimodal systems, we can have one M-dimensional pattern resulted from a comparison with a target template and $(N-1)$ M-dimensional patterns resulted from comparisons with non-target templates. Therefore, by putting the most genuine-like pattern into the top rank, multimodal identification can be performed.

This approach considers the concatenated score vectors as new features and a pre-trained binary classifier discriminates them into two classes, genuine pattern and impostor pattern. Therefore, the fused score, f_i, is given by the distance from the decision boundary of the two classes and represents confidence rate of the classifier. Absolute decision results, such as $+1/-1$ or yes/no, are not necessary because we order the candidates using the relative magnitudes of the fused

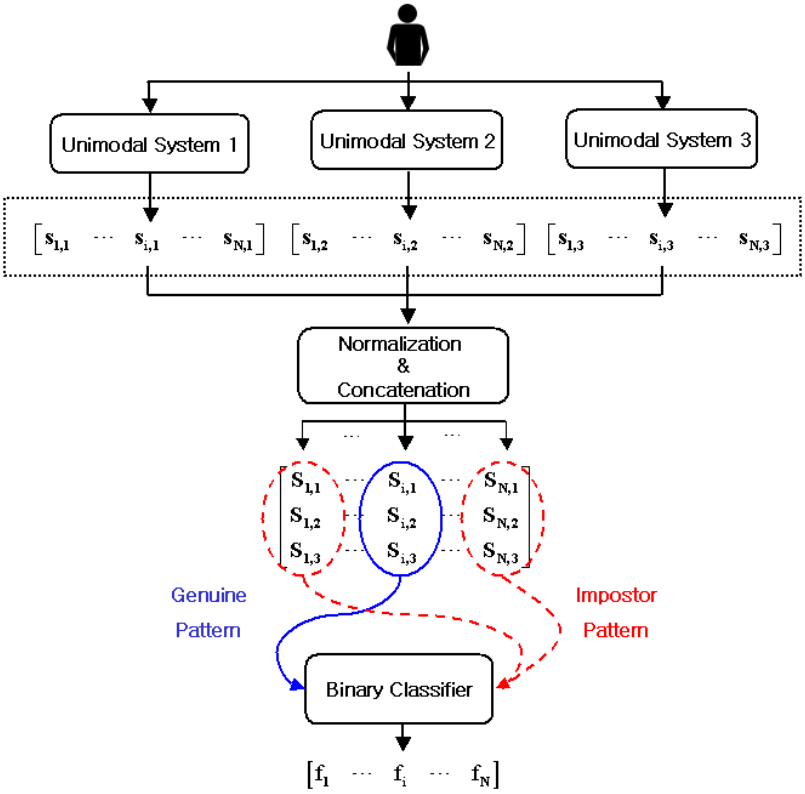


Fig. 1. Binary Classification Approach. $s_{i,1}$, $s_{i,2}$ and $s_{i,3}$ refers to scores for candidate i by unimodal system 1, 2 and 3, respectively. And $[S_{i,1}S_{i,2}S_{i,3}]^T$ represents a normalized score vector for candidate i. The final fused score, f_i, is given by the decision result of the binary classifier

scores. However, the absolute decision results can be used to indicate whether a candidate is in a watch list or not. And, by minimizing classification error, the classifier considers dependency between unimodal systems.

The classifier does not need to be dependent on a user. By treating the fusion as a two class problem and training the classifier using score vectors from a subset of users, we can easily integrate the identification results without restriction on the number of candidates generated by the unimodal systems. In case of three unimodal systems, a schematic draw of the method is given in Fig. 1.

4 Experiment Results

4.1 Database

The database for fusion experiments consists of identification results of one fingerprint and two face systems. Because the pairs of the two biometric traits

from a single set of users are not available to us and they are supposed to be independent of each other, we construct a multimodal database from two distinct fingerprint and face databases. The fingerprint database is provided by KISIS [9] and the face database is XM2VTS, which is publicly available [10]. The final database consists of 287 users in which biometric information is not overlapped between users.

After constructing unimodal identification systems, four probes per person were available. Two probes with a known identity were used as training data for supervised fusion, and the rest two were used for test. For score level fusion, we transformed a dissimilarity measure, which is based on distance, into a similarity measure by multiplying -1. Then all scores were normalized between 0 and 1 by subtracting the minimum value and dividing the maximum value. The minimum and maximum values were selected from a training data set for supervised fusion methods.

The scatter plot of the normalized scores for training is shown in Fig. 2. The scores from the two face systems are highly correlated to each other, but the scores from the fingerprint system is not. In case of the fingerprint system, scores of genuine and impostor are separated clearly, but not the face systems. In the figure, *Euc* and *SVM* represent face identification using Euclidean distance and SVM as described in the previous section. And *Finger* refers to fingerprint identification.

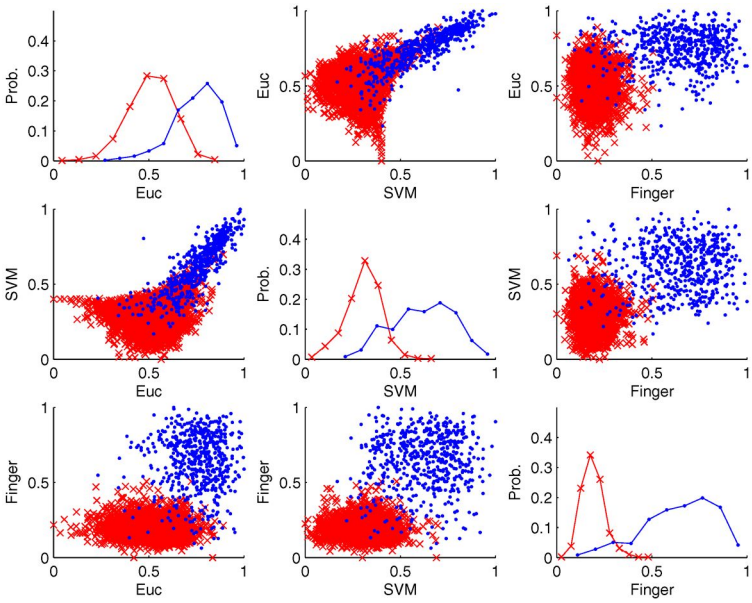


Fig. 2. Scatter plots of normalized scores of training data. *Euc* and *SVM* represent scores from the face systems using Euclidean distance and SVM, respectively. *Finger* refers to scores from the fingerprint system. Blue dots and red crosses stand for genuine and impostor. And the figures in the diagonal are score distributions

Table 1. Rank-N recognition rate(%). Face(Euc)+Face(SVM)

Rank	Unimodal		Fusion			
	Face(Euc)	Face(SVM)	Borda	Bayes	Sum	SVM
1	57.84	65.68	62.72	66.03	66.03	66.55
2	66.20	71.95	70.38	72.82	73.35	73.52
3	69.34	75.96	75.26	75.09	76.48	77.18
4	71.78	78.22	78.22	78.57	79.09	79.62
5	74.56	79.97	80.14	80.31	81.19	80.31
6	75.44	81.01	82.23	82.40	82.75	82.23
7	77.35	82.06	83.28	83.97	84.15	83.10
8	78.57	83.62	84.15	84.84	84.84	85.02
9	79.44	84.84	84.50	85.89	85.71	86.06
10	79.79	86.06	85.19	86.41	86.76	86.59

Table 2. Rank-N recognition rate(%). Face(Euc)+Finger

Rank	Unimodal		Fusion			
	Face(Euc)	Finger	Borda	Bayes	Sum	SVM
1	57.84	93.38	81.71	94.43	97.39	97.21
2	66.20	94.95	84.67	97.91	97.74	97.91
3	69.34	94.95	86.59	98.26	98.26	98.61
4	71.78	95.30	87.63	98.26	98.78	98.78
5	74.56	95.65	88.33	98.78	98.78	98.78
6	75.44	95.82	88.85	98.78	98.78	98.78
7	77.35	95.99	89.20	98.96	98.78	98.78
8	78.57	96.17	89.37	98.96	98.78	98.78
9	79.44	96.52	89.37	98.96	98.78	98.78
10	79.79	96.69	89.90	99.13	98.96	98.78

4.2 Experimental Results

Rank-N recognition rates of fusion and individual unimodal systems are shown in Table 1, 2, 3, and 4. The rank-N recognition rate is the proportion that the score of the correct gallery is within top N scores for each probe. For Binary Classification Approach, we used SVM with Gaussian kernel.

As shown in Table 1, there were no apparent improvements when two face systems were combined. By combining independent sources, performance gains were obtained. Although the recognition rate of the fingerprint system was very high, all methods, except for Borda Count, achieved higher recognition rates. We think that the reason for Borda Count's failure is in information loss by its uniform discretization of scores rather than its unsupervised learning. This is because the simple Sum Rule, which is an unsupervised score level method, achieved a better performance improvement than Bayes Fuse, which is a supervised rank level method. Moreover, the improvement of Sum Rule is competitive compared to SVM fusion, which is a supervised method.

Table 3. Rank-N recognition rate(%). Face(SVM)+Finger

Rank	Unimodal		Fusion			
	Face(SVM)	Finger	Borda	Bayes	Sum	SVM
1	65.68	93.38	85.71	94.77	97.21	97.74
2	71.95	94.95	89.37	98.08	98.61	98.78
3	75.96	94.95	90.24	98.78	98.96	98.96
4	78.22	95.30	90.42	98.78	99.30	99.48
5	79.97	95.65	91.12	99.13	99.48	99.65
6	81.01	95.82	91.81	99.30	99.65	99.65
7	82.06	95.99	91.81	99.30	99.65	99.65
8	83.62	96.17	92.16	99.48	99.65	99.65
9	84.84	96.52	92.33	99.48	99.83	99.65
10	86.06	96.69	92.68	99.48	99.83	99.65

Table 4. Rank-N recognition rate(%). Face(Euc)+Face(SVM)+Finger

Rank	Unimodal			Fusion			
	Face(Euc)	Face(SVM)	Finger	Borda	Bayes	Sum	SVM
1	57.84	65.68	93.38	82.85	87.81	95.99	98.26
2	66.20	71.95	94.95	86.24	97.39	97.56	98.61
3	69.34	75.96	94.95	86.93	98.43	97.91	98.78
4	71.78	78.22	95.30	88.15	98.78	98.26	99.13
5	74.56	79.97	95.65	88.68	98.78	98.43	99.30
6	75.44	81.01	95.82	89.55	98.96	98.43	99.65
7	77.35	82.06	95.99	89.90	98.96	98.61	99.65
8	78.57	83.62	96.17	90.07	99.30	98.96	99.83
9	79.44	84.84	96.52	90.59	99.30	99.13	99.83
10	79.79	86.06	96.69	90.94	99.30	99.48	99.83

For the rank 1, the best recognition rates of Borda Count, Bayes Fuse, Sum Rule, and SVM Fusion are 85.71%, 94.77%, 97.39%, 98.26%, which can be seen in Table 3, 3, 2 and 4 respectively. SVM fusion achieved the best performance when all three sources were combined, but the rest three methods obtained the best performance when only two sources from different biometric modalities were used. Bayes Fuse showed severe degradation of recognition rate in the rank 1 when combining three sources. This seems to be due to its explicit assumption of independence. Among various fusion methods and different combinations of sources, SVM fusion with three sources gave the best result but the difference of performance is not very significant in the top ranks except for Borda Count.

5 Conclusion

In this paper, we have investigated four fusion methods for multimodal identification using one fingerprint system and two face systems. The methods are (i) Borda Count, (ii) Bayes Fuse, (iii) Sum Rule and (vi) Binary Classification

Approach using SVM. These fusion methods can be categorized by a rank level or a score level method, and in addition they can be also classified into an unsupervised or a supervised method. Through the methods except for Borda Count, good performance improvements were obtained. We think that the reason for Borda Count's failure is in its information loss by uniform discretization of scores rather than its unsupervised learning. Level of fusion may be a more important factor for a performance gain than learning method in fusion for multimodal identification. However, in cases that similarity scores are not available, Bayes Fuse can be a good choice for its simplicity and relatively good performance. Binary Classification Approach has a potential advantage over the other methods. As it combines multiple sources, it can indicate whether a user is in a watch list or not at the same time. Future experiments will include a watch list using Binary Classification Approach.

References

1. Ross, A., Jain, A.K.: Multimodal biometrics: An overview. In: Proc. of 12th European Signal Processing Conf. (EUSIPCO), Vienna, Austria (2004) 1221–1224
2. Golfarelli, M., Maio, D., Maltoni, D.: On the error-reject trade-off in biometric verification systems. IEEE Trans. on Pattern Analysis and Machine Intelligence **19** (1997) 786–796
3. Ross, A., Jain, A.K.: Information fusion in biometrics. Pattern Recognition Letters **24** (2003) 2115–2125
4. Pan, S., Gil, Y., Moon, D., Chung, Y., Park, C.: A memory-efficient fingerprint verification algorithm using a multi-resolution accumulator array. ETRI Journal **25** (2003) 179–186
5. Turk, M.A., Petland, A.P.: Face recognition using eigenface. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Maui, Hawaii (1991)
6. Vapnik, V.: The Statistical Learning Theory. John Wiley & Sons, New York (1995)
7. Aslam, J.A., Montague, M.: Models for metasearch. In: Proc. of the 24th ACM SIGIR Conf. on Research and Developement in Information Retrieval, New Orleans, Louisiana (2001) 276–284
8. Aslam, J.A., Montague, M.: Bayes optimal metasearch: A probabilistic model for combining the results of multiple retrieval systems. In: Proc. of the 23rd ACM SIGIR Conf. on Research and Development in Information Retrieval. (2000) 379–381
9. KISIS: Korea information security industry support center. (http://www.kisis.or.kr)
10. XM2VTS: (http://xm2vtsdb.ee.surrey.ac.uk)

Between-Source Modelling for Likelihood Ratio Computation in Forensic Biometric Recognition

Daniel Ramos-Castro[1], Joaquin Gonzalez-Rodriguez[1], Christophe Champod[2], Julian Fierrez-Aguilar[1], and Javier Ortega-Garcia[1]

[1] ATVS (Speech and Signal Processing Group), Escuela Politecnica Superior
Universidad Autonoma de Madrid, E-28049 Madrid, Spain
{daniel.ramos,joaquin.gonzalez,javier.ortega}@uam.es
[2] Institut de Police Scientifique, Ecole de Sciences Criminelles
Universite de Lausanne, CH-1015, Lausanne, Switzerland

Abstract. In this paper, the use of biometric systems in forensic applications is reviewed. Main differences between the aim of commercial biometric systems and forensic reporting are highlighted, showing that commercial biometric systems are not suited to directly report results to a court of law. We propose the use of a Bayesian approach for forensic reporting, in which the forensic scientist has to assess a meaningful value, in the form of a likelihood ratio (LR). This value assist the court in their decision making in a clear way, and can be computed using scores coming from any biometric system, with independence of the biometric discipline. LR computation in biometric systems is reviewed, and statistical assumptions regarding estimations involved in the process are addressed. The paper is focused in handling small sample size effects in such estimations, presenting novel experiments using a fingerprint and a voice biometric system.

1 Introduction

The number of commercial applications of biometric systems has significantly increased in the last years. As a consequence, forensic applications of biometric systems arise then in a natural way. Forensic reporting in cases involving anthropomorphical or behavioral patterns can be assisted by using a biometric system. For example, a sample pattern is recovered at the scene of a crime (e. g., a fingermark) and a court of law requests an expert opinion on the comparison of such a *mark* with a template (e. g., a suspect's fingerprint) from a suspect. The aim of a forensic system in such a case is to report a *meaningful value* in order for the court to assess the *strength of the forensic evidence* in this context of identification of sources [1][2]. However, when a biometric system is used, this value cannot be given neither by a decision or a threshold nor directly by a similarity measure [1][3], because it may lead the forensic scientist to usurp the role of the court, responsible of the actual decision [4]. Our point is that commercial, score-based biometric systems are not suited for direct forensic reporting to a court of law as has been stated in previous work [1][3].

To overcome this difficulty, the application of a *likelihood ratio* (LR) paradigm suited for forensic evidence to score-based biometric system has been proposed

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1080–1089, 2005.
© Springer-Verlag Berlin Heidelberg 2005

[2][3][5][6]. In this paper we propose to use this LR framework. Following this approach the forensic scientist assesses and reports one meaningful value: the LR, that allows the court to progress to a posterior opinion starting from his prior opinion about the case before the forensic evidence analysis [5][7]. This logical Bayesian framework implies a change of opinion when new information is considered, i. e., when the weight of the evidence has been assessed [1]. LR Computation can be performed using the scores from any biometric system [3][8][9], the process being independent of the biometric discipline. Thus, the LR assessed from the system scores can be used for direct forensic reporting. Our recent work presents examples of LR computation using on-line signature, face and fingerprint biometric systems [9]. In [10], the ATVS forensic voice biometric system is presented and excellent results in NIST Speaker Recognition Evaluation 2004 [11] and NFI-TNO Forensic Evaluation 2003 [12] using robust LR computation algorithms are shown. LRs can be used to compare the strength of the evidence between different biometric systems and expert opinions, and allow the combination of evidence weights coming from different and independent systems [5].

The present paper describes briefly forensic interpretation and reporting using biometric systems by means of LR computation. Then the paper highlights statistical assumptions regarding estimations involved in LR computation [9][10]. The novel contribution is focused on small data set effects [13] using different estimation techniques. The paper is organized as follows: Sect. 2 describes the Bayesian analysis of forensic evidence and its motivation. In Sect. 3, the LR computation process is reviewed, statistical assumptions commonly considered are presented, and main approaches found in the literature in order to cope with them are reviewed. Sect. 4 presents new experiments regarding generalization against small data set effects using different estimation techniques for fingerprint and voice biometric systems. In Sect. 5, conclusions are extracted.

