Lecture Notes in Computer Science 4577

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
University of California, Los Angeles, CA, 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

Nicu Sebe Yuncai Liu
Yueting Zhuang Thomas S. Huang (Eds.)

Multimedia Content Analysis and Mining

International Workshop, MCAM 2007
Weihai, China, June 30-July 1, 2007
Proceedings

Springer

Volume Editors

Nicu Sebe
University of Amsterdam
Faculty of Science, The Netherlands
E-mail: nicu@science.uva.nl

Yuncai Liu
Shandong University, China
E-mail: whomliu@sjtu.edu.cn

Yueting Zhuang
Zhejiang University, China
E-mail: yzhuang@cs.zju.edu.cn

Thomas S. Huang
University of Illinois at Urbana-Champaign, USA
E-mail: huang@ifp.uiuc.edu

Library of Congress Control Number: 2007929552

CR Subject Classification (1998): H.5.1, H.3, H.5, C.2, H.4, I.2-4, E.3, K.6

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-73416-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73416-1 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

springer.com

© Springer-Verlag Berlin Heidelberg 2007
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12085408 06/3180 5 4 3 2 1 0

Preface

Welcome to the International Workshop on Multimedia Content Analysis and Mining, MCAM 2007. Our workshop gives a snapshot of the current worldwide research in multimedia analysis. Through recent advances in computing, networking, and data storage, multimedia will create new interesting technical possibilities in a wide range of fields, such as entertainment, commerce, science, medicine, and public safety. To benefit from this potential, developers need reliable techniques for the analysis, search, and management of multimedia data, as well as distributed system architectures in which these techniques can be embedded to effectively help the users.

Recently, there have been many workshops and meetings dedicated to Multimedia, but most of them followed the standard pattern consisting mainly of oral and poster presentations. We all feel that there is a need for a meeting that really is a workshop, i.e., one that provides a lot of opportunities for discussions and dissemination. Consequently, the format of the workshop consists of 4 panels on hot topics in multimedia followed by discussions and poster presentations. The topics of the panels are: Multimedia Analysis and Applications, Multimedia Search and Mining, P2P Streaming, and Security. This volume contains a number of invited contributions from experts in the area as well as the selected regular contributions.

This year 139 submissions from 13 countries were submitted and 46 were accepted for presentation at the conference after being reviewed by the Program Committee members.

We would like to thank all members of the Program Committee, as well as the additional referees listed below, for their help in ensuring the quality of the papers accepted for publication. We would also like to thank the Organizing Committee for all their efforts in making this workshop happen.

April 2007

Nicu Sebe
Yuncai Liu
Yueting Zhuang

International Workshop on Multimedia Content Analysis and Mining 2007 Organization

Organizing Committee

Honorary Chair	Tao Zhan, Shandong University, China
General Chairs	Thomans S. Huang, University of Illinois at Urbana-Champaign, USA Arnold Smeulders, University of Amsterdam, The Netherlands Dongfeng Yuan, Shadong University, China
Program Chairs	Yuncai Liu, Shanghai Jiao Tong University, China Nicu Sebe, University of Amsterdam, The Netherlands Yueting Zhuang, Zhejiang University, China
International Steering Committee	Wen Gao, Chinese Academy of Sciences, China Shih-Fu Chang, Columbia University, USA Kiyo Aizawa, University of Tokyo, Japan HongJiang Zhang, Microsoft Research Asia, China Alan Smeaton, Dublin City University, Ireland Ed Chang, University of California, Santa Barbara, USA & Google, China Bob Liang, Intel Research, USA

Program Committee

Kiyo Aizawa	University of Tokyo, Japan
Noboru Babaguchi	Osaka University, Japan
Susanne Boll	University of Oldenburg, Germany
Dick Bulterman	CWI, Amsterdam, The Netherlands
Shih-Fu Chang	Columbia University, USA
Homer Chen	National Taiwan University, Taiwan
Chang Wen Chen	Florida Institute of Technology, USA
Alberto Del Bimbo	University of Florence, Italy
Jana Dittman	Otto-von-Guericke-University, Germany
Wen Gao	Chinese Academy of Sciences, China
Forouzan Golshani	Wright State University, USA
William Grosky	University of Michigan, USA
Allan Hanbury	TU Vienna, Austria
Tiejun Huang	Chinese Academy of Sciences, China
Alejandro Jaimes	IDIAP, Martigny, Switzerland

Mingyan Jiang	Shandong University, China
John Kender	Columbia University, USA
Michael Lew	Leiden University, The Netherlands
Bob Liang	Intel, USA
Rainer Lienhart	University of Augsburg, Germany
Jiebo Luo	Eastman Kodak, USA
Ketan Mayer-Patel	University of North Carolina, USA
Nasir Memon	Polytech University, USA
Jan Nesvadba	Philips Research, The Netherlands
Yuhua Peng	Shandong University, China
Alan Smeaton	Dublin City University, Ireland
Ralf Steinmetz	Darmstadt University of Technology, Germany
Fengrong Sun	Shandong University, China
Qi Tian	University of Texas at San Antonio, USA
Marcel Worring	University of Amsterdam, The Netherlands
Changsheng Xu	Institute for Infocomm Research, Singapore

Additional Reviewers

Jehanzeb Abbas	University of Illinois at Urbana-Champaign, USA
Yen-Kuang Chen	Intel, USA
Charlie Dagli	University of Illinois at Urbana-Champaign, USA
Ivo Everts	University of Amsterdam, The Netherlands
Pablo Cesar Garcia	CWI, Amsterdam, The Netherlands
Jinhua Guo	University of Michigan, USA
Junfeng He	Columbia University, USA
Niels Henze	OFFIS, Oldenburg, Germany
Eva Hoerster	University of Augsburg, Germany
Simon Hoffmann	University of Augsburg, Germany
Huxiao Hu	University of Illinois at Urbana-Champaign, USA
Wei Jiang	Columbia University, USA
Bart Kroon	Delft University of Technology, The Netherlands
Aleksandar Ivanovic	University of Illinois at Urbana-Champaign, USA
Dawei Liang	Harbin Institute of Technology, China
Ming Liu	University of Illinois at Urbana-Champaign, USA
Jochen Lux	University of Augsburg, Germany
Noel O'Connor	Dublin City University, Ireland
Xi Shao	Nanjing University of Posts and Telecomm., China

Roberto Valenti	University of Amsterdam, The Netherlands
Jun Wang	Columbia University, USA
Dajun Wu	Institute for Infocomm Research, Singapore
Xun Xu	University of Illinois at Urbana-Champaign, USA
Xinguo Yu	Institute for Infocomm Research, Singapore
Eric Zavesky	Columbia University, USA
Zhenqiu Zhang	University of Illinois at Urbana-Champaign, USA
Xi Zhou	University of Illinois at Urbana-Champaign, USA

Table of Contents

Invited Contributions

Multimedia Analysis by Learning 1
Arnold W.M. Smeulders

Learning Concepts by Modeling Relationships 5
Yong Rui and Guo-Jun Qi

Emerging Issues for Multimedia Analysis and Applications 14
Kiyoharu Aizawa

The Real Problem of Bridging the "Semantic Gap" 16
John R. Smith

Multimedia Ontology Based Computational Framework for Video Annotation and Retrieval 18
Alberto Del Bimbo and Marco Bertini

Scalable Indexing for Perceptual Data 24
Arun Qamra and Edward Y. Chang

Searching One Billion Web Images by Content: Challenges and Opportunities 33
Zhiwei Li, Xing Xie, Lei Zhang, and Wei-Ying Ma

Challenges on Peer-to-Peer Live Media Streaming 37
Wen Gao and Longshe Huo

Video Streaming to Mobile Handheld Devices: Challenges in Decoding, Adaptation, and Browsing 42
Chang Wen Chen and Houqiang Li

Fixed-Mobile Convergence, Streaming Multimedia Services, and Peer-to-Peer Communication 52
Jason J. Yao and Homer H. Chen

Blind Passive Media Forensics: Motivation and Opportunity 57
Shih-Fu Chang

Security Models of Digital Watermarking 60
Qiming Li and Nasir Memon

Evolvement of DRM Schema: From Encryption to Interoperability and Monitoring 65
Tiejun Huang

Regular Contributions

Shape Retrieval Based on the Relativity of Chain Codes 76
Junding Sun and Xiaosheng Wu

3D Motion Segmentation from Straight-Line Optical Flow 85
Jing Zhang, Fanhuai Shi, Jianhua Wang, and Yuncai Liu

A Fast Adaptive ME Algorithm Based on H.264 95
Zhisheng Liu and Yuhua Peng

Shared Latent Dynamical Model for Human Tracking from Videos 102
Minglei Tong and Yuncai Liu

Panoramic Video Coding Using Affine Motion Compensated Prediction 112
Zheng Jiali, Zhang Yongdong, Shen Yanfei, and Ni Guangnan

Robust Bootstrapping of Speaker Models for Unsupervised Speaker Indexing 122
Fu ZhongHua

Reversible Data Embedding Technique for Palette Images Using De-clustering 130
Chin-Chen Chang, Yi-Hui Chen, and Yung-Chen Chou

The PT-2 Caching Algorithm in the Transcoding Proxy Cluster to Facilitate Adaptive Content Delivery 140
YongJu Lee, YuHyeon Bak, OkGee Min, HagYoung Kim, and CheolHoon Lee

Constructing and Application of Multimedia TV News Archives 151
H.T. Pao, Y.Y. Xu, S.C. Chung, and H.C. Fu

A Three-Level Scheme for Real-Time Ball Tracking 161
Xiaofeng Tong, Tao Wang, Wenlong Li, Yimin Zhang, Bo Yang, Fei Wang, Lifeng Sun, and Shiqiang Yang

Video Object Mining with Local Region Tracking 172
Arasanathan Anjulan and Nishan Canagarajah

Automatic Extraction of Semantic Relationships from Images Using Ontologies and SVM Classifiers 184
Jin-Woo Jeong, Kyung-Wook Park, Oukseh Lee, and Dong-Ho Lee

A Robust Caption Detecting Algorithm on MPEG Compressed Video 195
Yaowei Wang, Limin Su, and Qixiang Ye