2 Forensic Interpretation of the Evidence

2.1 Score-Based Biometric Systems vs. Forensic Interpretation

The aim of commercial score-based biometric systems is to output a similarity measure (score) between a user of the system, represented by a biometric test pattern, and a claimed identity, represented by a biometric template. Biometric verification is a classification problem involving two classes, namely *target users* of the system and *non-target users* or *impostors*. A decision is made by comparing the output score with a threshold. Assessment of these systems can be done by means of decision theory tools such as ROC or DET curves [3].

The aim of forensic interpretation is different. Forensic evidence is defined as the relationship between the suspect material (samples of biometric patterns obtained from the suspect) and the mark (biometric pattern generally left in association with a crime of disputed origin) involved in a case. The role of the forensic scientist is to examine the material available (mark and control material) and to assess the contribution of these findings with regards to competing propositions arising from the circumstances and often the adversarial nature of

the criminal trial [6]. In sources attribution issues [2] such as the ones considered here, the prosecutor view will suggest that the suspect left the mark whereas the defense will support that an unknown contributor is the source [1][3][6]. The LR framework we suggest bellow imply that the forensic scientist will only guide as to the degree of support for one proposition versus the other and not comment, probabilistically or otherwise, on the hypotheses themselves [1][6]. This role differs fundamentally from the natural objectives of a commercial biometric system (i. e., making a decision) [3][4].

2.2 Bayesian Analysis of Forensic Evidence

The problems described above are handled elegantly when using the Bayesian analysis of forensic evidences [1][5][6]. Following this approach, the interpretation of the forensic findings is based on two competing hypotheses, namely H_p (*the biometric trace originates from the suspect*, also called *prosecutor hypothesis*) and H_d (*the biometric trace originates from any other unknown individual*, also called *defence hypothesis*). The decision of the judge or jury (in one word the *fact finder*) is based on the probabilities of the two hypotheses given all the information of the case, that can be split into *forensic information* (E), and *background information* (I) (i. e., all other information related to the case). Using the Bayes Theorem [5], we can write in odds form:

$$\frac{\Pr(H_p|E,I)}{\Pr(H_d|E,I)} = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \cdot \frac{\Pr(H_p|I)}{\Pr(H_d|I)} \tag{1}$$

In this way, the posterior probabilities needed by the fact finder can be separated into prior probabilities, based only on the background information, and a *likelihood ratio* (LR) that represents the strength of the analysis of the forensic evidence in the inference from prior to posterior odds:

$$LR = \frac{\Pr(E|H_p,I)}{\Pr(E|H_d,I)} \tag{2}$$

The role of the forensic scientist lies therefore with the assessment of this LR. The meaning of the LR is in essence *independent of the forensic discipline* [5], and its assessment in a case can involve computation (such as in [8][9][3][14]) or informed judgements expressed as subjective probabilities [15].

Assessment of forensic systems performance can be made using Tippett plots [3][9] (see Fig. 3), which are cumulative distributions of LR for targets (when H_p is true) and non target (when H_d is true) respectively.

3 Likelihood Ratio Computation in Score-Based Biometric Systems

As noted in [1], the numerator of the LR (Eq. 2), is obtained from knowledge of the within-source variability (WS) of the suspect material. This distribution can be estimated using scores obtained by comparing biometric patterns

(*controls*) from the suspect to templates originating from the same suspect. On the other hand, the denominator of the LR is obtained from knowledge of the between-source (BS) distribution of the mark, which can be estimated from scores resulting from the comparison of the mark with a set of biometric templates from a relevant population of individuals. The *evidence score* is computed by comparing the mark with the suspect biometric template. Finally, the LR value will be the ratio of the density of the evidence score under respectively WS and BS [3][8], as is shown in Fig. 1. As the LR is conditioned by the prosecutor (H_p) and defence (H_d) hypothesis and background information (I), the forensic scientist has to estimate the WS and BS distributions based on the data available in the case. Evidence scores significantly different from the data set used in distribution estimations will give a non-informative LR value of one.

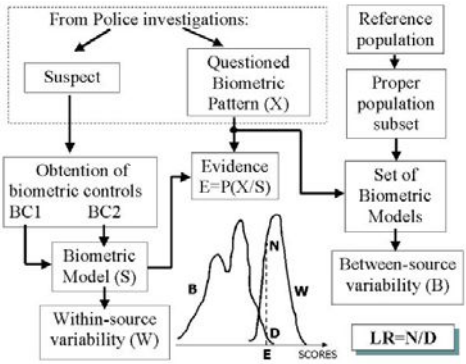


Fig. 1. LR Computation Steps

3.1 Statistical Assumptions

In the estimation of WS and BS distributions for LR computation, some assumptions have to be made. In order to estimate WS distribution, matching conditions between the suspect biometric template and controls (see Fig. 1) is needed [16][17]. However obtaining matching controls in real forensic casework can be a very difficult task, especially in some biometric disciplines, leading to a paucity of data. Therefore, generalization is desirable to avoid small sample size effects [13]. Approaches based on modelling WS distributions using databases can be found in [16]. More robust techniques based on additional knowledge about the system behavior are shown in [10], in which they can be also found procedures to optimize the use of the suspect data.

BS estimation problems related to mismatch between the considered relevant population and the conditions of the mark have been explored in [10] and [17] for voice biometric systems. In [18], corpus-based techniques are applied to reduce the mismatch between the population and suspect templates. Also, the nature of the population is conditioned to the circumstances of the case (I). The relevant population can then be reduced, either according to I, or because of the lack of

databases matching the conditions of the case under study. If the population size is small, non-matching conditions between population templates and questioned patterns can seriously degrade system performance.

The novel contribution of this work focuses on small sample size effects, not related to forensic issues of the partiality, poor quality or degradation of marks. Thus, the scenarios explored will postulate marks of quality comparable with the control material. In that sense, we have a symmetry in term of amount of information between the mark and the suspect material.

3.2 Between-Source Distribution Modelling Techniques

In this paper, we concentrate on BS modelling. We propose to assess two different estimation techniques, one parametric and one non-parametric, to model BS distributions. The parametric approach, proposed in [3], consist in modelling BS with a one-dimensional mixture of gaussian components:

$$p(x) = \sum_{m=1}^{M} p_m \cdot b_m(x) \tag{3}$$

where M is the number of mixtures used, and p_m are restricted to:

$$\sum_{m=1}^{M} p_m = 1 \tag{4}$$

Maximum Likelihood (ML) estimation using this parametric model is carried out by the Expectation-Maximization (EM) algorithm [19].

On the other hand, Kernel Density Functions (KDF) [19] are used. In this non-parametric technique the score-axis is divided in regions (*bins*) of length h. If N samples are available, and k_N of these samples fall in a bin, the probability estimated for that bin will be k_N/N. So the corresponding density will be:

$$\hat{p}(x) \equiv \hat{p}(x_0) \approx \frac{1}{h}\frac{k_N}{N},\ |x - x_0| \leq \frac{h}{2} \tag{5}$$

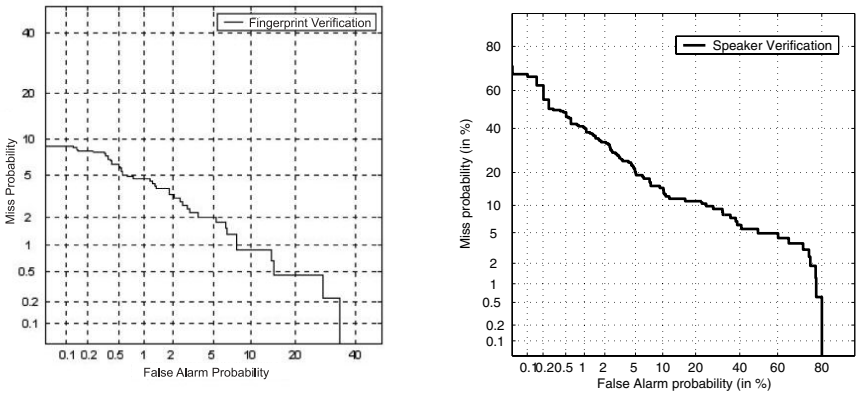
Using smooth functions ϕ, known as *kernels*, where $\phi \geq 0$ and:

$$\int_x \phi(x) \cdot dx = 1 \tag{6}$$

then the resulting estimated function is a legitimate pdf.

4 Experiments

In order to test the performance of the BS estimation techniques proposed, we present experiments using a fingerprint and a voice biometric system respectively.



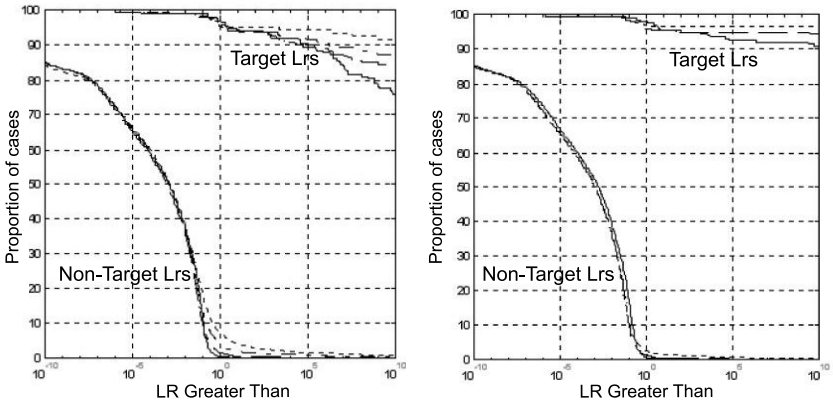
(a) (b)

Fig. 2. DET curve of: (a) ATVS reference fingerprint system with MCYT corpus, and (b) ATVS reference voice biometric system with NIST SRE 2004 sub-corpus

4.1 Databases, Experimental Protocol and Biometric Systems

For fingerprint experiments, the ATVS fingerprint recognition system based on minutiae comparison [20] has been used. A sub-corpus from the MCYT fingerprint database [21] has been selected, consisting of 50 users each one having 10 fingerprint samples. One sample per user will be used as reference biometric template. For score-based system performance assessment via DET plots, the 9 remaining samples will be used as test patterns (marks), so a total of $50 \times 9 = 450$ target trials and $50 \times 49 \times 9 = 22050$ non-target trials have been considered. For the forensic interpretation system, 5 (out of 9) fingerprint patterns have been used as biometric controls in order to obtain WS scores, and the remaining 4 will be used as marks. No technique will be used to predict degradation in WS distribution, as fingerprint biometric patterns are all acquired in the same conditions. Therefore, a total of $50 \times 4 = 250$ target trials and $50 \times 49 \times 4 = 9800$ non-target trials will be used for Tippett plot computation. Population data has been taken from the same corpus too.

For voice biometric system experiments, the ATVS UBM-MAP-GMM system [10] has been used. The scores used in the LR computation experiments are extracted from the ATVS results in the NIST Speaker Recognition Evaluation 2004 [11], using only a male subcorpus of 50 users, and all the trials defined in the evaluation for these users in the core condition, i. e., one conversation side (5 minutes) for training and one for testing. Strong mismatch on channel and language conditions is present in this data set, and it exists variability in the amount of speech per conversation side (as silence removal has not been performed). As only one speech segment is used as suspect biometric material, jackknife and prediction techniques described in [10] are used to perform robust WS estimation. In summary, a total of 163 target trials and 1969 non-target trials are performed. Population data consists of a channel-balanced English set of GMM models obtained from development corpora from past NIST SRE.



(a) (b)

Fig. 3. Tippett plots for fingerprint system estimating BS distribution with different techniques. (a): ML with M gaussian mixtures: M=1 (solid), M=3 (dashed), M=10 (dash-dot) and M=30 (dotted). (b): KDF with bin size h=10 (solid), h=3 (dashed), h=1 (dotted)

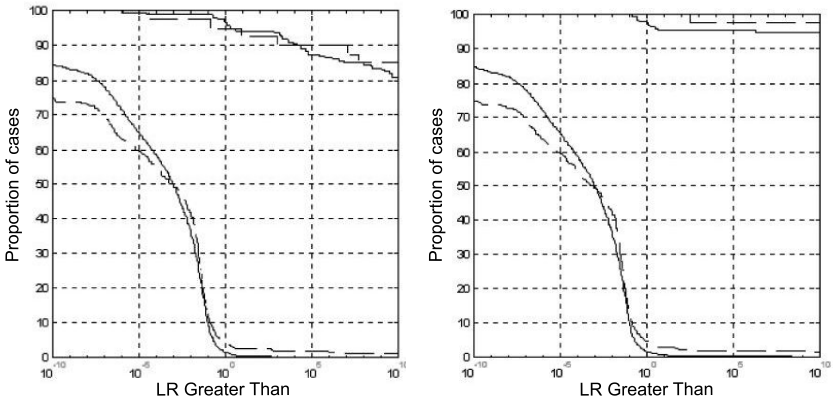
4.2 Results

Fig. 2 show the performance of the score-based biometric systems in the scenarios described. The performance of the forensic fingerprint system using the two techniques described is shown in Fig. 3. As can be seen in Fig. 3(a), performance of the forensic system in non-target trials degrades as the number of mixtures (M) increases. For M=30 mixtures, a small but not negligible proportion of non-target trials have values of LR greater than 100.000, which is alarming because the rate of misleading evidence for the non-target curve is critical in forensic systems [10]. The same conclusion can be extracted for KDF in Fig. 3(b), when the bin size h is small. This effect is due to an over-fitting effect of the BS model on the available data set.

Generalization against small sample size effects is inferred from Fig. 4. Two populations of $L = 50$ and $L = 10$ (obtained by sampling) biometric templates has been used. It can be seen that KDF and ML estimation presents very similar performance when the data set size is reduced. However, performance of targets for KDF estimation is better when population size decreases, which means over-estimation of target LRs due to over-fitting in BS distribution estimation.

In the experiments presented using voice biometrics system, ML estimation is performed to model BS distribution. In Fig. 5, the same effect noticed in Fig. 3(a) can be observed, i. e., the proportion of non-target trials having LR values greater than one grows as M increases.

Generalization for the voice biometric system is shown in Fig. 6. ML estimation of BS distribution with M=1 and M=8 has been used. It can be seen that as population sample size decreases, over-fitting in the data (M grows) implies degradation on system performance (i.e., bigger proportion of non-target $LRs > 1$, and over-estimated target LR).



(a) (b)

Fig. 4. Analysis of generalization effects with small sample-size data for the fingerprint biometric system. Population size: L=50 (solid) and L=10 (dotted). (a): ML with M=3 gaussian mixtures, (b): KDF with bin size h=3

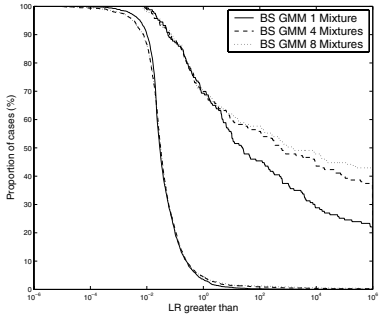
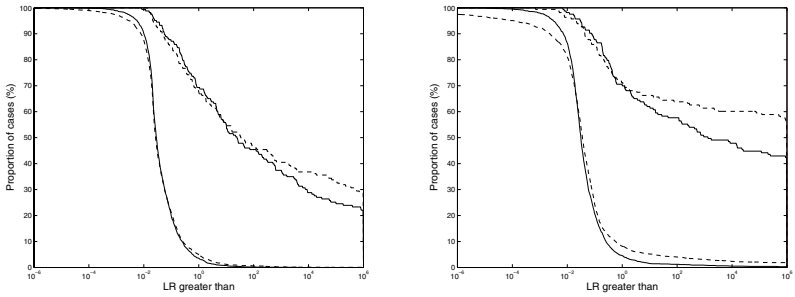


Fig. 5. Tippett plots for voice biometric system estimating BS distribution with ML and different number of gaussian mixtures



(a) (b)

Fig. 6. Analysis of generalization effects using ML for voice biometric system. Population size L=60 (solid) and L=12 (dotted) (a): M=1 gaussian; (b): M=8 gaussian

5 Conclusions

In this paper, we have shown how biometric systems can be used in forensic applications, using a LR as a measure of the strength of the evidence computed from the scores. The need of proper forensic reporting has been highlighted, as the fact finder needs a meaningful value to assist his decision making. Direct reporting using score-based biometric systems has been shown in the literature to be misleading, and we promote a LR based reporting system. Bayesian analysis of forensic evidence has been referred as the logical way for evaluating forensic findings. LR computation process has been reviewed, highlighting that it can be performed using any score-based biometric system, regardless of the biometric discipline. Statistical assumptions regarding estimations involved in the LR computation process have been discussed. The main contribution of the paper are the experiments regarding small sample size effects in BS estimation, which can appear in forensic casework when the relevant population is reduced, either because of the background information on the case (I) or the availability of databases matching the suspect biometric template conditions. It has been shown that the performance of the system degrades when BS distribution overfits the data set when its size is small, and misleading evidence in non-target trials can increase, which is a highly undesirable effect in forensic systems. LRs for target trials might also be over-estimated in these conditions.

Acknowledgements

This work has been partially supported by the Spanish MCYT projects TIC03 09068-C02-01 and TIC2000 1669-C04-01. D. R.-C. and J. F.-A. are also supported by a FPI scholarship from Comunidad de Madrid.