A Hybrid Approach for Authenticating MPEG-2 Streaming Data 203
Meiqin Wang, Lin Li, S.M. Yiu, Lucas C.K. Hui, C.F. Chong, K.P. Chow, W.W. Tsang, H.W. Chan, and K.H. Pun

HOS-Based Image Super-Resolution Reconstruction 213
Jianping Qiao and Ju Liu

Story Unit Segmentation with Friendly Acoustic Perception 223
Longchuan Yan, Jun Du, Qingming Huang, and Shuqiang Jiang

Decomposition in Hidden Markov Models for Activity Recognition 232
Weidong Zhang, Feng Chen, Wenli Xu, and Zisheng Cao

Image Inpainting with Improved Exemplar-Based Approach 242
Qiang Chen, Yingxiang Zhang, and Yuncai Liu

Fingerprinting Codes for Live Pay-Television Broadcast Via Internet ... 252
Shuhui Hou, Tetsutaro Uehara, Yoshitaka Morimura, and Michihiko Minoh

A New Type of Proxy Ring Signature Scheme with Revocable Anonymity and No Info Leaked 262
Chengyu Hu, Pengtao Liu, and Daxing Li

The Research of an Embedded Processor Element for Multimedia Domain 267
Lai Mingche, Guo Jianjun, Lv Yasuai, Dai Kui, and Wang Zhiying

Speeding Up Scalar Multiplication Using a New Signed Binary Representation for Integers 277
Bang-ju Wang, Huan-guo Zhang, Zhang-yi Wang, and Yu-hua Wang

Color-Texture Image Segmentation by Combining Region and Photometric Invariant Edge Information 286
Shengyang Yu, Yan Zhang, Yonggang Wang, and Jie Yang

Multimedia Authoring Tool for Real-Time Facial Animation 295
Hae Won Byun

Object Re-detection Using SIFT and MPEG-7 Color Descriptors 305
Philipp Schügerl, Robert Sorschag, Werner Bailer, and Georg Thallinger

Interactive Boosting for Image Classification 315
Yijuan Lu, Qi Tian, and Thomas S. Huan

FBSA: A Self-adjustable Multi-source Data Scheduling Algorithm for P2P Media Streaming 325
Shuangjia Chen, Longshe Huo, Qiang Fu, Rui Guo, and Wen Gao

An Optimized Topology Maintenance Framework for P2P Media Streaming 334
Rui Guo, Longshe Huo, Qiang Fu, Shuangjia Chen, and Wen Gao

Building Large Scale 3D Face Database for Face Analysis 343
Yuxiao Hu, Zhenqiu Zhang, Xun Xu, Yun Fu, and Thomas S. Huang

Senary Huffman Compression – A Reversible Data Hiding Scheme for Binary Images 351
Chung-Chuan Wang, Chin-Chen Chang, Xinpeng Zhang, and Jinn-Ke Jan

Managing and Searching Distributed Multidimensional Annotations with Large Scale Image Data 361
Tian Xia, Fusheng Wang, Peiya Liu, and Sridharan Palanivelu

A Closed-Form Solution of Reconstruction from Nonparallel Stereo Geometry Used in Image Guided System for Surgery 371
Jianhua Wang and Yuncai Liu

SIEVE—Search Images Effectively Through Visual Elimination 381
Ying Liu, Dengsheng Zhang, and Guojun Lu

Virtual Community Based Secure Service Discovery and Access for 3D Video Steaming Applications 391
Shudong Chen, Igor Radovanovic, Johan Lukkien, Richard Verhoeven, Melissa Tjiong, and Remi Bosman

QoS Adaptive Data Organizing and Delivery Framework for P2P Media Streaming 398
Longshe Huo, Wen Gao, Qiang Fu, Rui Guo, and Shuangjia Chen

Adaptive Interpolation for Error Concealment in H.264 Using Directional Histograms 408
Toan Nguyen, Mi Seon Park, Won Taek Lim, Deok Jae Choi, Guee Sang Lee, and Jae Myung Yoo

Players and Ball Detection in Soccer Videos Based on Color Segmentation and Shape Analysis 416
Yu Huang, Joan Llach, and Sitaram Bhagavathy

Efficient Image Retrieval Using Conceptualization of Annotated Images 426
Miyoung Cho, Chang Choi, Hanil Kim, Jungpil Shin, and Pankoo Kim

3-D Camera Modeling and Its Applications in Sports Broadcast Video Analysis 434
Jungong Han and Peter H.N. de With

Face Recognition by Matching 2D and 3D Geodesic Distances 444
S. Berretti, A. Del Bimbo, P. Pala, and F.J. Silva Mata

A Blind Watermarking Scheme Based on Visual Model for Copyright Security 454
Cong Jin, Liang-Gang Pan, and Ting Su

Fast Mode Decision by Exploiting Spatio-temporal Correlation in H.264 464
Sung-Hoon Jeon, Tak-Gi Lee, Kwang-Mu Shin, Sung-Min Kim, and Ki-Dong Chung

SVM-Based Audio Classification for Content-Based Multimedia Retrieval 474
Yingying Zhu, Zhong Ming, and Qiang Huang

Moving Object Tracking in H.264/AVC Bitstream 483
Wonsang You, M.S. Houari Sabirin, and Munchurl Kim

A Prediction Error Compression Method with Tensor-PCA in Video Coding 493
Jian Liu, Fei Wu, Lei Yao, and Yueting Zhuang

On the Robustness of Parametric Watermarking of Speech 501
Aparna Gurijala and J.R. Deller Jr.

Author Index 511

Multimedia Analysis by Learning

Arnold W.M. Smeulders

Intelligent Systems Lab Amsterdam
University of Amsterdam
Amsterdam, The Netherlands
smeulders@science.uva.nl

Abstract. In this presentation for the panel at MCAM07, I put forward the transition of modeling the world as was done on a large scale in computer vision before the year 2000, to the current situation where there have been considerable successes with multimedia analysis by learning from the world. We make a plead for the last type of learned features, modeling only the scene accidental conditions and learning the object or object class intrinsic properties. In this paper, in respect to contributions by many others, we illustrate the approach of learning features by papers from our lab at the University of Amsterdam.

Keywords: multimedia analysis, invariants, learning, concepts.

In any sensory data analysis system, defining the problem at hand may already be half the solution. If looking into the history of computer vision, scene segmentation was redefined many times both in its place in the processing chain as in what it actually is. It can be said that any essentially new type of segmentation algorithms brings along a new definition of what segmentation actually is. Every new formulation of the problem brings along the appropriate class of data to work on, the list of goals to achieve precisely, the appropriate evaluation measure(s), and the appropriate set of previous methods to compare it to.

Let us consider the most trivial algorithm for segmentation: thresholding the intensity values. The goal is usually to localize the edges of the object precisely. Effectively this solves the problem of finding concatenated blobs of globally conspicuous intensity value(s). It could work on data like multi-spectra data, infra-red cameras, uniformly colored suits, non-black objects on a black conveyer belt, and other circumstances when there is a proper reason why a single set of intensity values would stand out from the rest. A proper evaluation measure is the number and position relative to the edge of the result relative to the reality. The list of methods to compare it to these days would include edge-based methods like snakes of Witkin and Duncan with built in edge smoothness criteria – when the result has to be accurate - and probabilistic shape profiles of Cootes and Taylor with built in edge knowledge when the result has to have a high probability of belonging to a class of shape with known variety. Alternatively, a region-based method can be chosen when the object has a fixed shape and a uniform background but carries high levels of noise. What is important here is to conclude that each of these various sophistications solve a slightly different problem depending on the amount of a priori knowledge or a priori information of the task or of the data which go into the solution.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 1 – 4, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In general, the current distinction is between the following two alternatives:

- an a priori model instantiated with problem-dependent constraints to which the solution in its match to the data should adhere in order to qualify for a result, or
- a learned set of observed instances to which an instance matched to the data demonstrates similarity, sufficiently so to call it a good result.

A good example for the first one is the snake model with a priori but generic integral constraints on smoothness and edge strengths. A good example for the second class is the probabilistic shape model assigning a probability to the shape. See also [1] for a combination by functional data analysis.

In the literature, a shift can be observed from modeling to learning. And the shift comes for good reasons. Modeling in essence is an abstract generic rule set derived from reality. To understand the world in a small set of rules is the ideal since the birth of civilization, hence a very praiseworthy endeavor. But for sensory data processing it remains to be seen whether the model as assumed is indeed valid in all circumstances or whether such kinds of models exist at all. When learning from sufficiently large datasets [6], in any rate it is clear that the data are realistic and so is the solution. So, by choosing the road of learning to analysis at any rate the solution is reality proof.

The shift from modeling to learning comes in a time when ordinary sensors turn digital, storage is no limitation, work processes in routine use build large scale digital repositories and even the processing power to perform the machine learning on have become available.

What is valid for the example of segmentation also holds for compression, although here the list of different goals per type of method is smaller. The same holds for image processing tasks as image enhancement and image repair. Where machine learning always provides a growing capacity in their sets of tools, they have achieved a level of general applicability in computer vision. Given the achievements on these disciplines, the attention shift towards features once more, as features are the essence of analysis.

There is no reason to deviate from the usual division of features. In the core we have intensity, color and multi-dimensional intensity feature values. They may be supported by Gaussian smoothing over a neighborhood of N x N pixels reducing their noise characteristics but not their intent to measure a point property. Spatial filtering is necessary for obtaining differential Taylor expansions of the image field but still it is best to think of it as a point property.

In the second place we have texture features measuring image distributions orderly – as a regular pattern usually by Gabor filters – or un-orderly by statistics of histograms of the above point properties [3] The region over which the texture field is measured is usually data driven expanding a seed as long as the data show spatial coherence or alternatively a mesh of fixed regions is imposed on the image field a priori. SIFT by Lowe and related features by Schmid can be seen as a sparse feature set of regions measured around conspicuous points in the image.

In the third place we have global features. Some of them are arrived at through a global transform taking the average over the image. This may be done over the point feature values as hinted at above, or it may be a summarization or averaging of regional feature values.