References

1. Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech Communication, **31** (2000) 193-203
2. Champod, C.: Identification/Individualization: Overview and Meaning of ID. Encyclopedia of Forensic Science, J. Siegel, P. Saukko and G. Knupfer, Editors. Academic Press: London (2000) 1077-1083
3. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ortega-Garcia, J.: Forensic Identification Reporting Using Automatic Speaker Recognition Systems. Proc. ICASSP (2003)
4. Taroni, F., Aitken, C.: Forensic Science at Trial. Jurimetrics Journal **37** (1997) 327-337
5. Aitken, C., Taroni, F.: Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Chichester (2004)
6. Evett, I.: Towards a uniform framework for reporting opinions in forensic science casework. Science and Justice **38(3)** (1998) 198-202
7. Kwan, Q.: Inference of Identity of Source. Department of Forensic Science, Berkeley University, CA (1977)

8. Meuwly, D.: Reconaissance de Locuteurs en Sciences Forensiques: L'apport d'une Approache Automatique. Ph.D. thesis, IPSC-Université de Lausanne (2001)
9. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J.: Bayesian Analysis of Fingerprint, Face and Signature Evidences with Automatic Biometric Systems Forensic Science International (2005) (accepted)
10. Gonzalez-Rodriguez, J., et al.: Robust Estimation, Interpretation and Assessment of Likelihood Ratios in Forensic Speaker Recognition. Computer, Speech and Language (2005) (submitted)
11. Home page of NIST Speech Group: http://www.nist.gov/speech
12. Van-Leeuwen, D., Bouten, J.: Results of the 2003 NFI-TNO Forensic Speaker Recognition Evaluation. Proc. of ODYSSEY (2004) 75-82
13. Raudys, S., Jain, A.: Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans. on PAMI **13(3)** (1991) 252-264
14. Curran, J.: Forensic Applications of Bayesian Inference to Glass Evidence. Ph.D. thesis, Statistics Department, University of Waikato, New Zealand (1997)
15. Taroni, F., et al.: De Finetti's Subjectivism, the Assessment of Probabilities and the Evaluation of Evidence: A Commentary for Forensic Scientists. Science and Justice **41(3)** (2001) 145-150
16. Botti, F., et al.: An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data. Proc. ODYSSEY (2004) 63-68
17. Gonzalez-Rodriguez, J., et al.: Robust Likelihood Ratio Estimation in Bayesian Forensic Speaker Recognition. Proc. Eurospeech (2003) 693-696
18. Alexander, A., et al.: Handling Mismatch in Corpus-Based Forensic Speaker Recognition. Proc. ODYSSEY (2004) 69-74.
19. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley (2001)
20. Simon-Zorita, D., et al.: Quality and Position Variability Assessment in Minutiae-Based Fingerprint Verification. IEE Proc. Vision, Image and Signal Processing **150(6)** (2003) 402-408
21. Ortega-Garcia, J., et al.: MCYT Baseline Corpus: A Bimodal Biometric Database. IEE Proc. Vision, Image and Signal Processing **150(6)** (2003) 395-401

The Effectiveness of Generative Attacks on an Online Handwriting Biometric

Daniel P. Lopresti and Jarret D. Raim

Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015, USA
{lopresti,jdr6}@cse.lehigh.edu

Abstract. The traditional approach to evaluating the performance of a behavioral biometric such as handwriting or speech is to conduct a study involving human subjects (naïve and/or skilled "forgers") and report the system's False Reject Rate (FRR) and False Accept Rate (FAR). In this paper, we examine a different and perhaps more ominous threat: the possibility that the attacker has access to a generative model for the behavior in question, along with information gleaned about the targeted user, and can employ this in a methodical search of the space of possible inputs to the system in an attempt to break the biometric. We present preliminary experimental results examining the effectiveness of this line of attack against a published technique for constructing a biometric hash based on online handwriting data. Using a concatenative approach followed by a feature space search, our attack succeeded 49% of the time.

1 Introduction

It is standard practice in biometric authentication to test a new system and report how well that system performs. In most cases, this information takes the form of FRR (False Reject Rate) and FAR (False Accept Rate) curves. Often, researchers perform studies with groups of university students and/or other volunteers playing the role of the attacker (*e.g.*, [3, 4]).

While such evaluations shed some light on the quality of the biometric, they do not always provide a full picture of the overall security provided by the system. In this paper, we examine a fundamentally different type of threat: the possibility that an attacker has access to a *generative model* for the behavior in question, *i.e.*, an algorithm which can be used to synthesize a signal that mimics true human input. Such models exist, for example, for speech and for handwriting, as well as for other physiological phenomena. By combining this with information gleaned about the targeted user (*e.g.*, samples of the user's speech or handwriting obtained surreptitiously), an adversary could conceivably conduct a methodical search of the space of possible inputs to the system in an attempt to break the biometric.

We present preliminary results for attacks on a published technique for constructing a biometric hash based on online handwriting data, and conclude by discussing possible areas for future exploration.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1090–1099, 2005.
© Springer-Verlag Berlin Heidelberg 2005

2 Related Work

Much work has been done in the area of testing biometric systems for security and performance. Brömme and Kronberg [1] proposed a system that integrates into state of the art operating systems such as Windows 2000 and Linux/UNIX. This framework allows the biometric system to log all information about its operation for later inspection. In this way, real-world conditions can be studied. The Brömme and Kronberg system is geared towards providing more accurate feedback to the current maintainers of an already-deployed biometric, however. It does not try to test a system from the point of view of a determined attacker, nor does it allow researchers to compare systems that have yet to be deployed.

Another technique to help researchers test biometrics based on handwriting has been developed by Vielhauer and Zoebisch [9]. This tool allows researchers to study forgeries generated by a human with access to static and dynamic representations of the true signal. The system presents the human forger with several different prompts containing increasing information about the targeted handwriting. The "attacker" first records a test sample with no information. He/she is then shown a static representation of the true writing and asked to input another test sample. Lastly, a dynamic representation of the handwriting is displayed and the user is allowed to input one more test sample. Depending on the characteristics of the test writer and the true signal, the accuracy of the forgeries will vary widely.

Another approach to more directly studying the security provided by a biometric system is presented by Monrose, *et al.* in [5]. This paper, which provides the primary motivation for our present work, describes several types of attacks against a speaker authentication system. The tested system extracts features from the user's voice, drawing entropy from both the passphrase spoken by the user and how the passphrase was spoken. These features are then used to extract a key from a data structure in which pieces of the true key are intermingled with random data. This process makes it difficult for an attacker in possession of the device to obtain any of the sensitive information stored on it.

The speech system in the above study was attacked using several methods. The first was a standard human impostor, whereby someone other than the true user tries to authenticate against the biometric. Next, a text-to-speech (TTS) system was used to generate sequences of phonemes for the passphrase, with various input parameters governing the type of speech produced. Lastly, a crude cut-and-paste attack was attempted, employing a large inventory of the true user's speech. Phonemes which had been manually labeled in the inventory were selected and concatenated to yield the targeted passphrase. Both the TTS and cut-and-paste attacks were able to out-perform random guessing, but did not work well enough to break the biometric. These results suggest, however, that as an attacker acquires more information, it becomes easier to breach the system.

In a test of another speech-based system, Masuko, *et al.* [4] attempted to use information about the pitch of voice samples to enable the system to reject synthetically created speech. They proved that current speech authentication

could be fooled over 20% of the time by using trained speech synthesis systems, and that pitch information was not useful in rejecting synthesized speech.

With generative models for handwriting appearing in the literature (*e.g.*, [2, 6]), we seek to adapt this style of investigation to the handwriting verification problem. We note that, as in [5, 8], we are not concerned with a user's one-and-only (*i.e.*, legal) signature, but rather the idiosyncratic way the user writes an arbitrary pre-selected passphrase of his/her own choosing.

3 Attack Models

To increase the amount of knowledge about the security provided by a given biometric, a model of the operation of the system is needed. This model must take into account all of the possible vulnerabilities of the system and provide ways for testing those vulnerabilities. By allowing a finer grained comparison of systems, individual components can be contrasted with one another. A system with low FRR and FAR might not be as secure as one with higher error rates, but a better-defined (more comprehensive and realistic) security model.

Because there are so many different kinds of information that could help an attacker breach a system, an exhaustive taxonomy is beyond the scope of this paper. We instead confine ourselves to exploring one line of attack, using techniques that should be generalizable to other scenarios.

For the present study, the types of handwritten inputs we consider include:

Class 1. Different User, Different Passphrase. Sometimes referred to as a "naïve forgery."
Class 2. Different User, True Passphrase. Different user writing the same passphrase as the true user.
Class 3. True User, Different Passphrase. True user writing something other than the passphrase.
Class 4. Concatenation Attack. Passphrase created from online samples of the true user writing non-passphrase material.
Class 5. True User, True Passphrase. The keying material, provided as a baseline for reference.

Certain of these input classes were chosen based on the types of attacks usually reported in the biometric literature. Class 1 is the typical brute-force type of attack while Class 2 is closer to a so-called "skilled forgery." In testing Class 3, we hope to show that even if the attacker has access to online samples of the true user's handwriting, more work must be done to use that information to reduce the possible search space.

The representative generative model in the current test is Class 4 (we plan to study other generative models in the near future). Here we employ samples of the user's handwriting collected separately from the passphrase. These samples are manually segmented into basic units, which can be individual characters, bigrams, trigrams, etc., and then labeled. The generative model accepts as input a labeled inventory and the targeted passphrase and produces a random sequence

of concatenated units that, when rendered, attempts to simulate the user writing the passphrase. Note that both the appearance of the writing as well as the dynamics are reproduced.

Lastly, Class 5 is provided as a baseline reference to contrast the other classes to the intended input. In most biometric systems, some allowances must be made to ensure that the true user is able to authenticate despite natural variations in handwriting.

As was the case in [5], our primary interest in this work is in offline attacks, a situation that might arise when the biometric is employed to generate a secure hash or cryptographic key to be used in protecting confidential information stored on a mobile device, for example. The handwritten input is provided to the system which generates a set of features as output. The range of acceptable inputs for the true user can be viewed as defining a subspace over the entire feature space. Ignoring the unlikely event of an exact match on the first attempt (a perfect forgery), the attacker's goal, then, is to explore the space around the feature vector returned by the forgery as rapidly as possible in the hopes of uncovering the correct setting. We assume, of course, that the attacker has no way of knowing whether the forgery is good enough to fall close to a true input until the match is actually found, but once that happens, the attacker is able to tell that the system has been broken. Hence, the attacker will conduct a methodical search, working outwards from the feature vector for a certain period of time before concluding that the forgery was not good enough and moving on to try another input.

While the attack models we have presented are quite simple, they are sufficient to motivate interesting tests of published biometrics, provide an indication of the associated combinatorics, and illustrate the difficulty (or ease) with which specific systems can be broken.

4 Experimental Evaluation

To examine the impact of the models described above, several example attacks were created. For testing purposes, we chose to implement the Vielhauer, *et al.* system [8] for biometric hashing. Based on a small data set, standard FRR and FAR measures were used to determine appropriate parameter settings for our later attempts at attacking the system. We then evaluated the effectiveness of each of the classes of inputs described in the previous section,

4.1 Data Sets

For our experiments, several small data sets were created. Two writers (the authors) wrote four different passphrases 20 or more times, which resulted in a total of 154 samples. The handwriting was collected using a Wacom Intuos digitizing tablet. While this data set is small in comparison to results typically reported in the literature, it is still possible to draw conclusions due to the specific nature of our study: we are not attempting to prove that a proposed

biometric is secure, rather, we are trying to examine whether attacks based on generative models can be successful. Since we are comparing the effectiveness of attack strategies and not the overall security of a system, the size of the data set is not a serious issue provided the phenomenon of interest, the breaking of the system, is seen to occur[1]. Two examples from this data set are shown in Fig. 1.

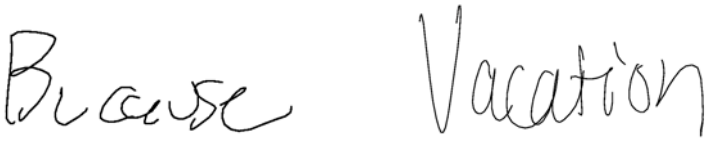
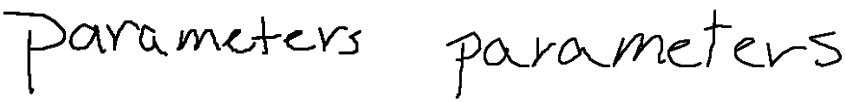


Fig. 1. Handwriting samples

As noted previously, to execute the concatenative attack (Class 4), it is assumed that the attacker has access to online handwriting samples of the targeted user as well as knowledge of the true passphrase. (It can also be assumed that the attacker has an offline image of the user's passphrase, but this was not used in our current study.) A separate set of online writing samples were labeled as to which stroke sequences corresponded to individual characters. This resulted in a corpus of possible n-gram combinations of the user's handwriting. To generate a synthetic handwritten passphrase, strokes were concatenated from the corpus to form the correct text of the passphrase. No scaling or smoothing was performed, however the individual stroke sequences were placed on a similar baseline and appropriate timestamps were recreated. An example of a passphrase synthesized using this approach appears in Fig. 2(b).



(a) Target passphrase. (b) Concatenative attack.

Fig. 2. Example of a concatenative attack

4.2 Biometric System

A detailed discussion of the Vielhauer, *et al.* biometric hash can be found in [8], but some knowledge of the system will be helpful for a greater understanding of the attacks discussed below. The system is based on 24 integer-valued features extracted from an online writing signal. The signal consists of $[x, y]$ position and timing information. Fourteen of the features are global, while the remaining

[1] A good analogy here are studies on the susceptibility of traditional password security systems to dictionary-based attacks. If a system with two passwords can be broken in such fashion, then certainly systems with larger numbers of passwords are even more susceptible. Nevertheless, we recognize the value of larger data sets and plan additional collection activities in the near future

ten features are concerned with segmented portions of the input obtained by partitioning the bounding box surrounding the ink into five equal-sized regions in the x- and y-dimensions. A listing of the features is provided in Table 1.

Table 1. Features employed in the Vielhauer, *et al.* biometric hash [8]

1. Number of strokes	13. Effective writing velocity in x
2. Total writing time (ms)	14. Effective writing velocity in y
3. Total number of samples (points)	15. Integrated area under x, segment 1
4. Sum of all local (x,y) minima and maxima	16. Integrated area under x, segment 2
5. Aspect ratio (x/y) * 100	17. Integrated area under x, segment 3
6. Pen-down / total writing time * 100	18. Integrated area under x, segment 4
7. Integrated area covered by x signal	19. Integrated area under x, segment 5
8. Integrated area covered by y signal	20. Integrated area under y, segment 1
9. Average writing velocity in x	21. Integrated area under y, segment 2
10. Average writing velocity in y	22. Integrated area under y, segment 3
11. Average writing acceleration in x	23. Integrated area under y, segment 4
12. Average writing acceleration in y	24. Integrated area under y, segment 5

To train the system to accept a given user, features are extracted and used to create a biometric hash where each feature generates a corresponding integer value. In addition, an interval matrix is created containing information needed for future testing. This information could be stored by the system itself or by the user in a portable format such as a USB key. A transitive enrollment system may be employed and would help in achieving a strongly-correlated set of sample data for the true user [7], but was not used for our experiments.

When a user attempts to authenticate, he/she provides a new handwriting sample and a claim to a certain identity. Features are extracted from the sample and passed through the hash generation. If a certain feature falls within the accepted range of values (plus some tolerance threshold), it generates the same integer hash. In our case, the size of the training set varied with the sample being tested, but ranged between 15 to 25 samples per class. To generate FRR and FAR curves for the system, the sample size for the training set cross-validation was varied between 5 and 10. Several potential tolerance values were also checked: 0.0, 0.01, 0.05, 0.1, 0.15, 0.2, and 0.3. The most promising graph is shown in Fig. 3, where the equal error rate tolerance threshold is found to be 0.15.

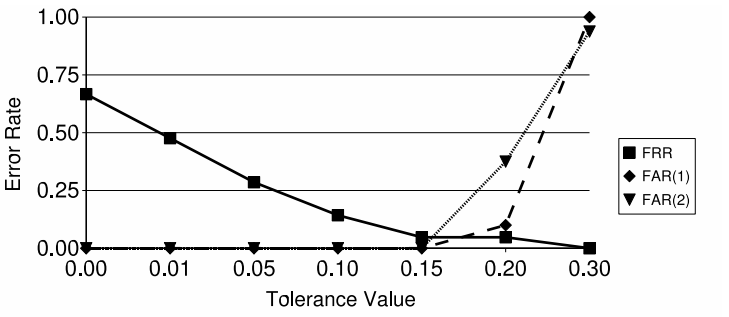


Fig. 3. Error rates for the Vielhauer, *et al.* hash on data used in our tests

As can be seen, the system performed quite well, with a FRR of 5%. In addition, two FAR curves were generated. The first is a naïve (Class 1) forgery, with one writer writing a passphrase that was used to test against the other writer writing a different passphrase. In this case, the FAR was 0%. In another test, one writer writing a passphrase was tested against the other writer writing the same passphrase (a Class 2 forgery). This case also resulted in a FAR of 0%.

To test how well each individual feature performed, a standard cross validation FAR test was run for all attack types. When a feature mismatch was detected, a note was made as to which feature failed and by what magnitude. This investigation showed that when a hash element is missed, the magnitude of the miss is generally ± 2 of the actual value. It was also interesting to see that several features yielded constant values for all of our handwriting samples. This could mean that these hash elements, and by extension the features that generated them, are not useful in the kinds of experiments we are performing. However, we believe more research is needed before drawing such conclusions.

4.3 Feature Space Search

As noted in Sect. 3, the search we performed attempts to find the true hash starting with the hash generated from the handwriting sample. The search begins with small alterations of the given hash and works outward until a predetermined time limit has been met. Our tests had a time limit of 60 seconds and were conducted on a Pentium 4 desktop PC running at 3.2 GHz with 1 GB of RAM. This machine was able to generate and check 540,000 search possibilities per second.

Results for the five different input classes are shown in Figs. 4 and 5. The first graph shows the min, mean, and max number of feature misses per sample, while the second graph shows how long it took the search process to correct the initial hash vector (when that was possible within the 60 second time limit).

It can be seen in Fig. 4 that the features used by the Vielhauer, *et al.* hash have several desirable qualities. Class 2 (Different User, Same Passphrase) and

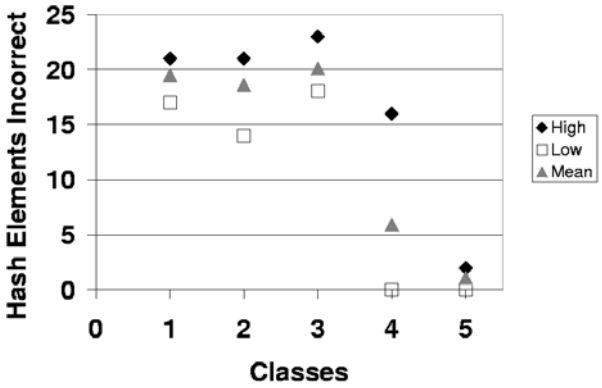


Fig. 4. Incorrect hash elements per input class

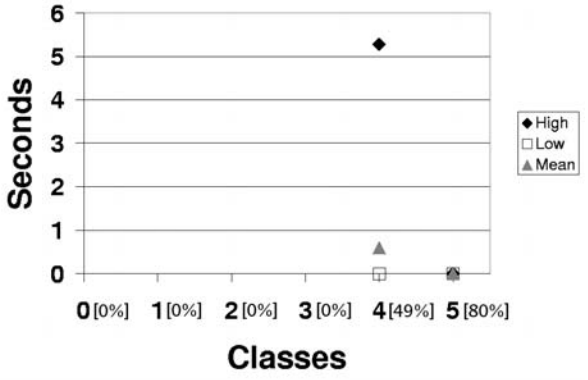


Fig. 5. Time to correct hashes per input class (with percent accepted)

Class 3 (Same User, Different Passphrase) have very similar means (18.6 vs. 20.1) and their maximum and minimum values are also comparable. This shows that some features of the biometric system are sensitive to the passphrase written, which accounts for the errors for Class 3, and some features are sensitive to the writer, which accounts for the errors seen in Class 2.

The mean incorrect number of hash elements for Class 1 is 21, which is higher than for all other classes, as would be expected. It is also of note that none of the Class 1, 2 or 3 hashes were broken in the 60 second search limit. While 60 seconds seems like a low bound, the number of possibilities for the search increases exponentially based on the number of incorrect hash elements. Still, more efficient search algorithms, faster machines, and longer runtimes could have an effect on the probability of the search finding the correct hash. As would be expected, Class 5 (keying material) was either accepted without any modifications or only required a maximum search time of 0.00019 seconds.

The concatenative attack, Class 4, displayed interesting behavior. Class 4 had the highest variance ranging from 0 to 16 hash elements incorrect, while none of the other classes had a range over 7. However, even with this distribution, the mean number of hash elements incorrect was still only 5.92, which put it below Classes 1, 2 and 3. The high variance means that many of the concatenated passphrases generated hashes with few to no hash elements incorrect. We note that 5% of Class 4 hashes were correct at the onset and required no search. When the standard 60 second search was allowed, 49% of the hashes were correctable, with an average search time of 5.28 seconds on those that were broken. Class 4 was the only class outside of the keying material (Class 5) that was able to generate hashes that required no search.

5 Conclusions

Popular measures for evaluating the performance of biometric systems may fail to capture certain kinds of threats. By limiting testing to human subjects and

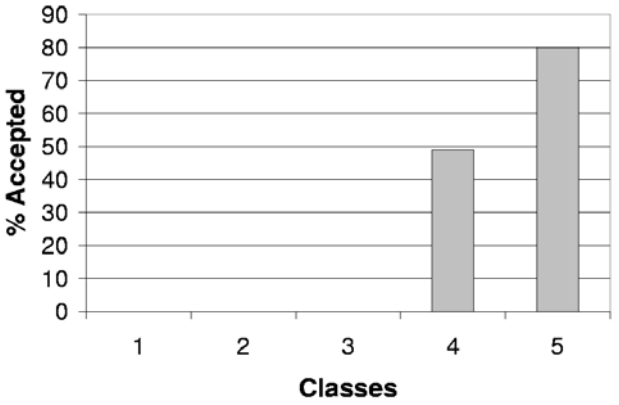


Fig. 6. Percentage of handwritten passphrases in each class accepted after search

reporting only results for FRR and FAR, the determined attacker is ignored. As more and more sensitive information is stored on portable computing devices, the incentives for breaking such systems becomes greater.

The attack models we have begun to study will help increase our understanding of potential flaws in biometric security, hopefully before they can be exploited. As can be seen in Fig. 6, we were able to achieve a 49% success rate using the concatenative attack described in this paper against a scheme for creating biometric hashes from online handwriting data[2].

The concatenative attack we have presented is only one possible avenue an adversary might take, as outlined in Sect. 3. We plan to study other forms of attack, including Plamondon's delta-log normal generative model [6] and Guyon's handwriting synthesis method [2]. These techniques will allow a full parameterization of the search space and may prove even more devastating. Other schemes for attempting to create secure hashes from a user's handwriting should likewise be evaluated in this fashion.

Testing with larger, more extensive data sets is also planned. By using tablet PC's and commercial signature capture tablets, we hope to better approximate a true distribution of users. These larger data sets will allow a more thorough examination of the feature space as well as the differences between handwritten passphrases and traditional "legal" signatures. Studying these issues in the context of other biometric measures, including speech, to build on the work that was first reported in [5], is another topic for future research.

Acknowledgments

The authors wish to thank Fabian Monrose for his helpful comments on an earlier draft of this paper.

[2] It is important to note that we are not singling out the system proposed by Vielhauer, *et al.* in [8] for criticism. The techniques presented in this paper should be applicable to many other behavioral biometrics with some amount of adaptation

References

1. A. Bromme and M. Kronberg. A conceptual framework for testing biometric algorithms within operating systems' authentication. In *Proceedings of the ACM Symposium on Applied Computing*, pages 273–280, 2002.
2. I. Guyon. Handwriting synthesis from handwritten glyphs. In *Proceedings of the Fifth International Workshop on Frontiers in Handwriting Recognition*, pages 140–153, 1996.
3. J. Lindberg and M. Blomberg. Vulnerability in speaker verification – a study of possible technical impostor techniques. In *EUROSPEECH*, pages 1211–1214, 1999.
4. T. Masuko, K. Tokuda, and T. Kobayashi. Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proceedings of the Sixth International Conference on Spoken Language Processing*, volume 2, pages 302–305, 2000.
5. F. Monrose, M. Reiter, Q. Li, D. Lopresti, and C. Shih. Towards speech-generated cryptographic keys on resource-constrained devices. In *Proceedings of the Eleventh USENIX Security Symposium*, pages 283–296, 2002.
6. R. Plamondon. A delta-lognormal model for handwriting generation. In *Proceedings of the Seventh Biennal Conference of the International Graphonomics Society*, pages 126–127, 1995.
7. C. Vielhauer, R. Steinmetz, and A. Mayerhöfer. Transitivity based enrollment strategy for signature verification systems. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, volume 2, pages 1263–1266, 2001.
8. C. Vielhauer, R. Steinmetz, and A. Mayerhofer. Biometric hash based on statistical features of online signatures. In *Proceedings of the Sixteenth International Conference on Pattern Recognition*, volume 1, pages 123–126, 2002.
9. C. Vielhauer and F. Zöbisch. A test tool to support brute-force online and offline signature forgery tests on mobile devices. In *Proceedings of the International Conference on Multimedia and Expo*, volume 3, pages 225–228, 2003.

Vulnerabilities in Biometric Encryption Systems

Andy Adler*

School of Information Technology and Engineering,
University of Ottawa, Ontario, Canada
adler@site.uOttawa.ca

Abstract. The goal of a biometric encryption system is to embed a secret into a biometric template in a way that can only be decrypted with a biometric image from the enroled person. This paper describes a potential vulnerability in such systems that allows a less-than-brute force regeneration of the secret and an estimate of the enrolled image. This vulnerability requires the biometric comparison to "leak" some information from which an analogue for a match score may be calculated. Using this match score value, a "hill-climbing" attack is performed against the algorithm to calculate an estimate of the enrolled image, which is then used to decrypt the code. Results are shown against a simplified implementation of the algorithm of Soutar et al. (1998).

1 Introduction

Traditional biometric technology tests for a match between a new image of an individual and the key biometric features of an original image stored in a biometric template. If the biometric software detects a match, further processing in a security system is activated. This often involves the release of security tokens or password codes to enable other applications. There are several potential concerns with such systems; in this paper we consider the concern that all the information needed to relase the codes must somehow be available to the software. It is therefore theoretically possible to compromise any traditional biometric system in order to gain secure access without presenting a biometric image [10]. At the same time, it may be possible to get information about the enrolled person from their biometric template [2][16].

Biometric encryption is designed to avoid these problems by embedding the secret code into the template, in a way that can be decrypted only with an image of the enrolled individual. [5][14]. Since the secret code is bound to the biometric template, an attacker should not be able to determine either the enrolled biometric image or secret code, even if they have access to the biometric software and hardware.

While such biometric encryption systems are not widely deployed, they appear to offer some compelling benefits for many applications [19]. The benefit of biometric encryption is perhaps most important for mobile applications of

* This work is supported by NSERC Canada

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1100–1109, 2005.
© Springer-Verlag Berlin Heidelberg 2005

biometrics, such as for cell phones or laptop computers, or in biometric-based identity cards, such as those designed into many new national passports. Another important application of biometric encryption is for control of access to digital content, with the primary interest being in preventing copyright infringement. Digital documents encoded with the biometric of the user(s) with approved access will presumably be subject to attacks, especially since both the documents and the software to access them will be widely distributed [10]. Finally, biometric encryption promises to help address the privacy concerns of biometric technology [17][19].

The primary difficulty in designing biometric encryption systems is the variability in the biometric image between data measurements. For example, a fingerprint image changes with applied pressure, temperature, moisture, sweat, oil, dirt on the skin, cuts and other damage, changes in body fat, and with many other factors. In the case of biometric encryption, this means that the presented biometric image cannot itself be treated as a code, since it varies with each presentation. For biometric encryption systems, this variability becomes especially difficult. An algorithm must be designed which allows an image from the enrolled person, with significant differences from the original, to decode the complete secret code. At the same time, an image from another person – which may only be slightly more different from the enrolled image – must not only not decode the secret, it must not be allowed to decode (or "leak") any information at all.

This paper develops one approach to attack biometric encryption algorithms, based on using any "leaked" information to attempt a "hill-climbing" of the biometric template. We show that this approach can successfully reconstruct a face image from a biometric encryption scheme based on [14][15]. We then discuss recent work in this area and some possible improvements to this attack.

2 Image Reconstruction from Biometric Templates

As discussed in [7], a biometric encryption system must have error tolerance, such that, for an enrolled image IM_{enroll}, it must be possible to perform the decryption for an input IM' which is sufficiently close (in which "close" is defined in some distance space appropriate to the biometric modality). For an IM' further from IM_{enroll} than some threshold, it must not only be infeasible to decrypt, but it must be impossible to obtain any statistical information about IM_{enroll}. The essence of the proposed attack on biometric encryption is to use this type of "leaked" information to iteratively improve an estimate of the enrolled biometric, which is then used to decrypt the secret code. Unfortunately, it is difficult to design an encryption algorithm to give complete information for a "close" answer, but no information for a slightly less accurate one [4][7][11][19].

In order to use the "leaked" information, it is necessary to construct a measurement which functions as a *match score*, ie. a measure which increases with the similarity of IM' to IM_{enroll}. Several authors have shown that, given access to match score data, it is possible to reconstruct a good estimate of an unknown enrolled image [16] from a fingerprint [9][20] or face recognition template [2].

These algorithms use a "hill-climbing" strategy. A test image is presented to a biometric algorithm and compared to an unknown enrolled image to obtain a match score. Then, iteratively, modifications are made to the input, and those that increase the match score are retained. Eventually, a best-match image is generated, which resembles the essential features of the unknown enrolled image, and is able to compare to it at high match score. In order to protect against this attack, the BioAPI [3] specifies that match scores should be quantized. However, recently, we have shown that the hill-climbing attack can be modified to overcome the effects of quantization [1] (for reasonable levels of quantization, ie. where one quantization level corresponds to a 10% change in match confidence).

Tests in this paper show that the modified hill-climbing algorithm is required for attacks against the biometric encryption algorithm. This appears to be because match scores calculated from biometric encryption algorithms are not easily related to traditional biometric match score values, and often it is only possible to calculate a quantized value. For example, with an error correcting code, the match score may be the number of bits that require correction, resulting in a heavily quantized score.

2.1 Quantized Hill-Climbing

This section describes the quantized hill climbing algorithm used to the attack the biometric encryption technique [1]. It has been shown to work successfully for face recognition systems; however, recent work [9][19] suggests that it is extensible to fingerprint biometrics. The algorithm has the ability to obtain match scores (MS) of the target compared to an arbitrarily chosen image (IM). We represent this function as:

$$MS = \mathsf{compare}(IM, IM_{enroll}) \tag{1}$$

A schematic diagram of this algorithm is shown in Fig. 1. It is implemented as follows:

1. *Local database preparation*: A local database of frontal pose face images is obtained. Images are rotated, scaled, cropped, and histogram equalized.
2. *Eigenface calculation*: Use a principle components analysis (PCA) decomposition to calculate an set of eigenimages (or eigenfaces) from the local image database [18], using the method of Grother [8]. Divide each image into four quadrants (Fig. 1, left). Quadrant eigenimages ($EF_{i,quadrant}$) are then defined to be equal to EF_i within the quadrant and zero elsewhere. The edge of each quadrant is then smoothed to provide a gradual transition over 10% of the image width and height.
3. *Initial image selection*: Choose an initial estimate (IM_0), which is subsequently iteratively improved in the next step. The selected image could be random, or could be the one with the largest MS.
4. *Iterative estimate improvement*: Iterate for step number i. Repeat iterations until MS is maximum, or there is no more improvement in MS.

(a) Randomly select a quadrant Q. The diametrically opposite quadrant is referred to as OQ.
(b) Randomly select an eigenimage, k; the component in Q is $EF_{k,Q}$
(c) Generate an image RN, consisting of random Gaussian noise in OQ and zero elsewhere.
(d) Calculate the amount of RN which reduces the quantized match score by one quantization level. Using a bisection search, calculate a minimum value n such that

$$\mathsf{compare}(IM_i, IM_{enroll}) > \mathsf{compare}(IM_i + nRN, IM_{enroll}) \tag{2}$$

(e) Iterate for j for a small range of values c_j

$$MS_j = \mathsf{compare}(IM_i + nRN + c_j EF_{k,Q}, IM_{enroll}) \tag{3}$$

(f) Select j_{max} as the value of j for the largest MS_j.
(g) Calculate

$$IM_{i+1} = IM_i + c_{j_{max}} EF_{k,Q} \tag{4}$$

(h) Truncate values to image limits (ie. 0 to 255) if any pixel values of IM_{i+1} exceed these limits.

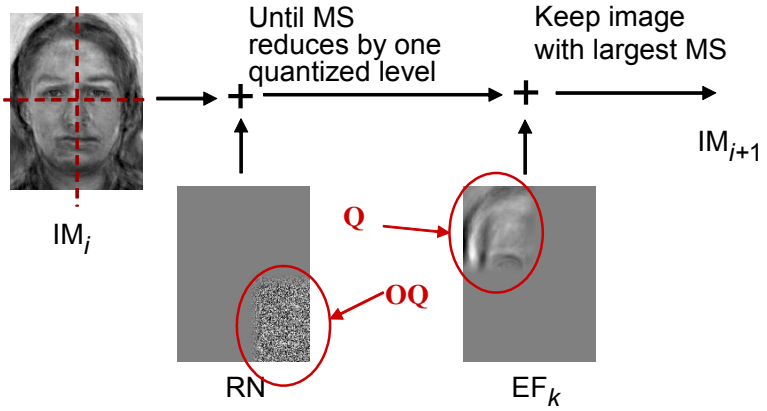


Fig. 1. Schematic diagram of the hill-climbing algorithm for quantized match scores. In each iteration, the candidate image is first "worsened" with the addition of random noise to a quadrant, until the match score is below a quantized level. Then a component of an eigenimage is added to the opposite quadrant, and the maximum match score output is retained

Because the quantized match score will not normally give information to allow hill climbing, a carefully chosen level of noise is introduced into the opposite image quadrant, in order to force the quantized score into a range where its information can once again be used. The local database does not need to resemble the target image, and may be one of the many freely available face image databases (for example [12][13]).

3 Biometric Encryption

This paper considers the fingerprint biometric encryption algorithm of Soutar et al. [14]. This algorithm was chosen because it represents a concrete system which has been implemented and for which the details are well described. Bioscypt Inc. (the employer of Soutar) has indicated that significant enhancements were made to this algorithm after the published version. However, this paper simply presents a framework for an attack, and not necessarily a break of a specific, implemented, algorithm. For a review of other recent biometric encryption systems, refer to [7][19].

Enrollment requires several sample images, and a secret code, and creates a template binding the code to the images. This differs for some other systems, such as that of Davida et al. [5][6], in which the biometric image forms a unique key. The system under consideration [14] calculates a template related to the input image by frequency domain correlations. We describe a simplified operation of this system, using slight variations in notation from [14]. During enrollment, an average image f_0 is obtained (with 2D Fourier transform $F_0(u)$) from multiple samples of the input fingerprint, after suitable alignment. In order to encode the secret, a random code is chosen and encoded as a phase-only function $R_0(u)$ such that the amplitude is one and the phase is $e^{\pm\pi/2}$ (selected randomly). Using F_0 and R_0, a filter function $H(u)$ is calculated based on a Wiener inverse filter, as

$$H_0 = \frac{F_0^* R_0^*}{F_0^* F_0 + N^2} \tag{5}$$

where * denotes the complex conjugate, and N^2 the image noise power. For this algorithm, N encodes the expected variability between images. As N increases, an image more dissimilar from the one enrolled can decrypt the code, at the expense of a smaller secret.