A relative new, forth step is the assignment of regions to intermediate types and to measure by similarity the combined support over all of them to the presence of objects or object classes in the image. The combination over the region may be performed by taking the maximum or the average. When taking the maximum, in effect we have an old-fashioned object template detector. When taking an average, we have a purely contextual feature detector. There are few papers discussing how to evaluate the combinatorics when the object is neither localized precisely nor entirely contextual [5,8].

Relative to the general consensus in the nineties, the amazing progression in this last approach is that for some visual concepts it appears easier to classify highly semantic concepts as an interview or a statesman. By the introduction contextual features and learning them from sufficiently many examples, it now is possible to learn a detector for these concepts without having a model of humans, their interaction and what the generic geometric model is for the appearance of an interview (or a statesman). In fact, it appears to be simpler to learn some of the highly semantic (but contextual) concepts then some of the low level ones, like the notoriously hard example of the object class of chairs. Chairs can have so many different appearances and forms, they are almost impossible to model or to learn from examples, yet their context is much more stable. They always appear in doors, well almost always. They appear on the ground, well etc. And they appear near a table (which is a much easier concept to model or learn).

In sensory features it is of crucial importance to make a distinction between the scene accidental influences in the feature values and the influence of the object or object class in them, see for example [2]. That is why features invariant to the scene accidental conditions are to be preferred over features which concentrate on the appearance of the object. Of course, when the target operation is to achieve an aim in an improved, restored or compressed appearance, under these circumstances it may be advantageous to use appearance features. But when the aim is to measure, classify or recognize an object or object class, it is by all means better to use features invariant to all relevant variations in the scene.

The general rule of thumb on the use of invariance is subscribed to by many. Nevertheless when measuring color in images, the use of color features invariant to the illumination intensity, direction or color is still modest. See for example [4,7] for robust color feature sets.

Not all adhere to the use of invariance features. They argue that if the data set is big enough, the accidental variance in the appearance due to the scene, the recording and the illumination will be averaged out in the data.

Similarly, progress has been made in the area of localizing features of saliency. The SIFT feature falls apart in a localization part, invariant to changes in the scale of the image, and generally stable to many variant conditions, and a feature value measurement of the local intensity layout, somewhat but less robust to explicit changes in the illumination conditions.

By combination of local or regional features if they invariant and on the basis of a sufficiently large dataset, the limits of what features can be learned from multimedia repositories has not yet seen the beginning yet.

References

[1] Ghebreab, S., Smeulders, A.W.M.: Strings: variational deformable models of multivariate continuous boundary features. IEEE Transactions Pattern Analysis Machine Intelligence 25(11), 1399–1410 (2003)
[2] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval: the end of the early years. IEEE Transactions Pattern Analysis Machine Intelligence 22(12), 1349–1380 (2000)
[3] Geusebroek, J.-M., Smeulders, A.W.M.: A six-stimulus theory for stochastic texture. International Journal of Computer Vision 62(1/2), 7–16 (2005)
[4] van den Weijer, J., Gevers, T., Smeulders, A.W.M.: Robust photometric invariant features from the color tensor. IEEE Transactions on Image Processing 15(1), 118–127 (2006)
[5] Snoek, C.G.M., Worring, M., Geusebroek, J.-M., Koelma, D.C., Seinstra, F.J., Smeulders, A.W.M.: The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. IEEE Transactions Pattern Analysis Machine Intelligence 28(10), 1678–1689 (2006)
[6] Geusebroek, J.-M., Burghouts, G., Smeulders, A.W.M.: The Amsterdam library of object images. International Journal of Computer Vision 61(1), 103–112 (2005)
[7] Geusebroek, J.-M., van den Boomgaard, R., Smeulders, A.W.M., Geert, H.: Color invariants. IEEE Transactions on Pattern Analysis Machine Intelligence 23(12), 1338–1350 (2001)
[8] van Gemert, J.C., Burghouts, G.J., Seinstra, F.J., Geusebroek, J.M.: Color invariant object recognition using entropic graphs. International Journal of Imaging Systems and Technology 16(5), 146–153 (2006)

Learning Concepts by Modeling Relationships

Yong Rui[1] and Guo-Jun Qi[2]

[1] Microsoft Corporation
Microsoft China R&D Group, 49 Zhichun Road, Beijing, 100080, China
yongrui@microsoft.com
[2] University of Science and Technology of China
Huang Shan Road, No.4, Hefei, Anhui, 230027, China
qgj@mail.ustc.edu.cn

Abstract. Supporting multimedia search has emerged as an important research topic. There are three paradigms on the research spectrum that ranges from the least automatic to the most automatic. On the far left end, there is the pure manual labeling paradigm that labels multimedia content, e.g., images and video clips, manually with text labels and then use text search to search multimedia content indirectly. On the far right end, there is the content-based search paradigm that can be fully automatic by using low-level features from multimedia analysis. In recent years, a third paradigm emerged which is in the middle: the annotation paradigm. Once the concept models are trained, this paradigm can automatically detect/annotate concepts in unseen multimedia content. This paper looks into this annotation paradigm. Specifically, this paper argues that within the annotation paradigm, the relationship-based annotation approach outperforms other existing annotation approaches, because individual concepts are considered jointly instead of independently. We use two examples to illustrate the argument. The first example is on image annotation and the second one is on video annotation. Experiments indeed show that relationship-based annotation approaches render superior performance.

Keywords: Relationship, Image Annotation, Video Annotation.

1 Introduction

Search is a hot topic these days. As search is getting more and more mature, it expands in multiple dimensions. From the content source perspective, there are web search, desktop search and enterprise search. From the domain perspective, there are general search and vertical search. From the media type perspective, there are text search and multimedia (e.g., image, video and audio) search. This paper will focus on general multimedia search. While the multimedia search accuracy is not as good as that of text search, it has been making sturdy progress. Early multimedia search, especially image search, can be traced back to the 70s. But it is the 90s that witnesses strong renaissance in multimedia search, with the emergency of QBIC, Virage and other systems. In general, there were two

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 5–13, 2007.
© Springer-Verlag Berlin Heidelberg 2007

paradigms to multimedia search, one is manual-annotation based and the other is content-based.

The manual-annotation based approach first annotates the multimedia content, e.g., images or video clips, by text and then use text-based search to search for multimedia content indirectly. However, there exist two major difficulties, especially when the size of content collections is large (tens or hundreds of thousands). One is the vast amount of labor required in manual annotation, and the other is the subjectivity of human perception of multimedia content. Opposite to the manual-annotation approach, the other extreme in the research spectrum is the pure automatic content-based approach. However, the difficult in the content-based approach is the semantic gap between the low-level features that we can automatically extract and the high-level concepts that are of interest to real-world users. In the past 8 years, researchers are gradually realizing that neither of the above two extremes gives good search/retrieval results. Instead, something in the middle strikes a better balance, and that is the automated annotation paradigm. This paradigm is not purely automatic, as people need to label/annotate some content. But it is not purely manual either - once a concept detector is trained based on the labeled data, the detector can automatically annotate the same concept for other video clips. This paper looks into this annotation paradigm. Specifically, this paper argues that within the annotation paradigm, the relationship-based annotation approach outperforms other existing annotation approaches, because individual concepts are considered jointly instead of independently.

The rest of the paper is organized as follows. In Section 2, we will review and discuss two different categories in image annotation, analyzing their strength and weakness. Then we will compare their performance and show that the category that models concept relationships gives better performance. Similarly, in Section 3, we will review and discuss three different categories in video annotation, analyzing their strength and weakness. Then we will compare their performance and show that the category that models concept relationships gives better performance. We give concluding remarks in Section 4.

2 Image Annotation

2.1 Different Categories

Before a learning machine can perform annotation, it needs to be trained first, and training samples need to be accurately labeled. The labeling process can be both time consuming and error-prone [13]. Fortunately, multiple instance learning (MIL) allows for coarse labeling at the image level, instead of fine labeling at pixel/region level, which significantly improves the efficiency of image categorization [5][15]. In the MIL framework, there are two levels of training inputs: bags and instances. A bag is composed of multiple instances. A bag is labeled positive if at least one of its instances falls within the concept, and it is labeled negative if all of its instances are negative. The efficiency of MIL lies in the fact that during training, a label is required only for a bag, not the instances in the

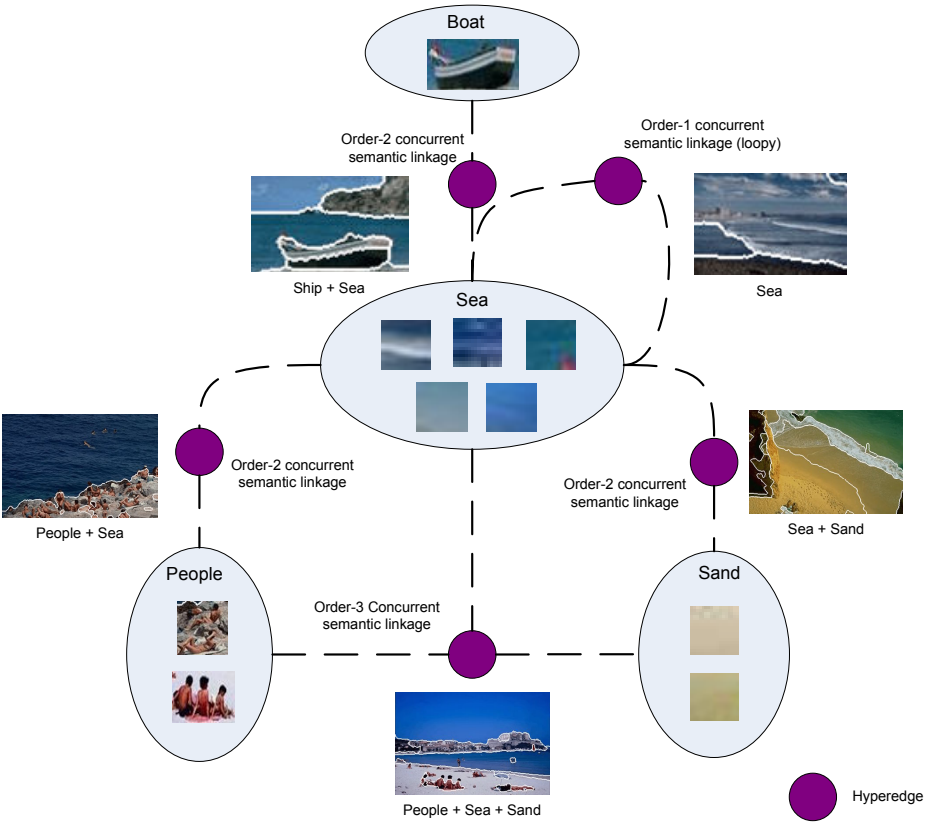


Fig. 1. Relationship between different concepts

bag. In the case of image annotation, a labeled image (e.g., a "beach" scene) is a bag, and the different regions inside the image are the instances [5]. Take Figure 1 as an example, while some of the regions are background and may not relate to "beach", other regions, e.g., sand and sea, do relate to "beach". If we exam more carefully, we can see that sand and sea tend to appear simultaneously in an image of "beach". Such a co-existence or concurrency can significantly boost the belief that an instance (e.g. the sand, the sea etc.) belongs to a "beach" scene. Therefore, in this "beach" scene, there exist order-2 concurrent relationship between the sea instance (region) and the sand instance (region). Similarly, in this "beach" scene, there also exist higher-order (order-4) concurrent relationship between instances, e.g., sand, sea, people, and sky (see Figure 1).