In order for biometric encryption to allow for variability in the input image, the secret code must be robustly encoded, using some sort of error correcting code (ECC) framework. [14] uses a simple ECC based on Hamming distances and majority decision. The secret is encoded by linking it with the sign of the complex component R_0. Each bit of the secret is associated with L locations in R_0 with the same phase angle. These associations are then stored in the template in a "link table". Majority decision requires that L be odd; [15] appears to recommend $L = 5$. For example, if the 4^{th} bit of the secret is a 1, position 4 of the link table will point to five positions in R_0 with a phase of $e^{+\pi/2}$, while if the bit is 0, position 4 will point to five positions with phase $e^{-\pi/2}$. The template is created containing the following information: H_0, the link table, a cryptographic hash of the secret, and an identifier. The cryptographic hash and identifier are to detect errors in storage and software processing, and do not concern us here.

During *key release*, a new image f_1 is acquired. This image is deconvolved with the filter H_0 to calculate R_1, an estimate of R_0.

$$R_1^* = sign(imag(H_0 F_1)) \tag{6}$$

It appears that the sign of the imaginary component of the phase of R_1 is the most robust feature of this calculation [15]. If F_1 is from the same individual as F_0, then R_1 should be a good estimate of R_0. The link table is used to extract the phase locations into which each bit is encoded. Since $R_1 \neq R_0$, some phase elements will be incorrect; however, if R_1 is sufficiently close, the use of majority decision should allow the correct value of the secret to be obtained.



Fig. 2. Sample images for an implementation of the biometric encryption technique of [14] applied for a face recognition. *Left*: Image f_0 averaged from five samples. *Right*: Template h_0 including the random phase encoded elements

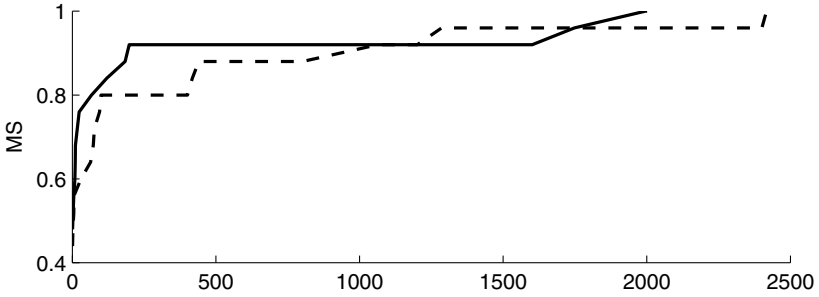


Fig. 3. Match score MS versus iteration number. The match score is calculated as the number of bit positions matching in the template. A MS of 1.0 indicates a perfect match. Solid and dashed line corresponds to top and bottom images in 4, respectively

4 Results

In order to apply the attack of section 2.1, it is necessary to create a match score from the template. For the biometric encryption system of [14] this is relatively straightforward. If $R_1 = R_0$, then all phases corresponding to each bit position in the link table will be equal, while for a random image, approximately half of the elements will match. We thus create a match score MS from the R_1 based on the difference between the number of ones and zeros in the link table, as

$$MS = \frac{1}{LB} \sum_{i=1}^{B} \left| \sum_{j=1}^{L} (LT_{ij} = 0) - \sum_{j=1}^{L} (LT_{ij} = 1) \right| \tag{7}$$



Fig. 4. Sample images of IM_k for as a function of iteration for two different initial images (top and bottom row). Left image is IM_0 and right image yields $MS = 1.0$

where LT_{ij} is the value of the link table entry for the j^{th} element of bit i, and B is the number of bits of secret. The maximum MS is 1; the minimum possible MS is $\frac{1}{L}$, and statistical considerations show a random image will typically give $MS = \frac{3}{8}$ of the maximum for $L = 5$.

We implemented the algorithm of section 3 for use with face recognition biometrics; the only modification required was to test which part of the Fourier transformed image F_0 produced reliable phase values to be encoded in the link table. The 13×13 low frequency 2D Fourier components appeared to be the most reliable for this application. The advantage of this implementation is that the framework and software previously developed for hill-climbing for face recognition in [1] would be applicable. On the other hand, such an algorithm is not realistic. Because face recognition data is not very distinctive, it would not be possible to encode many bits of a key (our initial results would suggest a maximum of about 20 bits). A template was created using 5 images from the NIST Mugshot Identification Database [12], and 20 secret bits were encoded using $L = 5$. In order to illustrate the power of the algorithm, an initial image intentionally different from the template was chosen. Fig. 2 shows an image of the averaged enrollment images from the template (f_0), and the encoded template (h_0). All images were scaled and rotated to have a common size and eye locations.

Results show that the template recreation algorithm is quickly able to attain a perfect match to F_0 ($MS = 1$), even though the resulting images are not very similar to the enrolled image. This is significantly larger than match values for other images of the enrolled individual (which were typically accurate to $MS = 0.82 - 0.86$). Fig. 3 shows the graph of MS versus iteration number for $L = 5$, while Fig. 4 shows a selection of images IM_k of the progress of the algorithm for $L = 5$ for two different initial images. There is an initial rapid

increase in *MS* after which the algorithm shows a more gradual improvement. It is interesting to note that *IM* begins to show some similar features to f_0 as iteration progresses. For example, the position of eyebrows, and shape of eyes, nose and chin and outline of the face begin to show a resemblance. One interesting aspect is that the hill-climbing algorithm does not seem to terminate with a final good estimate of the template image. Perhaps biometric encryption allows several possible variants of the enrolled image to match.

5 Discussion

This paper presents an approach to attack biometric encryption algorithms in order to extract the secret code with less than brute force effort. A successful result was obtained for a simplified version of the biometric encryption algorithm of [14]. Essentially, this attack requires that the some information be "leaked" from the biometric match for sample images very dissimilar from the enrolled one. This leaked information is used to construct a match score, which is subsequently used to iteratively improve an estimate.

While this work was implemented against a specific algorithm [14], several more recent systems have been proposed, which appear to be somewhat less susceptible to this vulnerability. For example, the fingerprint algorithm of [4], encodes the secret as the coefficients of a Galois field polynomial. Minutiae points are encoded as pairs (x_i, y_i) where x_i is a minutiae point, and y_i is a point on the polynomial. Additionally, numerous "chaff" points are encoded, in which the value of y_i is random. During key release, the minutiae of the new fingerprint image are calculated, and the points x_i closest to the minutiae are chosen. The y_i corresponding to these points are used to estimate the polynomial, using a Reed-Solomon error correcting code framework. If enough legitimate points are taken, the correct polynomial will be obtained and the correct secret decrypted. This encryption technique is based on the "fuzzy vault" technique of [11]. An interesting generalization of this scheme is given by the "secure sketches" of [7]. We believe that it may be possible to use the attacks of this paper against the biometric encryption technique of [4], even though Juels and Sudan [11] were able to give a proof of security. A key assumption for security proof is that the data held in the "fuzzy vault" are random. The data of [4], however, are not. Firstly, biometric data is inherently structured – otherwise hill-climbing wouldn't be possible. Secondly, the need to carefully place chaff minutiae points sufficiently far from legitimate ones is another source of non-randomness. However, at this time, we are not able to demonstrate an attack against this technique.

In their analysis, Uludag et al. [19] note that most proposed biometric encryption systems only appear to account for a "limited amount of variability in the biometric representation." In order to quantify this notion, experiments were conducted by them to estimate the variability in fingerprint minutiae. Matched fingerprint pairs were imaged and minutiae locations identified by a human expert, which was assumed to give an upper bound on system performance. Using these data, the algorithm of [4] was analyzed to estimate the *FMR*/*FNMR* trade-

off curve during key generation and key release. Results were surprisingly poor; an equal error rate of 6.3% can be estimated from the results, although the authors note that there are a limited number of feasible operating points. This means that such systems could be feasibly attacked by successively presenting biometric samples from a representative population.

In conclusion, this paper has presented a scheme that appears to show vulnerabilities in biometric encryption systems. The attacker can regenerate an estimate of the enrolled biometric image and use it to release the stored secret. The attacker considered here, who has access to biometric templates and authentication software, is quite plausible, as such biometric templates may be stored in standardized formats on identity documents or portable devices.

References

1. Adler, A.: "Images can be regenerated from quantized biometric match score data", *Proc. Can. Conf. Elec. Comp. Eng.* 469–472 (2004)
2. Adler, A.: "Sample images can be independently restored from face recognition templates" *Proc. Can. Conf. Elec. Comp. Eng.* 1163–1166 (2003)
3. BioAPI Consortium: *BioAPI Specification* `http://www.bioapi.org/BIOAPI1.1.pdf` 1163–1166 (2001)
4. Clancy, T.C., Kiyavash, N., Lin, D.J.: "Secure smartcard-based fingerprint authentication" *Proc. ACMSIGMM 2003 Multimedia, Biometrics Methods and Applications Workshop* 45–52. (2003)
5. Davida, G.I., Frankel, Y., Matt, B.J.: "On enabling secure applications through off-line biometric identification" *Proc. IEEE Symp. Privacy and Security* 148–157 (1998)
6. Davida, G.I., Frankel, Y., Matt, B.J., Peralta, R.: "On the relation of error correction and cryptography to an offline biometric based identification scheme" *Proc. Conf. Workshop Coding and Cryptography (WCC'99)* 129–138.
7. Dodis, Y., Reyzin, L., Smith, A.: "Fuzzy Extractors and Cryptography, or How to Use Your Fingerprints", Proc. Eurocrypt'04, (2004) `http://eprint.iacr.org/2003/235/`
8. Grother, P.: "Software Tools for an Eigenface Implementation" National Institute of Standards and Technology, (2000) `http://www.nist.gov/humanid/feret/`
9. Hill, C.J.: *Risk of Masquerade Arising from the Storage of Biometrics* B.S. Thesis, Australian National University, 2001 `http://chris.fornax.net/biometrics.html`
10. Kundur, D., Lin, C.-Y., Macq, B., Yu, H.: "Special Issue on Enabling Security Technologies for Digital Rights Management" Proc. IEEE **92** 879-882, (2004).
11. Juels, A., Sudan, M.: "A fuzzy vault scheme" *Proc. IEEE Int. Symp. Information Theory* 408 (2002)
12. National Institute of Standards and Technology (NIST): *NIST Special Database 18: Mugshot Identification Database (MID)* `http://www.nist.gov/srd/nistsd18.htm`
13. Phillips, P.J., Moon, H., Rauss, P.J., Rizvi, S.: "The FERET evaluation methodology for face recognition algorithms" *IEEE Trans. Pat. Analysis Machine Int.* **22** 1090–1104 (2000)
14. Soutar, C., Roberge, D., Stoianov, A., Gilroy, R., Vijaya, B.: "Biometric Encryption using image processing", Proc. SPIE Int. Soc. Opt. Eng., **3314** 178-188 (1998)

15. Soutar, C., Roberge, D., Stoianov, A., Gilroy, R., Vijaya, B.: "Biometric Encryption: enrollment and verification procedures", Proc. SPIE Int. Soc. Opt. Eng., **3386** 24-35 (1998)
16. Soutar, C., Gilroy, R., Stoianov, A.: "Biometric System Performance and Security", Conf. IEEE Auto. Identification Advanced Technol., (1999). `http://www.bioscrypt.com/assets/security_soutar.pdf`
17. Tomko G.: "Privacy Implications of Biometrics - A Solution in Biometric Encryption", 8th Ann. Conf. Computers, Freedom and Privacy, Austin, TX, USA, (1998).
18. Turk, M.A., Pentland, A.P.: "Eigenfaces for recognition" *J. Cognitive Neuroscience* **3** 71–86 (1991)
19. Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.K.: "Biometric Cryptosystems: Issues and Challenges", *Proc. IEEE* **92** 948–960 (2004)
20. Uludag, U.: "Finger minutiae attack system" *Proc. Biometrics Conference*, Washington, D.C. USA. Sept. (2004)

Securing Electronic Medical Records Using Biometric Authentication

Stephen Krawczyk and Anil K. Jain

Michigan State University, East Lansing MI 48823, USA
{krawcz10,jain}@cse.msu.edu

Abstract. Ensuring the security of medical records is becoming an increasingly important problem as modern technology is integrated into existing medical services. As a consequence of the adoption of electronic medical records in the health care sector, it is becoming more and more common for a health professional to edit and view a patient's record using a tablet PC. In order to protect the patient's privacy, as required by governmental regulations in the United States, a secure authentication system to access patient records must be used. Biometric-based access is capable of providing the necessary security. On-line signature and voice modalities seem to be the most convenient for the users in such authentication systems because a tablet PC comes equipped with the associated sensors/hardware. This paper analyzes the performance of combining the use of on-line signature and voice biometrics in order to perform robust user authentication. Signatures are verified using the dynamic programming technique of string matching. Voice is verified using a commercial, off the shelf, software development kit. In order to improve the authentication performance, we combine information from both on-line signature and voice biometrics. After suitable normalization of scores, fusion is performed at the matching score level. A prototype bimodal authentication system for accessing medical records has been designed and evaluated on a small truly multimodal database of 50 users, resulting in an average equal error rate (EER) of 0.86%.

1 Introduction

An increased need for a reliable authentication scheme has emerged in the health care industry as a result of the movement toward electronic medical records and the recently approved governmental regulations in the United States. Every year, billions of patients in the United States visit doctor's offices, clinics, Health Maintenance Organizations (HMO), hospitals, and other health care providers [2]. Each of these visits either generates a new medical record or adds to an existing one, necessitating the retrieval of a particular record. The procedure by which these records are stored and retrieved is undergoing a change toward a system that will better utilize modern technology. Security risks involved with this new system of archiving and retrieving patient records has brought about the onset of several government regulations pertaining to the protection and privacy of medical records which in turn has increased the need for a reliable user authentication scheme in this domain.

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1110–1119, 2005.
© Springer-Verlag Berlin Heidelberg 2005

1.1 Electronic Medical Records

A medical record can span hundreds of pages consisting of text, graphs, and images. It contains information on treatments received, medical history, lifestyle details, family medical history, medications prescribed, and numerous other items pertinent to an individual's health. In the interests of the integrity of the health care industry and good patient care, it is recommended that these records should be retained for as long as possible. For these factors alone, it is obvious that the move toward electronic data capture will greatly assist in the storage and management of patient records. Although this change is long overdue, the health care industry has only recently begun to convert their paper records to electronic form using electronic medical record (EMR) systems [3, 14].

1.2 Federal Regulations

The automation of health care information management has created increasing governmental and societal concerns about the security of computerized health care data. While the health care industry has incorporated electronic medical records, data repositories, networking, Internet access, and other new technologies into its various process, the corresponding security measures have not been enhanced. Many weaknesses have been identified in existing health care security measures from past operations [6]. The Health Insurance Portability and Accountability Act (HIPAA), which set the standards to ensure the security and integrity of patient information that is maintained or transmitted electronically, took effect in April 2003 [5]. Patients are assured, under HIPAA regulations, that their medical records will be used only by individuals directly involved in their medical treatments, payment of their bills, and health care operations. Any other individual or organization wishing to access a patient's medical record would require specific authorization by that patient. These regulations also attempt to ensure that when the medical records are properly disclosed, only the minimum amount of information necessary shall be released.

1.3 Tablet PC

Since it is convenient for a health care professional to have a patient's record readily available when prescribing or administering treatment, many health care facilities have adopted the use of tablet PCs as access devices to retrieve and edit a patient's record. The tablet PCs are easy to use and are able to access a patient's data through wireless access points. The widespread deployment of these wireless access points in hospitals and other facilities presents new security problems where only authorized users of the tablet PC are permitted to view the requested medical records.

1.4 Biometric Authentication

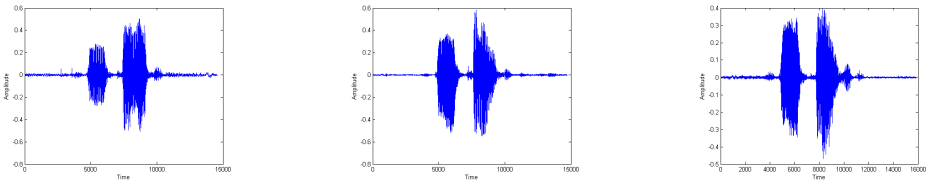
It is widely recognized that biometric authentication offers a number of advantages over traditional token-based (e.g. ID cards) or knowledge-based (e.g.

passwords) systems [12]. Several companies have realized these security benefits and have integrated biometrics into their EMR systems that use modalities such as the fingerprint and iris [1, 3, 14]. Additionally, multimodal biometric systems can overcome many of the limitations of a unimodal biometric system and will be the focus of this work [10]. In order to meet the guidelines of the HIPAA regulations, both health professionals and patients must be given access to medical records. Taking into account the requirements of both these groups (health professionals and patients), our biometric authentication system uses the voice and signature modalities. These modalities are unobtrusive and emulate the current, already well accepted system whereby a patient authenticates herself when seeking treatment or visiting a doctor's office for consultation. A typical scenario consists of a patient telling his or her name to a receptionist and then signing a release form. In addition, health professionals are already beginning to use tablet PCs to access patient records which are equipped with a stylus/pen and an internal microphone. Using the voice and signature modalities, our biometric authentication system can be seamlessly integrated into a tablet PC without any extra hardware.

2 Voice and Signature Verification

2.1 Voice Verification

In our authentication system, both voice identification and verification are utilized. The difference between voice identification and verification is that voice identification involves identifying a speaker out of a group of templates (1 to N matching) whereas verification deals with verifying whether an utterance matches with a specific user's template (1 to 1 matching). A user template is visually depicted in figure 1, and, as shown, can contain high intra-class variance. The voice biometric is used for authentication in such companies as Banco Bradesco, the largest bank in Brazil, the United Kingdom government's Intensive Supervision and Surveillance Program for fighting crime, and other major financial institutions for access to personal accounts and information [4]. In this work, both voice identification and verification are performed using the Nuance Verifier SDK [11]. The Nuance recognition and verification engines use Hidden Markov Models (HMM) to provide a mapping from sampled speech to phonetic units. Continuous-density HMMs are utilized, where the relationship between acoustic frames and states is modeled using a mixture of Gaussians [13]. These HMMs are set up in a hierarchical fashion so that after sampled speech is mapped to phonetic units, the resulting phonetic sequence is then mapped to the corresponding word sequence. The probability from the last Markov chain in the sequence is used as the verification score. Verification is text-independent while identification is text-dependent. Our system uses the same utterance for both identification and verification and accordingly the same phrase used in enrollment must also be used for verification. A minimum of two utterances is needed to train the Markov model. Each voiceprint will usually require 20KB of memory. Typical accuracy figures of the verifier are reported as being 99% or higher.



(a) (b) (c)

Fig. 1. Voice intra-class variability. (a), (b), and (c) are three waveforms (amplitude vs. time) from a single user who spoke his first and last name three different times

2.2 On-Line Signature Verification

Handwritten signatures are frequently used to authenticate financial transactions or the contents of a document, where the verification of these signatures is usually done by human visual inspection. Much work has been done in the effort to automate the process of signature verification because of its long standing acceptance in many applications. The main disadvantage of using this biometric is its inherent high intra-class variability, as shown in figure 2. The signature verification algorithm used in this work is a modified version of the algorithm reported in [7] and the details are described in [15]. The input to the algorithm is both the dynamic (temporal) and spatial information of the writing. Features such as the change in x and y coordinates between subsequent points in the signature and the pen pressure are extracted to form a feature vector at each point. An input signature is compared with an enrolled signature by using dynamic time warping (DTW), to find an alignment between the points in the two signatures such that the sum of the differences between each pair of aligned points is minimal. The resulting difference value is used as the verification score. A training set of signatures is used to both calculate user-dependent statistics and to compare against an input signature. After performing user-normalization and dimension reduction techniques, the resulting score is combined with a global feature system score to produce a final distance value. This global feature system extracts twenty global features and performs matching using the Mahalanobis distance. The size of the templates for each user is on average 30KB. The accuracy of the algorithm has an EER of 14.25% on skilled forgeries and 0.57% on random forgeries using the first 40 users from the SVC database [16].



(a) (b) (c)

Fig. 2. Signature intra-class variability. (a), (b), and (c) are three signatures from a single user

3 Biometric Fusion

Common problems that may occur in unimodal biometric systems are noise in sensed data, intra-class variations, inherent distinctiveness, and spoof attacks. Many of these limitations imposed by unimodal biometric systems can be either overcome or reduced by using multiple biometric traits. Multimodal systems have demonstrated higher accuracy due to the fact that they use multiple biometric modalities and combine independent evidence to make a more informed decision. If any of the limitations mentioned above is present in one extracted biometric trait, there will be other traits available to the system to use in its decision. Accordingly, it is necessary to determine a method by which the individual modalities are combined. There are three possible levels at which fusion can be performed; feature level, matching score level, and decision level. We are unable to perform fusion at the feature level because of the use of a commercial voice biometric system. Also, the matching scores provide much more information than the output decisions and, consequently, we will perform fusion at the matching score level. After having computed the signature and voice matching scores and before attempting to combine the scores, a normalization technique has to be applied. The signature score is a distance measure in the range $[0, \infty)$, where 0 indicates a perfect match and any non-zero value represents the degree of difference between the two signatures. The Nuance speech SDK produces a score as a similarity measure in the range $(-\infty, \infty)$, where a negative value represents a small similarity between the two voiceprints and a positive value represents a large similarity. The transformation $T_v = e^{-x_v}$ is used to convert the voice score to a distance measure, where x_v is the raw matching score and T_v is the normalized score. After this transformation, both the modalities have a similar range of $[0, \infty)$.

The problem of combining the scores from the voice and signature modalities for a given test sample T with scores (T_v, T_s) can be considered as a two-class classification problem. The sample T can fall into either the impostor (w_i) or genuine (w_g) class. A Bayesian approach would assign T to w_i if

$$P(w_i|T_v, T_s) > P(w_g|T_v, T_s) \tag{1}$$

and w_g otherwise. In the above equation, T_v and T_s are the normalized voice and signature scores, respectively, and $P(w|T_v, T_s)$ denotes the posteriori probability of class w given the voice and signature scores. The strategy used in our system is the simple sum rule described in Jain and Ross [9]. This rule assumes statistical independence among the two modalities and also assumes that the posteriori probabilities computed by the individual classifiers do not deviate much from the prior probabilities [8]. The weighted sum rule assigns a test sample $T = (T_v, T_s)$ to w_i if

$$W_v P(w_i|T_v) + W_s P(w_i|T_s) > W_v P(w_g|T_v) + W_s P(w_g|T_s) \tag{2}$$

and w_g otherwise. In equation (2), W_v and W_s are the weights assigned to the voice and signature scores, respectively. Figure 3 shows the genuine and impostor

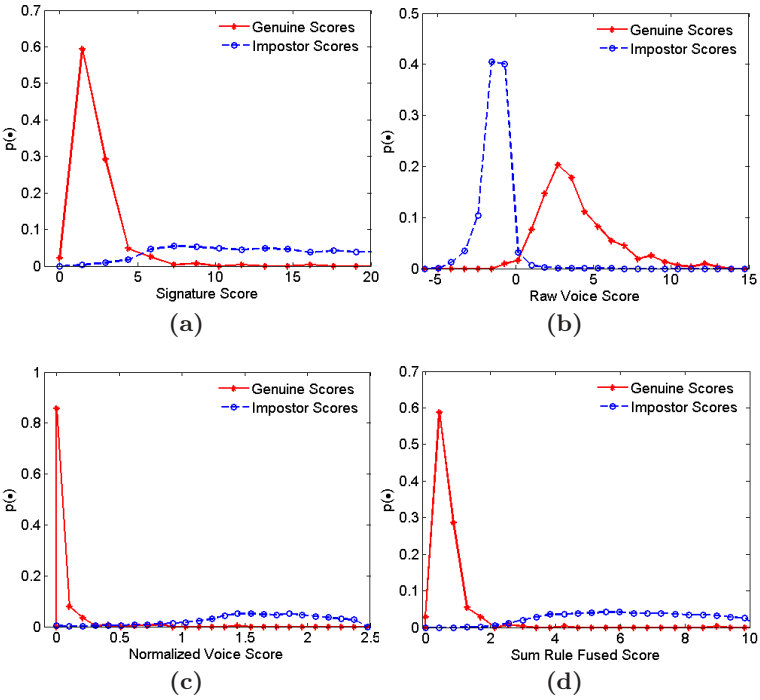


Fig. 3. Distribution of genuine and impostor scores from one trial of cross validation; (a) Signature (distance score), (b) Raw voice (similarity score), (c) Normalized voice (distance score), (d) After sum rule fusion (distance score)

distributions of the signature, raw voice, normalized voice, and fused matching scores using equal weights.

4 Results

4.1 Database

The data used for the evaluation of the authentication system was gathered from 50 individuals, each contributing 10 voiceprints and 10 signatures. The data was collected in a single session from students in various laboratories on our campus with significant ambient noise. Each individual was asked to speak his or her full name and provide a genuine signature. A Toshiba Protege tablet PC was used to perform the data collection for both the voice and signature using the stylus for the signature and the internal microphone for the voice.

4.2 Performance

The database was divided into training and testing sets by using three randomly selected voice and signature samples as the training set and the remaining seven samples as the testing set. The training voice and signature samples were used

for enrollment for each user, creating user templates for each modality. The testing samples are then used to generate authentic scores for each user. Random impostors for a user are generated by using the signature and voice samples from all the other users. The corresponding receiver operator characteristic (ROC) curves are shown in figure 4. After performing ten-fold cross validation, the average equal error rate of voice alone is 1.60% versus 3.62% for signature alone. The variance of the equal error rates of the individual voice and signature systems is 0.05 and 0.31, respectively. The combination of the two modalities using the weighted sum rule (with equal weights) has an average equal error rate of 0.86% and a variance of 0.01.

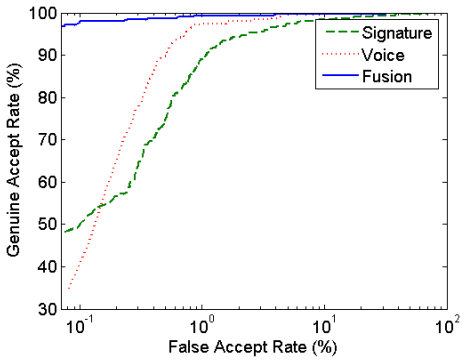


Fig. 4. ROC curves showing the results of the unimodal and multimodal systems from one trial of cross validation. The verical axis is the genuine accept rate and the horizontal axis is the false accept rate, drawn on a logarithmic scale

Figures 5, 6, and 7 show some specific examples of incorporating multiple modalities into the final decision. Figure 5 displays an example of an error in the signature verification algorithm being corrected by fusion. Here, the template and query signatures are very similar and, therefore have a low matching score. However, because the voice verification algorithm found the two voiceprints to be dissimilar, the multimodal system was able to classify the query correctly as an impostor. Figure 6 displays a situation where the query voiceprint contained a significant amount of noise and was incorrectly matched with the template. On the other hand, the signature verification algorithm found the user to be an impostor and this was able to help the system classify the query correctly. Finally, figure 7 displays an example of an error that was unable to be resolved by the mulitmodal system. Both voiceprints are greatly influenced by noise and the verification provides a misleadingly low distance score. The signatures also seem to follow the same pattern and the verification process found them to be similar. Both modalities gave wrong results and, consequently, the fusion system was unable to correctly classify the query as an impostor.

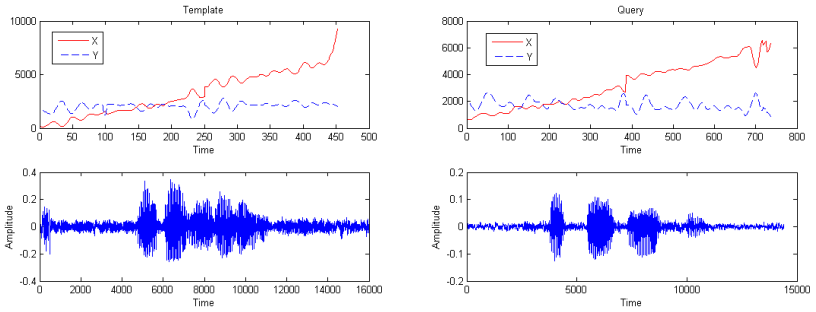


Fig. 5. Signature error resolved by fusion. The graphs show the x and y signals of signature and amplitude of voice samples plotted against time of two different users (a) and (b). The signature signals are the upper plot while the voice waveforms are depicted below. The signature score between (a) and (b) is 0.77, indicating a genuine signature. The normalized voice score between (a) and (b) is 1.5, indicating an impostor voice sample. Fusing the scores together shows the user to be an impostor

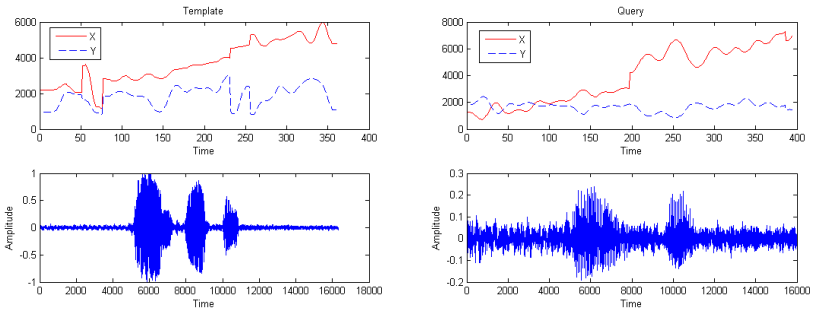
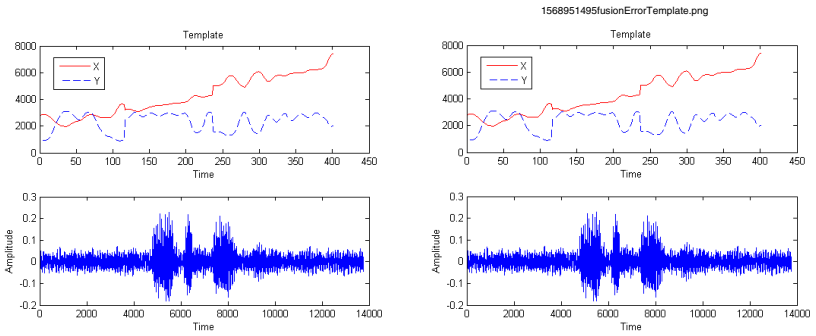


Fig. 6. Voice error resolved by fusion. The graphs show the x and y signals of signature and amplitude of voice samples plotted against time of two different users (a) and (b). The signature signals are the upper plot while the voice waveforms are depicted below. The signature score between (a) and (b) is 1.853, indicating an impostor signature. The normalized voice score between (a) and (b) is 0.02, indicating a genuine voice. Fusing the scores together shows the user to be an impostor

5 Conclusions

We have designed and implemented an authentication system based on the fusion of voice and signature data. This system was motivated by the health care industry and is designed to interact well with both patients and health care professionals. The authentication system will help medical facilities comply with the HIPAA regulations regarding protection and privacy of medical records and accountability issues. The HIPAA regulations require all patient data access to be logged. This is done in order to provide accountability (audit trail); anyone who accesses the patient records is held responsible for what they see and do.



(a) **(b)**

Fig. 7. Unresolved error after fusion. The graphs show the x and y signals of signature and amplitude of voice samples plotted against time of two different users (a) and (b). The signature signals are the upper plot while the voice waveforms are depicted below. The signature score between (a) and (b) is 0.979, indicating a genuine signature. The normalized voice score between (a) and (b) is 0.06, indicating a genuine voice. Fusing the scores together shows the user to be genuine

Accordingly, this system gives a much higher confidence in the access logs because it is very likely that the individual who logged into the system is the same as the enrolled user. To combine the voice and signature modalities, we used fusion at the matching score level and, in particular, used the weighted sum rule. Using both the modalities gives higher accuracy than either individual modality and also makes spoofing of the system a much more difficult task. Thresholds can be adjusted in this system in order to achieve the desired security in this application domain.

Acknowledgments

This work was supported by the MSU CyberSecurity Initiative. We would like to acknowledge the help of Dr. Michael Zaroukian for providing us excellent guidance during the course of this project.

References

1. A^4 Health Systems. A^4 Health Systems Electronic Medical Record Solutions. http://www.a4healthsystems.com/.
2. George J. Annas. *The Rights of Patients*. Southern Illinois University Press, Carbondale, Illinois, 2004.
3. BCBSRI. Blue Cross Blue Shield of Rhode Island. https://www.bcbsri.com.
4. Business Wire. http://www.businesswire.com
5. D'Arcy Guerin Gue. The HIPAA Security Rule (NPRM): Overview. http://www.hipaadvisory.com/regs/securityoverview.htm.
6. HHS. Protecting the Privacy of Patients' Health Information. http://www.hhs.gov/news/facts/privacy.html.

7. A. K. Jain, Friederike D. Griess, and Scott D. Connell. On-line Signature Verification. *Pattern Recognition*, 35(12):2963–2972, December 2002.
8. A. Jain, K. Nandakumar, A. Ross. Score Normalization in Multimodal Biometric Systems. To appear in *Pattern Recognition*, 2005.
9. A. K. Jain and A. Ross. Information Fusion in Biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, September 2003.
10. A. K. Jain and A. Ross. Multibiometric Systems. *Communications of the ACM*, 47(1):34–40, January 2004. Special Issue on Multimodal Interfaces.
11. Nuance. Nuance Corporation http://www.nuance.com.
12. S. Prabhakar, S. Pankanti, and A.K. Jain. Biometric Recognition: Security & Privacy Concerns. *IEEE Security & Privacy Magazine*, 1(2):33–42, March-April 2003.
13. Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of IEEE*, vol. 77, No. 2, 257-286, Feb. 1989
14. University of South Alabama Health System. http://www.southalabama.edu/usahealthsystem/.
15. S. Krawczyk. User Authentication using On-line Signature and Speech. MS Thesis, Michigan State University, Dept. of Computer Science and Engineering (May 2005).
16. D. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and Gerhard Rigoll. SVC2004: First International Signature Verification Competition. *Proceedings of the International Conference on Biometric Authentication (ICBA)* Hong Kong, 15-17 July 2004.

A Novel Approach to Combining Client-Dependent and Confidence Information in Multimodal Biometrics

Norman Poh and Samy Bengio

IDIAP Research Institute, Rue du Simplon 4, CH-1920 Martigny, Switzerland
{norman,bengio}@idiap.ch

Abstract. The issues of fusion with client-dependent and confidence information have been well studied separately in biometric authentication. In this study, we propose to take advantage of both sources of information in a discriminative framework. Initially, each source of information is processed on a per expert basis (plus on a per client basis for the first information and on a per example basis for the second information). Then, both sources of information are combined using a second-level classifier, across different experts. Although the formulation of such two-step solution is not new, the novelty lies in the way the sources of prior knowledge are incorporated prior to fusion using the second-level classifier. Because these two sources of information are of very different nature, one often needs to devise special algorithms to combine both information sources. Our framework that we call "Prior Knowledge Incorporation" has the advantage of using the standard machine learning algorithms. Based on $10 \times 32 = 320$ intramodal and multimodal fusion experiments carried out on the publicly available XM2VTS score-level fusion benchmark database, it is found that the generalisation performance of combining both information sources improves over using either or none of them, thus achieving a new state-of-the-art performance on this database.