Existing MIL-based image annotation algorithms can be classified into two categories: the one that assumes that the instances in a bag are independent, and the one that explores concurrent relations. In the first category, Yang et al. [15] proposed an Asymmetric Support Vector Machine-based MIL (ASVM-MIL) algorithm by introducing asymmetric loss function for false positives and false

negative to exploit the instance label. An alternative approach, diverse density (DD) [5][16] uses a scaling and gradient search algorithm to find the prototype points in the instance space with the highest DD value.

Although this independence assumption significantly simplifies the modeling and computing procedure, it does not take into account the hidden information encoded in the semantic linkage among instances, as we described in the above "beach" example. The Concurrent MIL (ConMIL) scheme [8] is an approach in the second category that models the concurrent relationship between instances. Specifically, it uses concurrent tensor to model the semantic linkage between the instances. Then Rank-1 tensor factorization [17] is applied to decompose this concurrent tensor to obtain the label of each instance in the Reproducing Kernel Hilbert Space (RKHS) [18]. Since these obtained instance labels come from the encoded concurrent relations in the concurrent tensor, the formulated prediction model will reflect the semantic linkage between different instances and, accordingly, this model can be more accurate than those ones that predict instance label in isolation.

2.2 Experimental Comparison Between Different Categories

In this section, we will compare between approaches that do and do not model concurrency between instances on the most widely used COREL 5000 benchmark dataset. The COREL 5000 dataset has 50 semantically diverse categories, with each category containing 100 images. A set of low-level features is extracted from each region to represent an instance, including color correlogram, color moment, region size, wavelet texture and shape (normalized inertia of order 1, 2, 3) [1].

During the experiments, images within each category are randomly partitioned into two halves to form the training and the testing sets. There are various measurements for evaluating performance, including ROC curve, precision-recall curve, etc. The most widely measurement in recent years is AUC (area under the ROC curve) [4]. We will use AUC in this paper. The ROC curve plots the true positive rate (i.e. the recall) as a function of the false positive rate, and AUC measures the probability that a randomly chosen positive image will be ranked higher than a randomly chosen negative image.

We next compare the performance of ConMIL with that of representative approaches that do not model concurrency, e.g., EM-DD and ASVM-MIL, and results are summarized in Table 1 and figure 2. The following observations can be made:

Table 1. Average AUC for COREL 5000 dataset by EM-DD, ASVM-MIL, ConMIL

	Avg. AUC for COREL dataset
EM-DD [16]	0.775
ASVM-MIL [15]	0.836
ConMIL [8]	0.916

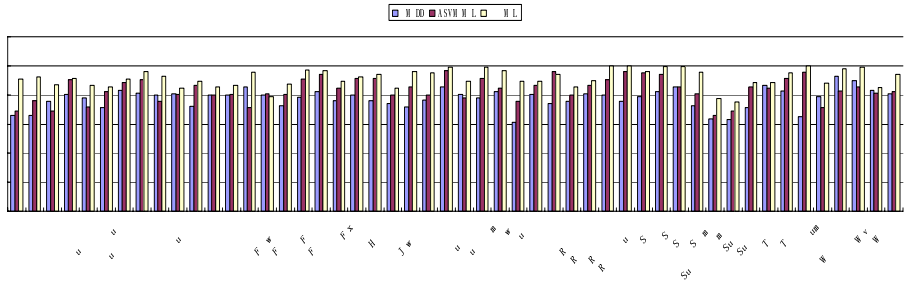


Fig. 2. AUC value of 50 categories by EM-DD, ASM-MIL and ConMIL

– Overall, ConMIL achieves the best results at AUC = 0.916.
– ConMIL performs well on categories with complex objects (e.g. see "building") or with complex scene (e.g. see "beach"). This is because ConMIL considers the semantic linkage of these concurrent regions and encodes this information into the inference of the instance labels.

3 Video Annotation

3.1 Different Categories

The video concepts of interest include a wide range of categories such as scenes (e.g., urban, sky, mountain, etc.), objects (e.g., airplane, car, face, etc.), and events (e.g., explosion-fire, people-marching, etc.) [11][12]. Research on video annotation evolved through three categories: independent concept detection, fusion-based concept detection, and integrated concept detection.

First Category: Independent Concept Detection. In this category, multiple video concepts are detected *individually* and *independently* without considering correlations between them. That is, the multi-label video annotation is translated into a set of binary detectors for each concept. A typical approach is to independently train a concept model using a classifier, e.g., support vector machine (SVM). The first-category approaches only achieved limited success. In real world, video concepts do not exist in isolation. Instead, they appear correlatively and naturally interact with each other at the semantic level. For example, the presence of "crowd" often occurs with the presence of "people" while "boat ship" and "truck" commonly do not co-occur. Furthermore, while simple concepts can be modeled directly from low level features, it is quite difficult to individually learn the models of complex concepts, e.g., "people marching", from the low-level features. Instead, the complex concepts can be better inferred based on the correlations with the other concepts. For instance, the presence of "people marching" can be boosted if both "crowd" and "walking running" occurs in a video clip.

Second Category: Fusion-Based Concept Detection. One of the most well-known approaches in this category is to refine the detection results of the individual detectors with a *Context Based Concept Fusion* (CBCF) strategy. For instance, Wu et al. [14] used an ontology-based multi-classification learning for video concept detection. Each concept is first independently modeled by a classifier, and then a predefined ontology hierarchy is used to improve the detection accuracy of the individual classifiers. Smith et al. [10] presented a two-step Discriminative Model Fusion (DMF) approach to mine the unknown/indirect relationship of specific concepts by constructing model vectors based on detection scores of individual classifiers. A SVM is then trained to refine the detection results of the individual classifiers.

Although it is intuitively correct that contextual relationship can help improve detection accuracy of individual detectors, experiments of the above CBCF approaches have shown that such improvement is not always stable. The unstable performance is due to the following reason. CBCF methods are built on top of the individual independent binary concept detectors with a second step to fuse them. However, the output of the individual independent detectors can be unreliable and therefore their detection errors can propagate to the second fusion step. As a result, the final annotations can be corrupted by these incorrect predictions. Indeed, the CBCF approaches do not follow the *principle of Least-Commitment* by D. Marr [6]. Instead, they are prematurely committed to irreversible individual predictions in the first step which may not be corrected in the second fusion step.

Third Category: Integrated Concept Detection. To address the difficulties faced in the first and second categories, a third category emerged [9]. The key of this category is to simultaneously model both the individual concepts and their interactions in a single formulation. The approach follows the *Principle of Least-Commitment* [6]. Because the learning and optimization is done in a single step for all the concepts simultaneously, it does not have the error propagation problem as in CBCF. Specifically, it develops a unified *Correlative Multi-Label* (CML) SVM to simultaneously classify concepts and model their correlations in a new feature space which encodes both the concept models and their interactions together.

To summarize, the first category does not address concept correlation. The second category attempts to address it by introducing another separate correlation step. The third category, on the other hand, addresses the correlation issue at the root in a single step.

3.2 Experimental Comparisons Between Different Categories

In this section, we compare the three categories on a widely used benchmark video data set from TRECVID [12]. The data set contains about 170 hours international broadcast news in Arabic, English and Chinese. These news videos are first automatically segmented into 61,901 subshots. All subshots are then processed to extract several kinds of low-level features, including

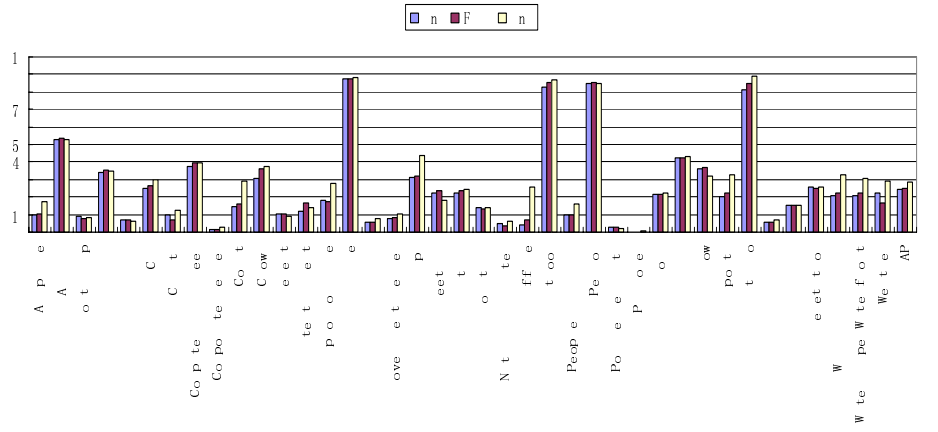


Fig. 3. The performance comparison of Ind, Fuse and Int

– Block-wise Color Moment in Lab color space;
– Co-occurrence Texture;
– Wavelet Texture;
– Edge Distribution Layout;

and some mid-level features

– Face - consisting of the face number, face area ratio, the position of the largest face.