1 Introduction

Previous studies have shown that combining several biometric authentication systems is a potential way to improve the overall system accuracy [1]. It has also been shown that fusion with client-dependent and confidence information can *further* improve the system performance. Studies using *client-dependent information* include client-dependent threshold [2], model-dependent score normalisation [3] or different weighing of expert opinions using linear [4] or non-linear combination [5] on a per client model basis. Some of the existing approaches to incorporate the *confidence or quality information* are a multivariate polynomial regression function [6], a statistical model (that reconciles expert opinions) [7] and a modified Support Vector Machine algorithm [8]. Specific to speaker authentication, in [9], the first formant of speech was used as an indicator of quality to weigh the Log-Likelihood Ratio (LLR) of each speech frame. Thus, instead of taking the average LLR as commonly done, a weighted average LLR was used. These studies have shown that incorporation of client-dependent and confidence information are important means to improve multimodal biometric systems.

In this study, we would like to verify whether fusion using both of these sources of information is more beneficial than using either one or none at all. To the best of

T. Kanade, A. Jain, and N.K. Ratha (Eds.): AVBPA 2005, LNCS 3546, pp. 1120–1129, 2005.
© Springer-Verlag Berlin Heidelberg 2005

our knowledge, this issue has not been examined before. This is perhaps because these two sources of information are very different, and strategies employed to integrate one source of information is completely different from or incompatible with the other. We propose a novel way to fuse these two sources of information in two steps: first incorporate the prior knowledge on a per expert basis and then combine them using a second classifier. The idea of using a second classifier is not new. This strategy is called post-classification in [10]. However, deriving ways to incorporate the prior knowledge into the scores, on a per expert basis, prior to fusion is new. This framework is called "Prior Knowledge Incorporation" (PKI). It should be noted that the prior knowledge incorporated scores, on their own, may not necessarily be very useful if not further combined with other scores. The advantage of this technique is that, due to PKI scores, (the first step), information sources can be combined independently. In terms of implementation, this means modular integration is possible. Secondly, the second-level classifier can be implemented using standard off-the-shelf machine-learning algorithms, thus eliminating the need to create a specific fusion algorithm for this purpose. In principle, any sources of prior knowledge can be combined this way. In practice, the amount of prior knowledge possibly employed is limited by the information given by the baseline expert systems.

In order to verify this hypothesis, three sets of fusion control experiments were carried out, i.e., fusion using the original expert scores, fusion using client-dependent normalised scores and fusion using confidence. These baseline experiments are then compared to fusion using all the available information sources. Based on 32 fusion data sets taken from the publicly available XM2VTS score fusion benchmark database [11], it is concluded that fusion with both sources of information is more beneficial than using either one or none of them.

This paper is organised as follows: Sections 2 and 3 discuss briefly how the client-dependent information and confidence information can be computed, on a per expert basis. Section 4 discusses how these seemingly different sources of information can be fused together using the PKI framework. The database and results are presented in Sections 5 and 6, respectively. They are followed by conclusions in Section 7.

2 Deriving Client-Dependent Information

There exists a vast literature in this direction. A survey can be found in [12, Sec. 2]. There are two families of approaches, namely, score normalisation and threshold normalisation. The former aims at normalising the score such that a global decision threshold can be found easily. The latter manipulates the decision threshold directly. It has been shown that [12] both families are dual forms of each other. The disadvantage of the latter category is that it is dependent on a specific cost of false acceptance and false rejection while the former does not have to be. Hence, client-dependent score normalisation methods are considered here.

Examples of existing methods are Z-, D- (for Distance), T- (for Test) and more recently, F-Norms (for F-ratio). In the terms used in [3, 13], Z-Norm [13] is impostor-centric (i.e, normalisation is carried out with respect to the impostor distributions calculated "offline" by using additional data), T-Norm [13] is also impostor-centric (but

with respect to a given utterance calculated "online" by using additional cohort impostor models). D-Norm [14] is neither client- nor impostor-centric; it is specific to the Gaussian Mixture Model (GMM) architecture and is based on Kullback-Leibler distance between two GMM models. In [2], a client-centric version of Z-Norm was proposed. However, this technique requires as many as five client accesses. Due to user-friendliness aspect, one often does not have many client-specific biometric samples. To overcome this problem, F-Norm was proposed [12]. It is client-impostor centric. Based on the experiments reported, as few as two client scores are needed to perform this normalisation. It was shown that F-Norm is superior over Z-Norm because F-Norm uses the client-specific impostor information in addition to the client-specific information.

In this study, as an extension of [12], F-Norm is used. Suppose that the score of a system is y. It indicates how likely that a given biometric sample belongs to a client. Let $\mu^k(j)$ be the mean score of client with the unique identity j given that the true class-label $k = \{C, I\}$ (either a client or an impostor) is known (from a development set). Let the (class-dependent but) client-independent mean be μ^k, for $k = \{C, I\}$. The resultant F-ratio transformed normalisation is:

$$y^F = A(j)(y - B(j)), \tag{1}$$

where,

$$A(j) = \frac{2a}{\beta(\mu^C(j) - \mu^I(j)) + (1-\beta)(\mu^C - \mu^I)}, \tag{2}$$

and

$$B(j) = \gamma\mu^I(j) + (1-\gamma)\mu^I \tag{3}$$

The terms $A(j)$ and $B(j)$ are associated to client j (client-dependent) and are derived from F-ratio. They are each controlled by the parameters $\beta \in [0, 1]$ and $\gamma \in [0, 1]$ on a per fusion experiment basis. The term $2a$ determines the "desired" distance between the client-specific mean and the client-specific impostor mean. a is a constant and is fixed to 1. β and γ adjust between the client-dependent and client-independent information. When $\beta = 0$ and $\gamma = 0$, it can be shown mathematically that F-ratio normalisation is equivalent to no normalisation at all. In biometric authentication, one often has abundant client-specific (simulated) impostor information. Preliminary experiments in [12] show that $\gamma = 1$ is always optimal. The experimental results confirm that due to abundant client-specific impostor information, the shift in $B(j)$ can always be estimated reliably. As a consequence, the only parameter needs to be optimised, on a per experiment and per expert basis, is the β parameter. It can be optimised using different approaches, among which the direct approach is to use the line search procedure [15, Sec. 7.2].

3 Deriving Confidence Information

It has been shown in [16] that confidence can be derived from a "margin". The margin can be defined from False Acceptance (FA) Rate (FAR) and False Rejection (FR) Rate (FRR) with respect to a threshold Δ. FAR and FRR are defined as follows:

$$\text{FAR}(\Delta) = \frac{\text{number of FAs}(\Delta)}{\text{number of impostor accesses}}, \tag{4}$$

$$\text{FRR}(\Delta) = \frac{\text{number of FRs}(\Delta)}{\text{number of client accesses}} \ . \tag{5}$$

Replacing Δ by the associated expert score y, the margin of the score y is defined as:

$$q = |\text{FAR}(y) - \text{FRR}(y)| \tag{6}$$

Hence, when incorporated into an existing discriminant function, q modifies the discriminant function *dynamically*, i.e., *a per example basis*. Suppose that y_i is the score of expert $i = 1, \ldots, N$. Linear combination of $\{y_i, q_i y_i\}$ from different expert systems, with weight $w_{1,i}$ associated to y_i and $w_{2,i}$ associated to $q_i y_i$, is equivalent to computing $y_i \times (w_{1,i} + q_i w_{2,i})$, for all i [16]. Note that from the term $(w_{1,i} + q_i w_{2,i})$, it is obvious that q_i has a direct influence on the gradient of the resultant discriminative function on a *per example basis*. Hence, $\{y_i, q_i y_i\}$, can be seen as a form of Prior Knowledge Incorporation (PKI). Using equal weight in linear combination, in [16], it was shown that fusion with $\{q_i y_i | \forall_i\}$ has a better generalisation performance than fusion without the margin information (the classical way), i.e., $\{y_i | \forall_i\}$. Furthermore, fusion with $\{y_i, q_i y_i | \forall_i\}$ *consistently* outperforms $\{q_i y_i | \forall_i\}$, even though the generalisation performance is not always significant based on the HTER significance test [17].

4 Combing Both Sources of Information: A Prior Knowledge Incorporation (PKI) Framework

In the previous sections, the client-dependent and confidence information are employed on a per expert basis, independently of the other expert scores. The concept of PKI was introduced when discussing how confidence (based on margin) can be combined. In this section, we extend this concept to incorporate the client-dependent information as well, i.e., using $\{y_i, q_i y_i, y_i^F | \forall_i\}$. In principle, we could combine any other sources of information or prior knowledge this way. The only limit is the amount of prior knowledge captured by the available data (scores in this case).

Suppose that a linear combination is used to fuse $\{y_i, q_i y_i, y_i^F | \forall_i\}$. Let $w_{1,i}$, $w_{2,i}$ and $w_{3,i}$ be weights associated to y_i, $q_i y_i$ and y_i^F, respectively, for all i. Let the bias term be $-\Delta$, where Δ is the final decision threshold. Note that in this study, a separate training procedure of the Δ parameter is employed to minimise Weighted Error Rate (WER) *on the development set*. WER is defined as:

$$\text{WER}_\alpha(\Delta) = \alpha \text{FAR}(\Delta) + (1 - \alpha)\text{FRR}(\Delta), \tag{7}$$

where $\alpha \in [0, 1]$ balances between FAR and FRR. This procedure requires the computation of fused scores on both the development and evaluation sets. In this way, during testing , based on a specified WER, the obtained threshold from the development set can be applied to the evaluation set. A separate threshold estimation procedure is necessary because algorithms that optimise the parameters of the fusion classifiers (weights in the linear combination case) *do not* necessarily optimise WER. For instance, SVM maximises the margin; Fisher discriminant maximises the Fisher-ratio criterion, etc.

The fused score can be written as:

$$\begin{aligned} y_{COM} &= \sum_i \left[y_i w_{1,i} + q_i y_i w_{2,i} + y_i^F w_{3,i} \right] - \Delta \\ &= \sum_i \left[y_i w_{1,i} + q_i y_i w_{2,i} + B(j)(y_i - A(j)) w_{3,i} \right] - \Delta \\ &= \sum_i \left[y_i \left(\underbrace{w_{1,i}} + \underbrace{q_i w_{2,i}} + \underbrace{B(j) w_{3,i}} \right) \right] - \sum_i \left[\underbrace{B(j) A(j) w_{3,i}} \right] - \underbrace{\Delta}, \end{aligned} \quad (8)$$

where Eqn. (1) was used to replace the term y_i^F. The first underbraced term is the *global weight* on a *per expert basis*; the second is the weight contribution due to the confidence information on a *per example basis*; and the third is the weight contribution due to the client-dependent information source on a *per client basis*. These three weights are *linearly* combined to weight the score y_i. Then the fourth underbraced term introduces the client-dependent shift on a *per expert and per client basis*. Finally, the last underbraced term introduces the *global shift* to the final discriminative function. This term (Δ) is optimised by minimising WER for a given α value. From fusion point of view, the first three underbraced terms introduce tilt and while the last two underbraced term introduces shift to the decision hyperplane.

Although the PKI scores are simple to obtain, their linear combination can be a very complex function as shown here. It should be noted that even though non-linear combination can also be used (using the SVM algorithm with non-linear kernels , polynomial expansion of the terms $\{y_i, q_i y_i, y_i^F | \forall_i\}$, etc), simple linear solution is preferred to avoid overfitting. Furthermore, most of the non-linear part of the problem should have been solved by the base experts, thus eliminating the need for a complex second-level classifier.

5 Database and Evaluation

The publicly available[1] XM2VTS benchmark database for score-level fusion [11] is used. There are altogether 32 fusion data sets and each data set contains a fusion task of two experts. These fusion tasks contain multimodal and intramodal fusion based on face and speaker authentication tasks. For each data set, there are two sets of scores, from the *development* and the *evaluation* sets. The development set is used *uniquely* to train the fusion classifier parameters, including the threshold (bias) parameter, whereas the evaluation set is used uniquely to evaluate the generalisation performance. They are in accordance to the two originally defined Lausanne Protocols [18]. The 32 fusion experiments have 400 (client accesses) $\times$ 32 (data sets)= 12,800 client accesses and 111,800 (impostor accesses) $\times$ 32 (data sets) = 3,577,600 impostor accesses.

The most commonly used performance visualising tool in the literature is the Decision Error Trade-off (DET) curve [19]. It has been pointed out [20] that two DET curves resulting from two systems are not comparable because such comparison does not take into account how the thresholds are selected. It was argued [20] that such threshold

[1] Accessible at http://www.idiap.ch/∼norman/fusion

should be chosen *a priori* as well, based on a given criterion. This is because when a biometric system is operational, the threshold parameter has to be fixed *a priori*. As a result, the Expected Performance Curve (EPC) [20] was proposed. This curve is constructed as follows: for various values of α in Eqn. (7) between 0 and 1, select the optimal threshold Δ on a development (training) set, apply it on the evaluation (test) set and compute the HTER on the evaluation set. This HTER is then plotted with respect to α. The EPC curve can be interpreted similarly to the DET curve, i.e., the lower the curve, the better the generalisation performance. In this study, the *pooled* version of EPC is used to visualise the performance. The idea is to plot a single EPC curve instead of 32 EPC curves for each of the 32 fusion experiments. This is done by calculating the *global* false acceptance and false rejection errors over the 32 experiments for *each* of the α values. The pooled EPC curve and its implementation can be found in [11].

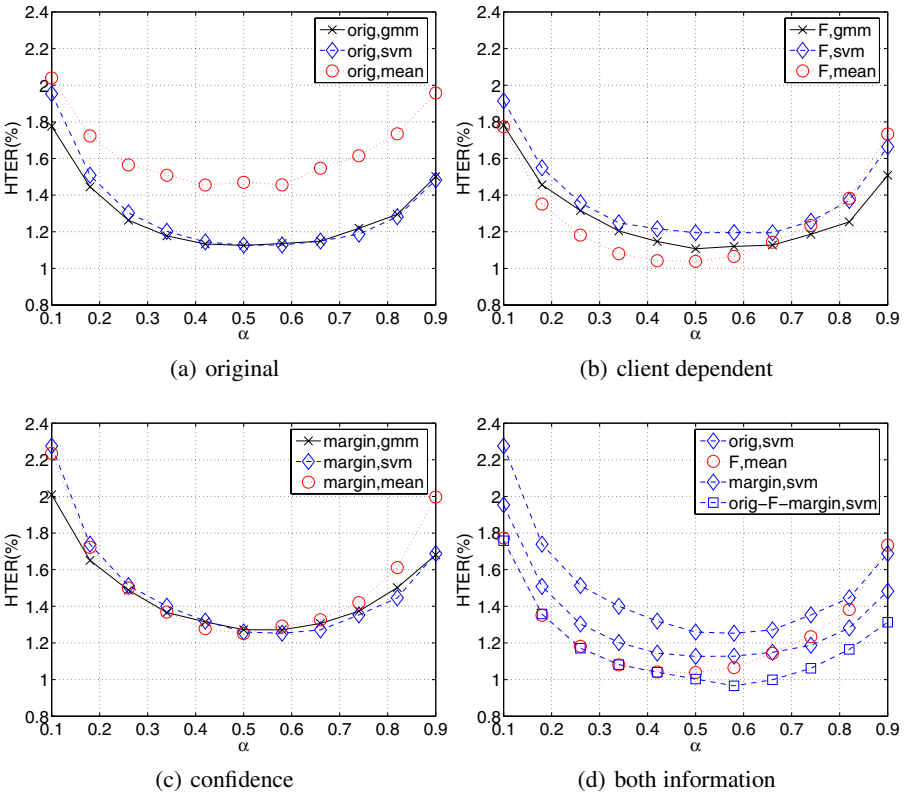
6 Experimental Results

The client-*dependent* setting is used to derive F-Norm transformed scores. On the other hand, the client-*independent* setting is used to derive the margin scores. Three sets of control experiments are performed, namely with original scores $\{y_i|\forall_i\}$, F-Norm transformed scores $\{y_i^F|\forall_i\}$ and margin-derived confidence scores $\{y_i q_i|\forall_i\}$. For each set of experiments, three types of fusion classifiers are used, namely, a Gaussian Mixture Model (GMM), a Support Vector Machine (SVM) with a linear kernel and the mean operator. Both GMM and SVM employed are using standard algorithms, without any particular modification. The hyper-parameters are selected automatically via cross-validation. Figures 1(a)–(c) show the generalisation performance of these three sets of control experiments. Each curve is a pooled EPC curve over 32 fusion multimodal and intramodal datasets. Figure 2 complements Figure 1 by showing the corresponding ROC curves.

To compare these three control experiments with the ones fusing all sources of information, i.e., $\{y_i, y_i q_i, y_i^F|\forall_i\}$, we plotted the best of each pooled EPC curves in (a)–(c) on (d). As can be seen in (d), fusion with all sources of information using SVM has the best generalisation performance, bringing a new state-of-the-art overall performance on this benchmark data set. Considering significant performance improvement with respect to the 3×3 sets of control experiments, for large range of α values (> 0.6 for the best pooled EPC curve of the 9 control experiments over 32 fusion data sets), one can conclude that fusion using client dependent and confidence information sources via PKI is a feasible approach.