For each subshot, 39 concepts are multi-labeled according to LSCOM-Lite annotations [7]. These annotated concepts consist of a wide range of genres, including program category, setting/scene/site, people, object, activity, event, and graphics.

We next compare three different approaches. The first-category approach, called Ind, is the combination of multiple binary encoded SVMs which are trained independently on each concept. The second- category approach is developed by adding a contextual fusion level on the detection output of the first approach [3]. In our implementation, we use the SVM for this fusion level. We denote this second-paradigm approach as Fuse in this section. The third- category approach [9] is denoted as Int.

The video data is divided into 3 parts with 65% (40,000 subshots) as training set, 16% (10,000 subshots) as validation set and the remaining 19% (11,901 subshots) as test set. For performance evaluation, we use the official performance metric Average Precision (AP) in the TRECVID tasks to evaluate and compare the algorithms on each concept. The AP corresponds to the area under a non-interpolated recall/precision curve and it favors highly ranked relevant subshots. We average the AP over all the 39 concepts to create the mean average precision (MAP), which is the overall evaluation result.

Figure 3 illustrates the performance of between Ind (first category), Fuse (second category), and Int (third category). We can observe that Int obtains about 0.038 and 0.03 overall improvement over Ind and Fuse respectively in terms of mean average precision (MAP). In other words, Int obtains about 15.4% and 12.2% relative improvements compared to Ind and Fuse. Such an improvement is statistically significant. Furthermore, between these 3 approaches, Int performs the best on 28 of the all 39 concepts.

4 Concluding Remarks

In this paper, we reviewed different paradigms for multimedia search. We argue that the semi-automatic annotation paradigm is the most fruitful. We further argue that within the annotation paradigm, the approaches that model concept relationships outperform the ones do not.

References

1. Chen, Y., Wang, J.Z.: Image Categorization by Learning and Reasoning with Regions. In: Journal of Machine Learning Research 5, 913–939 (2004)
2. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research 2, 265–292 (2001)
3. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, Springer, Heidelberg (2004)
4. 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)
5. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: Proc. International Conference on Machine Learning, pp. 341-349 (1998)
6. Marr, D.: Vision. In: W.H. Freeman and Company (1982)
7. Naphade, M.R., Kennedy, L., Kender, J.R., Chang, S.-F., Smith, J.R., Over, P., Hauptmann, A.: A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. In: IBM Research Technical Report (2005)
8. Qi, G.-J., Hua, X.-S., Rui, Y., et al.: Concurrent Multiple Instance Learning for Image Categorization. In: Proc. of CVPR 2007 (2007)
9. Qi, G.-J., Hua, X.-S., Rui, Y., et al.: Correlative Multi-Label Video Annotation. In: Pre-prints of ACM Multimedia 2007 (submission)
10. Smith, J.R., Naphade, M., Natsev, A.: Multimedia semantic indexing using model vectors. In: IEEE International Conference on Multimedia and Expo (2003)
11. Snoek, C., Worring, M., Gemert, J., Geusebroek, J.-M., Smeulders, A.: The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In: Proceedings of the ACM International Conference on Multimedia October 2006, pp. 421–430, Santa Barbara, USA (2006)
12. TRECVID. TREC video retrieval evaluation. `http://www. nlpir.nist.gov/projects/trecvid/`
13. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: Proc. of Advances in Neural Information Processing System (2005)

14. Wu, Y., Tseng, B.L., Smith, J.R.: Ontology-based multi-classification learning for video concept detection. In: IEEE International Conference on Multimedia and Expo (2004)
15. Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multi-instance learning. In: Proc. of IEEE International Conference on CVPR (2006)
16. Zhang, Q., Goldman, S.A.: Em-dd: An improved multiple-instance learning technique. In: Proc. of Advances in Neural Information Processing System (2001)
17. Kofidis, E., Regalia, P.: On the Best Rank-1 Approximation of Higher-order Supersymmetric Tensors. In: SIAM Journal on Matrix Analysis and Applications 23(3), 863–884 (2002)
18. Schölkopf, B., Herbrich, R., Smola, A.J.: A Generalized Representer Theorem. In: Proc. of the Annual Conference on Computational Learning Theory, pp. 416–426 (2001)

Emerging Issues for Multimedia Analysis and Applications

Kiyoharu Aizawa

Univ. of Tokyo, Dept. of E.E. and Dept. of Information and Communication Eng.
7-3-1 Hongo Bunkyo, Tokyo 113-8656 Japan
aizawa@hal.t.u-tokyo.ac.jp

1 Multimedia Is Everywhere

In multimedia community, multimedia content such as TV news, sports, movies have been mainly the target of the research and development. In addition to those traditional public ones, personal multimedia content is growing rapidly with advances in technology in computers, storage, sensors and networking. Affordable methods for capture, store and retrieval of human activities become available, and our life are digitized by many different ways. Personal cameras, camera installed mobile phones, and surveillance cameras are widely in use and produce a large amount of multimedia data. Not only audio-visual sensors but also various sensors such as motion sensors, acceleration sensors and RFIDs etc, give rise to many different types of data which reflect human activities. Emerging sensor networks will produce more.

In extreme cases, home-like environments such as Ubiquitous Home full of various sensing devices may capture lives of a family, or wearable sensing devices such as SenseCam, SenseWear etc, may be widely used to capture personal activities. The result of capturing human activities is a huge amount of multimedia data. New applications, different from those of traditional multimedia processing of TV and movies, will appear. Analysis of personal human activities will lead to variety of new applications such as multimedia e-Chronicle, healthcare, care for child and elderly, surveillance, behavioral studies, etc.

2 Challenges

There exist many challenging issues in order to develop new multimedia applications by capture of personal life. Difficulties in handling such data are

- Huge amount of data, which is continuously growing
- A large number of sources and multiple sensing modalities.
- Unstructured content
- Privacy.

The data captured is huge and much unstructured in its nature, differing to the well organized TV programs and movies, which makes retrieval hard. Privacy is always the concern for personal data. Navigation, summarization of personal activities, mining of important events in them and connecting them using a number of

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 14 – 15, 2007.
© Springer-Verlag Berlin Heidelberg 2007

multimodality sensor data are challenging tasks. We tend to think we know ourselves, but in reality, we do not. In the panel, I would like to mention the importance of the personal multimedia analysis and show, as an example, our effort in capturing and retrieval of personal experiences in home-like environment etc.

The Real Problem of Bridging the "Semantic Gap"

John R. Smith

Intelligent Information Management Department
IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532 USA
jrsmith@watson.ibm.com

Abstract. The "semantic gap" is a well-know problem in multimedia. The challenge is to accurately classify and effectively search multimedia content from automatically extracted low-level audio-visual features. While much effort is focused on developing the best machine learning approaches, not enough attention is placed on the required semantic coverage and the real utility of the classifiers in multimedia systems. Exacerbating the problem is the tremendous dearth of reliable training data across the diversity of semantics required for effective multimedia search in practice. Until we address the problems of what classifiers are needed in the large scale and how they can be sufficiently trained, research on bridging the "semantic gap" will not produce any real impact.

Video search is the current hot topic on the Internet. Today's category leaders in video search are doing their best to exploit speech, text, user tags and other types of metadata to enable search. So far, the results are not good. There are some obvious limitations. For example, almost all of today's video search is limited to narrow pockets of content. Broad video search across the Internet does not exist. The quality of video search is also an issue. Today's metadata is too sparse and unreliable to provide effective search. For example, a query for "basketball" returns hundreds of thousands of matches on the most prominent sites with no rhyme or reason to why matches are at top or bottom. Some video clips show basketball games, others discuss basketball, or others have very loose correlation. This query typifies video search today, which cannot find the most relevant matches and does not give user meaningful ways to refine search to find what they want.

Clearly video search can be improved. Bridging the semantic gap is a good way to do it. Accordingly, the topic is of great interest to the research community. However, the research is focusing too much on core technology and not enough making it work in practice. The real semantic gap is the lack of understanding on what semantics are important for enabling effective video search. Clearly, the overall semantic space is broad, e.g., covering sports, news, entertainment, personal, etc. Potentially, the space is also deep, e.g., covering goals, penalty kicks, tackles, etc. What is not known is the required breadth and depth of the semantic space required for effective video search. Even if this was known, there is no obvious source for training data to build classifiers. Further unknown is how the quality of the classifiers would contribute to the overall video search performance.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 16 – 17, 2007.
© Springer-Verlag Berlin Heidelberg 2007

There are some notable recent research efforts on the semantic spaces for video classifiers, such as the LSCOM lexicon of 1,000 semantic concepts for broadcast news video (http://www.ee.columbia.edu/ln/dvmm/lscom/) and the MediaMill challenge of 101 detectors (http://www.science.uva.nl/research/mediamill/challenge/). These are proving to be valuable resources for research. They were designed for and experimenting with algorithms for video classification and not for enabling effective video search in practice. On this dimension, they have yet to prove their value.

To bridge the semantic gap, we don't need more algorithms. We need to know what to model and where to get the data to model it. The research community if full of great ideas but it is not backed with the necessary data to make great things happen. The more we can focus on building out the semantic spaces for video content, the closer our research will come to having real impact.

Multimedia Ontology Based Computational Framework for Video Annotation and Retrieval

Alberto Del Bimbo and Marco Bertini

Università di Firenze - Italy
Via S. Marta, 3 - 50139 Firenze
delbimbo@dsi.unifi.it

Abstract. Ontologies are defined as the representation of the semantics of terms and their relationships. Traditionally, they consist of concepts, concept properties, and relationships between concepts, all expressed in linguistic terms. In order to support effectively video annotation and content-based retrieval the traditional linguistic ontologies should be extended to include structural video information and perceptual elements such as visual data descriptors.

These extended ontologies (referred in the following as multimedia ontologies) should support definition of visual concepts as representatives of specific patterns of a linguistic concept. While the linguistic part of the ontology embeds permanent and objective items of the domain, the perceptual part includes visual concepts that are dependent on temporal experience and are subject to changes with time and perception. This is the reason why dynamic update of visual concepts has to be supported by multimedia ontologies, to represent temporal evolution of concepts.

1 Position

The relevance of audio-visual data and more in general of media other than from text in modern digital libraries has grown in the last few years. Modern digital libraries are expected to include digital content of heterogeneous media, with almost the same percentage of text and visual data, and to a lower extent of audio and graphics, with the perspective that visual data will have the major role in many specialized contexts very soon. In this scenario, digital video is the media that has probably the highest relevance. There are, in fact, huge amounts of digital video daily produced for news, personal entertainment, educational, and institutional purposes, by broadcasters, media companies, government institutions and individuals. However, video presents important challenges to digital librarians, mainly due to the size of files, the temporal nature of the medium, and the lack of bibliographic methods that leverage non-textual features. In principle, similarly to images, video can be accessed according to either perceptual or semantic cues of the sequence content. But indeed, search and retrieval at syntactic level is impractical. Few examples using the QBE approach for content-based retrieval at perceptual level have been presented for video [1, 2, 3], none of them of practical interest. In fact, creating dynamic examples is utmost complex and

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 18–23, 2007.
© Springer-Verlag Berlin Heidelberg 2007

unfeasible for common users. According to this, librarians have traditionally indexed video collections with textual data (expliciting the producer's name, the date and time, and a few linguistic concepts that summarize video content at semantic level), so that they can be accessed almost in the same way as textual documents. Effective examples of retrieval by content of video clips using textual keywords have been presented for news [4, 5, 6] and sports video domains [7, 8], among the many others.

In this context, a crucial importance is given to the availability of appropriate metadata schemes and cross-language access techniques [9, 10], and appropriate tools for the organization of concepts into knowledge structures, also referred to as ontologies [11], so as to support effective access to video content for annotation and content based retrieval [12]. Ontologies are defined as the representation of the semantics of terms and their relationships. Traditionally, they consist of concepts, concept properties, and relationships between concepts, all expressed in linguistic terms. In particular, video ontologies can describe either the video content domain - in this case they are static descriptions of entities and highlights that are displayed in the video and their relationships, as codified by human experience - or the structure of the media - in this case they describe the component elements of the video, the operations allowed on its parts, and the low-level video descriptors that characterize their content.

However, traditional domain ontologies, are substantially inadequate to support complete annotation and retrieval by content of video documents. In fact, concepts and categories expressed in linguistic terms, are not rich enough to fully describe the diversity of the visual events that are present in a video sequence and cannot support video annotation and retrieval up to the level of detail of a pattern specification. A simple example may help to understand this point. Let's consider the video of a soccer game, and attack action highlights. The highlights that can be classified as attack actions have many different patterns. The patterns may differ each other by the playfield zone where the action takes place, the number of players involved, the players' motion direction, the speed and acceleration of the key player, etc. Although we are able to distinguish between them, and make clustering of attack actions into distinct classes, nevertheless, if we would like to express each pattern in linguistic terms we should use a complex sentence, explaining the way in which the action was developed. The sentence indeed should express the translation of the user's visual memory of the action into a conceptual representation where concepts are concatenated according to spatio-temporal constraints. In this translation, some visual data will be lost (we typically make a synthesis that retains only the presumed most significant elements), some facts will not be reported, and, probably most important, it will be possible that the appropriate words to distinguish one pattern from the other are not found.

Some early attempts to solve the inadequacy of traditional linguistic ontologies to support modeling of domain knowledge up to the level of pattern specification, have guessed the need that video domain ontologies incorporate both conceptual and perceptual elements. The basic idea behind all these researches is that,

although linguistic terms are appropriate to distinguish between broad event and object categories in generic domains, they are substantially inadequate when they must describe specific patterns of events or entities, like those that are represented in a video, and more in general in any perceptual media.

In [13], Jaimes et al. suggested to categorize the concepts that relate to perceptual facts into classes, using *modal keywords*, i.e. keywords that represent perceptual concepts in several categories, such as visual, aural, etc. Classification of the keywords was obtained automatically from speech recognition, queries or related text. In [14], perceptual knowledge is instead discovered grouping previously annotated images into clusters, based on their visual and text features, and extracting semantic knowledge by disambiguating the senses of the words in the annotations with WordNet and image clusters. Visual prototypes instances are then manually linked to the domain ontology. In [15, 16], a linguistic ontology for news videos is used to find recommendations for the selection of sets of effective visual concepts. A taxonomy of almost 1000 concepts was defined providing a multimedia controlled vocabulary. The taxonomy was mapped into the Cyc knowledge base [17] and exported to OWL.

Other authors have explored the possibility of extending ontologies so that also structural video information and visual data descriptors could be accounted for. In [18], a Visual Descriptors Ontology and a Multimedia Structure Ontology, respectively based on MPEG-7 Visual Descriptors and MPEG-7 Multimedia Description Schema, are used together with a domain ontology in order to support video content annotation. A similar approach was followed by [19] to describe sport events. In [20], a hierarchy of ontologies was defined for the representation of the results of video segmentation. Concepts were expressed in keywords using an *object ontology*: MPEG-7 low-level descriptors were mapped to intermediate level descriptors that identify spatio-temporal objects.

In all these solutions, structural and media information are still represented through linguistic terms and fed manually to the ontology. In [21] three separate ontologies modeling the application domain, the visual data and the abstract concepts were used for the interpretation of video scenes. Automatically segmented image regions were modelled through low-level visual descriptors and associated to semantic concepts using manually labeled regions as a training set. Text information available in videos, obtained through automatic speech recognition and manual annotation, and visual features were automatically extracted and manually assigned to concepts or properties in the ontology in [22]. In [23], qualitative attributes that refer to perceptual properties like color homogeneity, low-level perceptual features like model components distribution, and spatial relations were included in the ontology. Semantic concepts of video objects were derived from color clustering and reasoning. In [24] the authors have presented video annotation and retrieval based on high-level concepts derived from machine learned concept detectors that exploit low level visual features. The ontology includes both semantic descriptions and structure of concepts and their lexical relationships, obtained from WordNet.

Other researchers have proposed integrated multimedia ontologies that incorporate both linguistic terms and visual or auditory data and their descriptors. In [25] a multimedia ontology, referred to as "pictorially enriched ontology", was proposed where concepts with a visual counterpart (like entities, highlights or events) were modeled with both linguistic terms and perceptual media, like video and images. Perceptual media were associated with descriptors of their structure and appearance. In [26] a taxonomy was defined for video retrieval, where visual concepts were modeled according to MPEG-7 descriptors. Video clips were manually classified according to the taxonomy and unsupervised clustering was employed to cluster clips with similar visual content.

However, none of these researches has fully exploited the distinguishing aspects and the consequences of accounting, in the ontology, both the perceptual and conceptual facts. Within multimedia ontologies, the linguistic and perceptual parts have substantially different characteristics and behaviors. On the one hand, the linguistic part of the ontology embeds permanent, self-contained, objective items of exact knowledge, that can be taken as concrete and accepted standards by which reality can be considered. On the other hand, the perceptual part includes concepts that are not abstractions, but are mere duplicates of things observed in reality, and correlated to entities. Differently from the abstract concepts, these concepts are dependent on temporal experience, and hence subjected to changes with time and perceptions. They are not expressed in linguistic terms but rather by a set of values that are attributed to a few features that are agreed to model the perceptual content of the concept. These values have a substantially different dynamics than linguistic terms. In that they are representatives or prototypes of patterns in which real objects or facts manifest, they might change in dependency of the patterns observed and their number. In other words they depend on the "experience" of the context of application.

According to this, multimedia ontologies, in that they permit to link concepts to their real manifestations, should adjust the prototypes of the perceptual facts to the changes and modifications that may occur through time, using proper mechanisms for updating the prototypes of perceptual patterns, as the result of knowledge evolution.

In the reported contributions, perceptual elements of the ontology are instead regarded as static entities, similarly to the abstract concepts expressed in linguistic terms. These elements are defined at the time in which the ontology is created and no mechanism that supports their modification through time is conceived in the ontology.

We believe instead that evolutionary multimedia ontologies architectures must include include both abstract concepts and descriptors of perceptual facts and must support mechanisms for temporal evolution and update of the visual concepts. We have experimented a novel ontology architecture, for video annotation and search, where the visual prototype acts as a bridge between the domain ontology and the video structure ontology. While the linguistic terms of the domain ontology are fed manually according to the agreed view of the context of application and typically present little or no changes through time, representatives

of perceptual facts are fed automatically and may change on the basis of the observations. Unsupervised pattern clustering is the mechanism used to define and update the perceptual part of the ontology. Patterns in which perceptual facts manifest are distinguished each other by clustering their descriptors, and the centers of the clusters that are obtained are assumed as pattern prototypes in the multimedia ontology, i.e. as representatives of classes of patterns that have distinct appearance properties. New patterns that are analyzed for annotation are considered as new knowledge of the context. Therefore, as they are presented to the ontology, they modify the clusters and their centers.

We discuss examples with application to the soccer domain considering highlight clustering. Effects of adding new knowledge, and particularly the changes induced in the clusters (some patterns can move from one cluster to another and their cluster centers are redefined accordingly) are discussed and presented with reference to the soccer championship domain. We also discuss the way in which use of OWL DL for the multimedia ontology description allows to use a reasoning engine to extend and exploit the knowledge contained in the ontology to perform extended annotations and retrieval by content according to video semantics.