7 Conclusions

In this study, we proposed to fuse two seemingly different sources of information using the Prior Knowledge Incorporation (PKI) framework. These sources of information are client-dependent and confidence information. Although fusion with both sources of information has been studied separately in biometric authentication, to the best of our knowledge, fusing both information sources has not been well investigated before. Because these information sources are of different nature, intuitively, a *new* combination

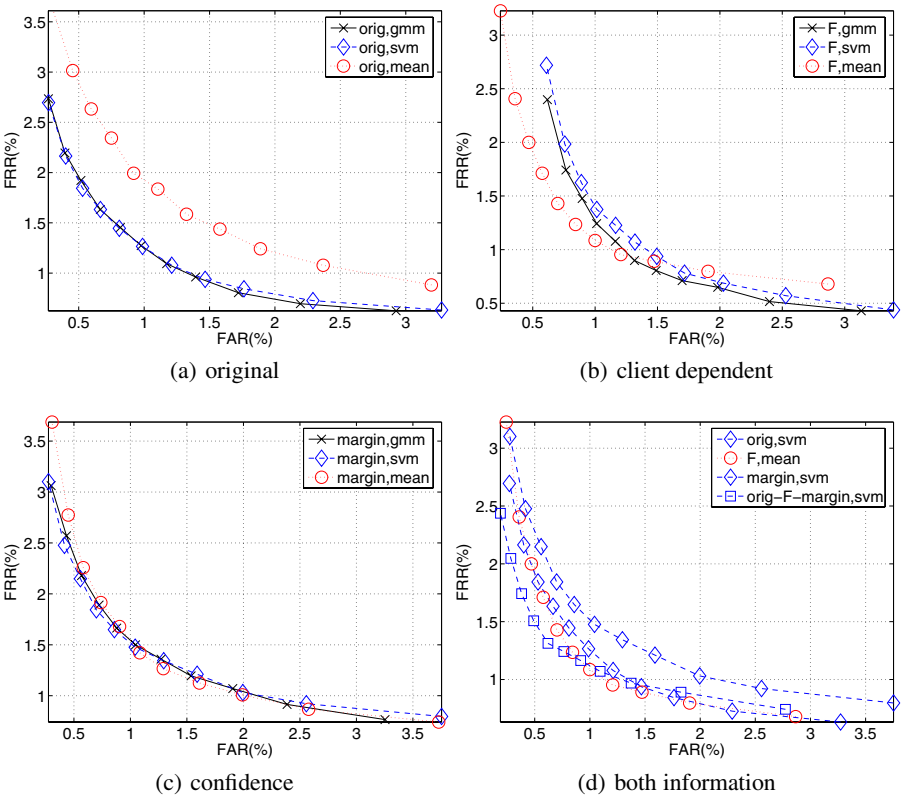


(a) original (b) client dependent

(c) confidence (d) both information

Fig. 1. Pooled EPC curves from 32 XM2VTS benchmark fusion data sets of three baseline experiments (a)–(c) and fusion with all information sources (d). (a) is fusion with the original scores, $\{y_i|\forall_i\}$, (b) is fusion with F-ratio transformed scores, $\{y_i^F|\forall_i\}$,and (c) is fusion with margin-derived confidence, $\{y_i q_i|\forall_i\}$, each using a GMM, an SVM with linear kernel and the mean operator. The best three pooled EPC curves in (a)–(c) are plotted in (d) (the top three in the legend), together with fusion with all sources of information, i.e., $\{y_i, y_i q_i, y_i^F|\forall_i\}$ using an SVM with linear kernel, denoted as "orig-F-margin,SVM". The pooled EPC of this curve is compared to the "best overall fusion" (lowest HTER in the EPC curve across different α) in each of (a)–(c). "orig-F-margin,SVM" is better than "F-mean" for $\alpha > 0.6$ according to the HTER significance test at 90% of confidence. Below $\alpha = 0.6$, both EPC curves are not *significantly different*

algorithm would be necessary. However, using the proposed PKI framework, we show that these information sources can be combined at the score level by a linear transformation, for each source of prior knowledge. The advantage is modularity: prior knowledge can be incorporated on a per expert basis (the first step) and the resultant PKI scores can be fused by a second-level classifier using standard machine learning algorihtms (the second step). Thus, this eliminates the need to devise specific fusion algorithms for this purpose. Based on the experiments carried out on 32 intramodal and multimodal fusion data sets taken from the publicly available XM2VTS benchmark database, over 10 fusion classifiers (3 fusion baselines on the original scores; 3 with client-dependent fusion baselines; 3 with margin-enhanced confidence baselines; and a final fusion with all in-



(a) original (b) client dependent

(c) confidence (d) both information

Fig. 2. Pooled ROC curves from 32 XM2VTS benchmark fusion data sets of three baseline experiments (a)–(c) and fusion with all information sources (d). (a) is fusion with the original scores, $\{y_i|\forall_i\}$, (b) is fusion with F-ratio transformed scores, $\{y_i^F|\forall_i\}$,and (c) is fusion with margin-derived confidence, $\{y_i q_i|\forall_i\}$, each using a GMM, an SVM with linear kernel and the mean operator. The "best" three pooled ROC curves (i.e., the EPC curve with the *lowest* HTER value across different α values) in (a)–(c) are plotted in (d), together with the one that fuses all sources of information, i.e., $\{y_i, y_i q_i, y_i^F|\forall_i\}$ using an SVM with linear kernel, denoted as "orig-F-margin,SVM". This figure complements Figure 1. As confirmed by the HTER significance test, for FRR above 1.2%, "orig-F-margin,SVM" is significantly different (and better) than "F-mean" but below 1.2%, their difference is *insignificant*. This phenomenon is due to few client accesses as compared to impostor accesses. As a result, low FRR values cannot be interpreted reliably compared to low FAR values

formation sources), fusion with both information sources using the PKI framework has the best generalisation performance and its performance is significant over large values of operating (false acceptance/false rejection) costs as compared to the most competing technique, i.e., fusion with client-dependent information.

Acknowledgment

This work was supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the

Swiss Federal Office for Education and Science (OFES) and the Swiss NSF through the NCCR on IM2. This publication only reflects the authors' view.

References

1. J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, "Combining Evidence in Personal Identity Verification Systems," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, 1997.
2. J.R. Saeta and J. Hernando, "On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 215–218.
3. J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Target Dependent Score Normalisation Techniques and Their Application to Signature Verification," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 498–504.
4. A. Jain and A. Ross, "Learning User-Specific Parameters in Multibiometric System," in *Proc. Int'l Conf. of Image Processing (ICIP 2002)*, New York, 2002, pp. 57–70.
5. A. Kumar and D. Zhang, "Integrating Palmprint with Face for User Authentication," in *Workshop on Multimodal User Authentication (MMUA 2003)*, Santa Barbara, 2003, pp. 107–112.
6. K-A. Toh, W-Y. Yau, E. Lim, L. Chen, and C-H. Ng., "Fusion of Auxiliary Information for Multimodal Biometric Authentication," in *Springer LNCS-3072, Int'l Conf. on Biometric Authentication (ICBA)*, Hong Kong, 2004, pp. 678–685.
7. J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal Biometric Authentication using Quality Signals in Mobile Communnications," in *12th Int'l Conf. on Image Analysis and Processing*, Mantova, 2003, pp. 2–11.
8. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, "Kernel-Based Multimodal Biometric Verification Using Quality Signals," in *Defense and Security Symposium, Workshop on Biometric Technology for Human Identification, Proc. of SPIE*, 2004, vol. 5404, pp. 544–554.
9. D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "On the Use of Quality Measures for Text Independent Speaker Recognition," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 105–110.
10. C. Sanderson and K. K. Paliwal, "Information Fusion and Person Verification using Speech and Face Information," IDIAP-RR 22, IDIAP, 2002.
11. N. Poh and S. Bengio, "Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication," Research Report 04-44, IDIAP, Martigny, Switzerland, 2004, Accepted for publication in *AVBPA 2005*.
12. N. Poh and S. Bengio, "Improving Single Modal and Multimodal Biometric Authentication Using F-ratio Client Dependent Normalisation," Research Report 04-52, IDIAP, Martigny, Switzerland, 2004.
13. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independant Speaker Verification Systems," *Digital Signal Processing (DSP) Journal*, vol. 10, pp. 42–54, 2000.
14. M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo Method For Score Normalization in Automatic Speaker Verification Using Kullback-Leibler Distances," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, 2002, vol. 1, pp. 689–692.
15. C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1999.
16. N. Poh and S. Bengio, "Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks," Research Report 04-63, IDIAP, Martigny, Switzerland, 2004, Accepted for publication in *AVBPA 2005*.

17. S. Bengio and J. Mariéthoz, "A Statistical Significance Test for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 237–244.
18. J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of Face Verification Results on the XM2VTS Database," in *Proc. 15th Int'l Conf. Pattern Recognition*, Barcelona, 2000, vol. 4, pp. 858–863.
19. A. Martin, G. Doddington, T. Kamm, M. Ordowsk, and M. Przybocki, "The DET Curve in Assessment of Detection Task Performance," in *Proc. Eurospeech'97*, Rhodes, 1997, pp. 1895–1898.
20. S. Bengio and J. Mariéthoz, "The Expected Performance Curve: a New Assessment Measure for Person Authentication," in *The Speaker and Language Recognition Workshop (Odyssey)*, Toledo, 2004, pp. 279–284.

Author Index

Abhyankar, Aditya 301
Adler, Andy 860, 1100
Ahn, Dosung 1071
Akarun, Lale 1019
Akkermans, Anton H.M. 436
Alba-Castro, José Luis 51
Allano, Lorène 23, 494
Aradhye, Hrishikesh 879

Bazen, Asker M. 436
Beattie, Michael 406
Bengio, Samy 474, 1059, 1120
Bhanu, Bir 355, 416, 533, 748
Bicego, Manuele 329
Bigun, Josef 171
Bistarelli, Stefano 279, 464
Bogoni, Luca 1
Bolles, Robert 879
Boult, Terrance 850
Bourlai, Thirimachos 289
Boutin, Mireille 693
Bowyer, Kevin W. 41, 503
Bunke, Horst 191
Butakoff, Costantine 365

Cantó, Enrique 683
Canyellas, Nicolau 683
Carter, John N. 386, 597
Cha, Sung-Hyuk 823
Chai, Xiujuan 956
Champod, Christophe 1080
Chen, Hui 355, 748
Chen, Jian-Sheng 376
Chen, Jie 208
Chen, Xilin 937, 956, 966
Chen, Xinjian 151, 665
Chen, Yi 160
Chiang, Chung-Shi 14
Chikkerur, Sharat 841, 977
Choi, Hee-seung 260
Choi, Kyoungtaek 260
Choi, Tae Young 656
Choi, Wooyong 1071
Chu, Nancy 607
Cruz, Santiago 365

Cukic, Bojan 484

Dass, Sarat C. 160, 1049
Daum, Henning 900
Demirkır, Cem 339
Deng, Huimin 270
Dorizzi, Bernadette 23, 494
Drygajlo, Andrzej 426

Ejima, Toshiaki 767
Elgammal, Ahmed 395
El Hannani, Asmaa 580
Enokida, Shuichi 767

Fan, Wei 122
Fancourt, Craig 1
Fang, Yuchun 637
Fierrez-Aguilar, Julian 523, 1080
Fischler, Martin 879
Flynn, Patrick J. 41, 544, 627
Forte, Giuseppe 683
Fox, Niall A. 777, 787
Frangi, Alejandro 365
Frassi, Stefano 464
Fukui, Kazuhiro 71
Fukui, Yutaka 758
Furui, Sadaoki 588

Gao, Wen 208, 937, 956, 966
Garcia-Salicetti, Sonia 23, 494
Geman, Donald 637
Gibbons, Michael 823
Gil, Younhee 1071
Gökberk, Berk 1019
González-Jiménez, Daniel 51
Gonzalez-Rodriguez, Joaquin 1080
Govindaraju, Venu 833, 841, 977
Grosso, Enrico 329
Guan, E 1010
Gui, Zhenghui 1029
Guo, Yanlin 1

Han, Ju 416, 533
Han, Youngchan 710
Hanna, Keith 1
Hao, Pengwei 201

Harthattu, Ashwath 607
Hernando, Javier 572
Hi, Yuliang 730
Hild, Michael 239
Hongo, Yasunori 455
Huang, Ping S. 14
Huo, Qiang 270
Hurley, David J. 386

Ishimura, Toshiyuki 767
Itoh, Yoshio 758
Iwano, Koji 588

Jain, Anil K. 160, 310, 1049, 1110
Jain, Uday 1
Jang, Jain 141
Jee, Hyungkeun 1071
Jiang, Li-Jun 320
Jung, Ho-Choul 987

Kamel, Mohamed 447
Kang, Byung Jun 31
Kevenaar, Tom A.M. 436
Kim, Dong-Hun 674
Kim, Hakil 710
Kim, Jaihie 141, 260
Kim, Woong-Sik 702
Kim, Yong-Guk 184
Kittler, Josef 289, 617
Kong, Adams 447
Krawczyk, Stephen 1110
Krichen, Emine 23
Kryszczuk, Krzysztof 426
Kumar, Ajay 813
Kumar, B.V.K. Vijaya 61, 406, 607
Kuzui, Ryo 239

Lee, Chan-Su 395
Lee, Joon-Jae 693
Lee, Kangrok 141
Lee, Kyunghee 219, 1071
Lee, Sang-Woong 112, 987
Lee, Sanghoon 141
Lee, Seong-Whan 102, 112, 987
Lee, Sung-Oh 184
Lee, Yillbyung 513
Lee, Yongjin 219, 1071
Lee, Yonguk 909
Lei, Zhenchun 797
Li, Jiangwei 229
Li, Liang 730
Li, Ming 929
Liang, Ji-Ren 14
Lie, Agus Santoso 767
Liu, Chengjun 1039
Liu, Li 563
Liu, Tong 201
López, Mariano 683
Lopez-Peñalba, Jaime 523
Lopresti, Daniel P. 1090
Lucey, Simon 406

Ma, Yan 484
Maeda, Takuji 945
Maltoni, Davide 523
Mammone, Richard 607
Mansukhani, Praveer 833
Matsumoto, Takashi 455
Matsushita, Masahito 945
Mellakh, Mohamed Anouar 494
Messer, Kieron 289, 617
Mhatre, Amit 841
Min, Jaesik 41
Mitra, Sinjini 61
Miyajima, Chiyomi 739
Miyazaki, Taro 588
Moon, Song-Hyang 112
Moon, Yiu-Sang 376
Mueller, Klaus 1010
Muramatsu, Daigo 455
Myers, Gregory 879

Na, Sangsin 656
Nakanishi, Isao 758
Nandakumar, Karthik 1049
Nanni, Loris 523
Nayak, Sankalp 977
Neuhaus, Michel 191
Nilsson, Kenneth 171
Nishiyama, Masashi 71
Nixon, Mark S. 386, 597
Novikov, Sergey 250

O'Mullane, Brian A. 777, 787
Ordás, Sebastián 365
Ortega-Garcia, Javier 523, 1080
Ozawa, Koji 739

Pamudurthy, Satprem 1010
Pan, Feng 320
Pan, Sungbum 219, 1071

Pankanti, Sharath 310
Papatheodorou, Theodoros 997
Park, Chul-Hyun 693
Park, Gwi-Tae 184
Park, Jeong-Seon 102
Park, Kang Ryoung 31, 141
Patel, Ankur 607
Petrovska-Delacrétaz, Dijana 580
Phillips, P. Jonathon 869
Podilchuk, Christine 607
Poh, Norman 474, 1059, 1120

Qing, Laiyun 956

Rafailovich, Miriam 1010
Raim, Jarret D. 1090
Ramos-Castro, Daniel 1080
Ranganath, Surendra 320
Reilly, Richard B. 777, 787
Reisman, James 720
Riopka, Terry 850
Ross, Arun 720
Rueckert, Daniel 997
Ruifrok, Arnout 891
Ryu, Choonwoo 710

Saeta, Javier R. 572
Sakaguchi, Shohei 767
Sakamoto, Hiroyuki 758
Salah, Albert Ali 1019
Samaras, Dimitris 91
Sankur, Bülent 339
Santini, Francesco 279
Sasakawa, Koichi 945
Savvides, Marios 61, 607
Scheenstra, Alize 891
Schrijen, Geert-Jan 436
Schuckers, Michael E. 860
Schuckers, Stephanie 301
Shan, Shiguang 937, 956
Shih, Peichung 1039
Shimomoto, Ryo 767
Shin, Hyungchul 909
Short, James 617
Singh, Harshinder 484
Smith, Mark J.T. 693
Sohn, Kwanghoon 909
Son, Byungjun 513
Song, Hwanjong 909
Song, Wei 320
Sridharan, Karthik 977
Su, Qi 151
Sukno, Federico 365

Takahashi, Naomi 1
Takeda, Kazuya 739
Tan, Tieniu 122, 229
Tang, Xiaofang 929
Tappert, Charles 823
Tewes, Andreas 81
Tian, Jie 151, 665, 730
Tistarelli, Massimo 329
Toh, Kar-Ann 919
Tonguz, Ozan K. 406
Tuyls, Pim 436

Uludag, Umut 310, 720
Ushmaev, Oleg 250

Vaccarelli, Anna 279, 464
Veldhuis, Raymond N.J. 436
Veltkamp, Remco C. 891
Venkateswarlu, Ronda 919
Veres, Galina V. 597
von der Malsburg, Christoph 81

Wada, Tomohito 767
Wakita, Toshihiro 739
Wang, Haoshu 627
Wang, Jian-Gang 919
Wang, Kuanquan 346, 555
Wang, Rong 355, 748
Wang, Ruiping 208
Wang, Sen 91
Wang, Yanrong 647
Wang, Yunhong 122, 229
Wildes, Richard 1
Woodard, Damon L. 544
Wu, Hong-tao 131
Wu, Shi-Qian 320
Wu, Xiangqian 555
Wu, Zhaohui 797, 804
Würtz, Rolf P. 81

Xie, Chunyan 607
Xie, Shou-Lie 320

Yagi, Yasushi 945
Yamaguchi, Osamu 71
Yan, Ping 503
Yan, Shengye 208

Yang, Jie 131
Yang, Pu 804
Yang, Ukil 909
Yang, Xin 151, 665, 730
Yang, Xiukun 647
Yang, Yingchun 797, 804
Yau, Wei-Yun 320
Yeung, Hoi-Wo 376
Yin, Yilong 647
Yoo, Weon-Hee 702
Yoon, Sungsoo 823
Yu, Kyung Deok 656
Yuan, Baozong 929

Zhang, Baochang 966
Zhang, Chao 201, 1029
Zhang, David 346, 447, 555, 563, 813
Zhang, Hongming 937
Zhang, Lei 91
Zhang, Wenchao 937
Zhang, Yong-liang 131
Zhou, Xiaoli 533
Zuo, Wangmeng 346