References

1. Naphade, M.R., Wang, R., Huang, T.S.:Multimodal pattern matching for audio-visual query and retrieval. In: Proc. of SPIE Storage and Retrieval for Media Database (2001)
2. Zhang, H., Wang, A., Altunbask, Y.: Content based video retrieval and compression: A unified solution. In: Proc. of the IEEE International Conference on Image Processing (1997)
3. Chang, S.F., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: Videoq: An automated content based video search system using visual cues. In: Proc. of the IEEE International Conference on Image Processing (1997)
4. Eickeler, S., Muller, S.: Content-based video indexing of tv broadcast news using hidden markov models. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 2997–3000 (1999)
5. Sato, T., Kanade, T., Hughes, E.K., Smith, M.A.: Video ocr for digital news archive. In: IEEE International Workshop on Content–Based Access of Image and Video Databases CAIVD' 98, pp. 52–60 (1998)
6. Hauptmann, A., Witbrock, M.: Informedia: News–on–demand multimedia information acquisition and retrieval. Intelligent Multimedia Information Retrieval, 213–239 (1997)
7. Ekin, A., Tekalp, A.M., Mehrotra, R.: Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing 12(7), 796–807 (2003)
8. Yu, X., Xu, C., Leung, H., Tian, Q., Tang, Q., Wan, K.W.: Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In: ACM Multimedia 2003. Berkeley, CA (USA), 4-6 November 2003, vol. 3, pp. 11–20 (2003)
9. Marchionini, G., Geisler, G.: The open video digital library. D-Lib Magazine 8(11) (December 2002)
10. European cultural heritage online (echo). Technical report (2002), `http://echo.mpiwg-berlin.mpg.de/home`

11. Gruber, T.: Principles for the design of ontologies used for knowledge sharing. Int. Journal of Human-Computer Studies 43(5-6), 907–928 (1995)
12. Athanasiadis, T., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., Kompatsiaris, Y.: Using a multimedia ontology infrastructure for semantic annotation of multimedia content. In: Proc. of 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot '05), Galway, Ireland (November 2005)
13. Jaimes, A., Tseng, B., Smith, J.: Modal keywords, ontologies, and reasoning for video understanding. In: Int'l Conference on Image and Video Retrieval (CIVR) (July 2003)
14. Benitez, A., Chang, S.F.: Automatic multimedia knowledge discovery, summarization and evaluation. IEEE Transactions on Multimedia (Submitted) (2003)
15. Kender, J., Naphade, M.: Visual concepts for news story tracking: Analyzing and exploiting the nist trecvid video annotation experiment. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 1174–1181 (2005)
16. Naphade, M., Smith, J., Tesic, J., Chang, S., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE Multimedia 13(3), 86–91 (2006)
17. Lenat, D., Guha, R.: In: Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Reading (1990)
18. Strintzis, J., Bloehdorn, S., Handschuh, S., Staab, S., Simou, N., Tzouvaras, V., Petridis, K., Kompatsiaris, I., Avrithis, Y.: Knowledge representation for semantic multimedia content analysis and reasoning. In: European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (November 2004)
19. Vembu, S., Kiesel, M., Sintek, M., Bauman, S.: Towards bridging the semantic gap in multimedia annotation and retrieval. In: Proc. First International Workshop on Semantic Web Annotations for Multimedia (SWAMM), Edinburgh (Scotland) (May 2006)
20. Mezaris, V., Kompatsiaris, I., Boulgouris, N., Strintzis, M.: Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. IEEE Transactions on Circuits and Systems for Video Technology 14(5), 606–621 (2004)
21. Simou, N., Saathoff, C., Dasiopoulou, S., Spyrou, E., Voisine, N., Tzouvaras, V., Kompatsiaris, I., Avrithis, Y., Staab, S.: An ontology infrastructure for multimedia reasoning. In: Proc. International Workshop VLBV 2005, Sardinia (Italy) (September 2005)
22. Jaimes, A., Smith, J.: Semi-automatic, data-driven construction of multimedia ontologies. In: Proc. of IEEE Int'l Conference on Multimedia & Expo. (2003)
23. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.K., Strintzis, M.G.: Knowledge-assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology 15(10), 1210–1224 (2005)
24. Snoek, C., Huurnink, B., Hollink, L., de Rijke, M., Schreiber, G., Worring, M.: Adding semantics to detectors for video retrieval. IEEE Transactions on Multimedia(Pending minor revision) (2007)
25. Bertini, M., Cucchiara, R., Del Bimbo, A., Torniai, C.: Video annotation with pictorially enriched ontologies. In: Proc. of IEEE Int'l Conference on Multimedia & Expo. (2005)
26. Grana, C., Bulgarelli, D., Cucchiara, R.: Video clip clustering for assisted creation of mpeg-7 pictorially enriched ontologies. In: Proc. Second International Symposium on Communications, Control and Signal Processing (ISCCSP), Marrakech, Morocco (March 2006)

Scalable Indexing for Perceptual Data

Arun Qamra[1] and Edward Y. Chang[2]

[1] Dept of Computer Science, University of California Santa Barbara
[2] Google Research

Abstract. In recent years, multimedia objects such as images, video, and audio are becoming increasingly widespread. Many applications require content-based retrieval to be performed, and measurement of distance is a key component in such scenarios. The nature of multimedia requires perceptual similarity to be captured when computing distance between objects. Measures such as the Euclidean distance, which utilize all attributes of a pair of objects, do not perform very well. Instead, distance measures that use partial matches between objects have been found to perform significantly better. This is because, two multimedia objects can be considered perceptually similar when some respects closely match, even when they are very different in other respects. Existing distance measures that capture partial similarity have limitations, such as their non-metric nature, which makes scalable indexing challenging. In this paper, we propose the Partial Match Function, a distance measure that performs well for perceptual data, and allows efficient indexing.

1 Introduction

In multimedia retrieval applications, multidimensional vectors are often used to represent multimedia objects For instance, an image may be represented by a vector whose coordinates represent various color and texture properties of the image. Such a representation can then be used to make judgments about the (dis)similarity of two objects. A common technique to measure dissimilarity in such multi-dimensional spaces is to compute the Euclidean distance function or some other L_p norm. However, such measures have some limitations.

First, Euclidean and like measures are unable to uncover partial similarities between objects, because they are cumulative measures based on all attributes for a pair of objects. In some cases, two objects may be very similar in some respects (attributes) but dissimilar in other respects. This partial similarity is significant, so it is desirable to uncover it in retrieval. To understand why the idea of partial similarity is important, suppose we are asked to name two places that are similar to England. Among several possibilities, Scotland and New England could be two reasonable answers. However, among all attributes for the places, England and New England are similar in one set of attributes, while Scotland and England are similar along a different set of attributes. If we compare England and New England using all attributes, the dissimilar aspects of the two will adversely impact the similarity score, which is not desirable. Thus, when judging

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 24–32, 2007.
© Springer-Verlag Berlin Heidelberg 2007

the similarity, a partial match based on the similar aspects or attributes must be used.

The second limitation of measures like the Euclidean distance is seen where one or more of the attributes of an object are incorrectly captured, resulting in a large difference along that attribute. Since all attributes are used in distance computation, the large difference adversely impacts the overall distance computation. Consider, for example, three objects A, B, and C. Let these objects be described by the attribute sets $\{0.01, 0.01, 0.01, 0.01, 0, 01, 0.01\}$, $\{0.02, 002, 0.02, 0.02, 0.02, 1\}$, and $\{0.1, 0.1, 0.1, 0.1, 0.1, 0.1\}$. While the last attribute of B may be so large only because it has been incorrectly captured, and object B is very close to A in all other attributes, Euclidean distance would find C to be closer to A.

Recently, the Dynamic Partial Function (DPF) [1] and Frequent k-n-match [2] have been proposed to measure dissimilarity between vectors while capturing partial similarities. These distance functions can uncover partial matches, and can thus be very useful in multimedia (e.g., image and video retrieval) where perceptual similarity may be partial [3]. However, both DPF and Frequent k-n-match are non-metric since they do not satisfy triangle inequality, which creates challenges for indexing, since most methods work by exploiting properties of metric spaces. DPF requires a data-dependent parameter m to be set, whereas Frequent k-n-match uses parameters n_0 and n_1 that must be tuned to control accuracy and cost. Also, Frequent k-n-match is a method designed for the retrieval scenario, and cannot be used to measure distance for clustering and classification. Indexing for DPF was proposed in [4], but retrieval efficiency is limited because triangle inequality cannot be used to prune candidate objects.

In this work, we propose a distance measure called Partial Match Function(PMF). PMF works by creating a binary transformation of the original vectors. This transformation is such that the dissimilar aspects of two objects being compared have limited effect, thus allowing the similar aspects to be better matched and partial similarities to be discovered. From the example used earlier, let the vectors $\{0.2, 0.2, 1\}$ and $\{0.1, 0.1, 0.1\}$ represent Scotland and England respectively. We see that the two places are similar in the first two attributes, and dissimilar in the third. This dissimilarity will clearly have a negative impact on the Euclidean distance, making the objects seem dissimilar. PMF can limit the effect of partial dissimilarity. Transforming the two vectors into the binary vectors 100101110 and 110100011, for instance, would cause the dissimilarity in the third attribute (which is captured by the last three bits) to be relatively reduced. As a result, the overall distance would be able to capture the partial similarity between the two objects.

Empirical results show that PMF is a powerful distance measure that does not have the limitations of measures proposed in [1] and [2], while modeling similarity as well as or better than these measures, and significantly better than the Euclidean distance. The PMF index allows efficient nearest neighbor retrieval. The PMF index, similar to the VA-file [5], works by storing a compressed representation of the object vectors. Unlike the VA-file, retrieval is performed using

only the compressed representation; the original object vectors are not used. This compressed representation can be cached in memory, or accessed sequentially on disk. While retrieval with tree-based index structures degenerates to the cost of a sequential scan as dimensionality becomes high, compression-based indexing methods are not affected by this *curse of dimensionality* [6,7,8]. Thus PMF is able to scale to very high D. Since the original object vectors are not used, only the compressed index needs to be stored, thus saving on storage requirement.

2 Related Work

Recently, some works have proposed to measure similarity based on a just a subset of dimensions. The set of dimensions used depends on the the pair of objects being compared, and is selected dynamically in a pairwise fashion. Doing so allows partial similarities to be captured. The first is the Dynamic Partial Function (DPF) measure which was presented in [4,9,1]. DPF models similarity based on the m most similar attributes of the pair of objects being compared. All other attributes are ignored, which allows pairwise distances to correspond to the similar aspects of two objects, and not be adversely affected by the dissimilar aspects. However, while DPF models similarity well, it has some shortcomings. First, DPF is non-metric, since it does not follow the triangle inequality. This is a limitation when indexing objects for retrieval, and also for clustering and classification,since most methods require the triangle inequality to be fulfilled. Secondly, DPF depends on an parameter m, which is selected in a data-dependent manner.

A second measure, Frequent k-n-match, was proposed in [2]. It is based on the n-match distance which measures similarity between a pair of vectors by computing the n^{th} smallest difference in dimension values across the pair, from among all dimensions. The k-n-match retrieves the k nearest neighbors for a given object, using the n-match distance. Frequent k-n-match extends this by retrieving the top-k NNs based on each of a range of values for n, and then aggregating results by selecting the k objects that appear most frequently. The lower and upper bounds of the range of n values are denoted by n_0 and n_1 respectively. Like DPF, Frequent k-n-match has the limitation of measuring distance in a non-metric fashion. Secondly, it incorporates the range of n values only at retrieval time, and hence cannot be used as a distance measure for clustering or classification. Third, the parameters n_0 and n_1 must be tuned for a given dataset to keep cost low. Also, reduction in cost is attained only if high accuracy can be achieved at low n_1.

3 Partial Match Function

In this section, we propose Partial Match Function (PMF). We describe PMF in Sect. 3.1. In Sect. 3.2, we focus on how indexing can be performed for PMF.

3.1 PMF

Let us first see how PMF is computed. We will then discuss how PMF is able to capture partial similarities.

Computing PMF. We wish to compute the distance between two D-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$; $\mathbf{x}, \mathbf{y} \in \Re^D$. Let $\mathbf{x_d}$ denote the d^{th} coordinate of $\mathbf{x}$. To simplify presentation, we assume that all coordinates of the vectors have been normalized to lie between 0 and 1.

The distance computation proceeds by transforming each real valued D-dim vector into a P-dim bit vector using a function $F_N : X \rightarrow B^P$.

DEFINITION 1. PMF Given two D-dimensional points $\mathbf{x}$ and $\mathbf{y}$, the PMF distance is given by the Hamming distance between $F_N(\mathbf{x})$ and $F_N(\mathbf{y})$.

Let $\overline{\mathbf{x}}$ denote the binary transformation of the vector $\mathbf{x}$. Thus, $\overline{\mathbf{x}} = F_N(\mathbf{x})$. The transformation $F_N(\mathbf{x})$ is performed using a 3-dimensional matrix $\mathbf{M}[D][H][N]$. $\mathbf{M}[D][H][N]$ is a matrix of randomly generated values drawn from between 0 and 1. D is the dimensionality of the vector x, and the resulting dimensionality of $\overline{\mathbf{x}}$ is controlled by H. This matrix is generated only once, before any vectors are transformed.

Let us now see how $\mathbf{M}[D][H][N]$ is used to create $\overline{\mathbf{x}}$. Each coordinate $\mathbf{x_d}$ of $\mathbf{x}$ is transformed into H independently computed bits, and all bits are concatenated to form a P dimensional bit-vector $\overline{\mathbf{x}}$, where $P = D \times H$. Each of the H bits $b_{dh}(\mathbf{x_d})$ corresponding to coordinate x_d are created by hashing together N bits $\{b_{dh1}(\mathbf{x_d}), \cdots, b_{dhn}(\mathbf{x_d}), \cdots, b_{dhN}(\mathbf{x_d})\}$. Only even values of N are used. Each bit $b_{dhn}(\mathbf{x_d})$ is generated based on one random value, $\mathbf{M}[d][h][n]$, from the matrix $\mathbf{M}$, as follows.

$$b_{dhn}(\mathbf{x_d}) = \mathbf{x_d} < \mathbf{M}[d][h][n] ? 0 : 1 \tag{1}$$

The bit is set to 0 or 1 depending on whether the value of $\mathbf{x_d}$ is less than or greater than the threshold in $\mathbf{M}[d][h][n]$.

The bit b_{dh} is then computed by an XOR operation on the N bits on which it is based:

$$b_{dh}(\mathbf{x_d}) = b_{dh1}(\mathbf{x_d}) \oplus \cdots b_{dhn}(\mathbf{x_d}) \oplus \cdots b_{dhN}(\mathbf{x_d}). \tag{2}$$

Once, the D-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ are transformed into $(D \times H)$-dimensional bit-vectors $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$, the PMF is computed as the Hamming distance between $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$.

$$\mathsf{PMF}(\mathbf{x}, \mathbf{y}) = \mathsf{HammingDistance}(F_N(\mathbf{x}), F_N(\mathbf{y})). \tag{3}$$

How PMF Works. Now let us explain how the PMF distance measure is able to capture partial similarities. First, consider the case when N is 1.

Case: $N = 1$

Here, each coordinate is represented by H bits, where each bit is 0 or 1 depending on whether it is less than or greater than a predefined random value. Consider the h^{th} bits generated for the d^{th} dimension of vectors $\mathbf{x}$ and $\mathbf{y}$, i.e., $b_{dh}(\mathbf{x_d})$, and $b_{dh}(\mathbf{y_d})$. Again, we assume that all coordinates have been normalized to lie between 0 and 1. That the probability that these two bits are different is given, by

$$Pr[b_{dh1}(\mathbf{x_d}) \neq b_{dh1}(\mathbf{y_d})] = |\mathbf{x_d} - \mathbf{y_d}|, \tag{4}$$

Since the probability of bits generated from the d^{th} dimension being different is proportional to the difference in the d^{th} dimension, the Hamming distance for H bits so generated would approximate the difference in the d^{th} dimension of the two vectors $\mathbf{x}$ and $\mathbf{y}$. By performing the same operation on each dimension of $\mathbf{x}$ and $\mathbf{y}$, the Hamming distance between $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ approximates the L_1 distance between $\mathbf{x}$ and $\mathbf{y}$.

Case: $N > 1$

Let us proceed to the case where $N > 1$. It can be shown that by performing an XOR operation each on sets of N bits generated from $\mathbf{x}$ and $\mathbf{y}$, the probability that the resulting bits for x and y are different, is given by

$$Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})] = 0.5(1 - (1 - 2 \times (|\mathbf{x_d} - \mathbf{y_d}|))^N). \tag{5}$$

To understand the behavior of these equations, let us look at the plot of how $Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})]$ varies with $|\mathbf{x_d} - \mathbf{y_d}|$, the difference in the corresponding coordinate values of $\mathbf{x}$ and $\mathbf{y}$. The plot is shown, for varying even N, in Fig. 1.

In these plots, for small $|\mathbf{x_d} - \mathbf{y_d}|$, $Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})]$ increases almost linearly with $|\mathbf{x_d} - \mathbf{y_d}|$. For larger $|\mathbf{x_d} - \mathbf{y_d}|$ upto 0.5, $Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})]$ stabilizes at a threshold. Note that for higher N, the $Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})]$ stabilizes sooner. For $|\mathbf{x_d} - \mathbf{y_d}| > 0.5$, (as seen in Fig. 1), $Pr[b_{dh}(\mathbf{x_d}) \neq b_{dh}(\mathbf{y_d})]$ becomes very small as $|\mathbf{x_d} - \mathbf{y_d}|$ increases.

Thus, small differences in attribute value are proportionately reflected in the overall PMF distance, while larger differences are thresholded to a maximum, thus limiting their effect. A very large difference in attribute value is minimized. This results in a distance that captures similarities and is not adversely affected by the dissimilar aspects, thus allowing a partial match.

We can see that the PMF distance measure is metric. This is because the distance is computed by computing the Hamming distance between two bitstrings, and Hamming distance is known to be metric. The metric property and triangle inequality being fulfilled are a significant advantage over other related distance measures.

3.2 PMF Index

For objects in a database, we wish to create an index that allows efficient retrieval using the PMF measure.

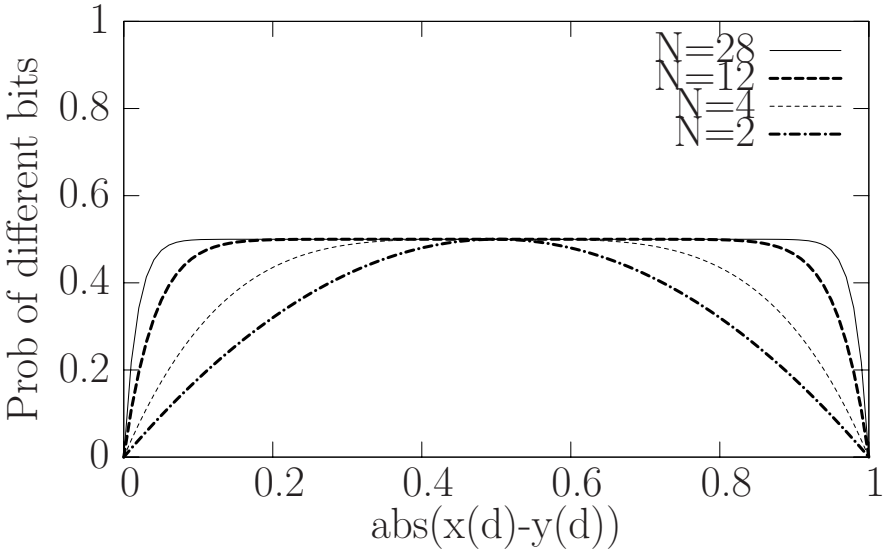


Fig. 1. How PMF works

Index Creation. To create the index, we compute the binary transformation of each object using the function F_N, and append it to a list. This list comprises the PMF index. Since the binary transformation utilizes only a few bits per coordinate, the size of the PMF index is only a fraction of the size of the database of object vectors, which use several bytes per coordinate.

PMF can then be readily computed by computing Hamming distance between the binary transformations stored in the index, allowing retrieval to proceed by simply performing a sequential scan on the compact PMF index.

Since the PMF index is much smaller than the database, it is more likely to fit entirely into main memory, as compared to the original data, or the index for Frequent k-n-match in [2] (which is as large as the original data). If the index is larger than available memory, it may simply be stored sequentially on disk, hence allowing efficient sequential access. The PMF index can be seen to be similar to compression based indexing approaches such as VA-file [5], since a much reduced representation of the database is used as index. However, the important distinction is that while most compression-based indexing methods require retrieving uncompressed versions of candidate objects for further evaluations, that is not true for the PMF index. The advantage of such an approach is that retrieval is not affected by the *curse of dimensionality*, and scales well as the number of dimensions becomes very high.

The cost of creation for PMF is of the order $O(D \times s)$, where s is the size of the database and D is the dimensionality. In contrast, for indexing with Frequent k-n-match [2], the cost of index creation is of the order $O(D \times s \times log\ s)$, since

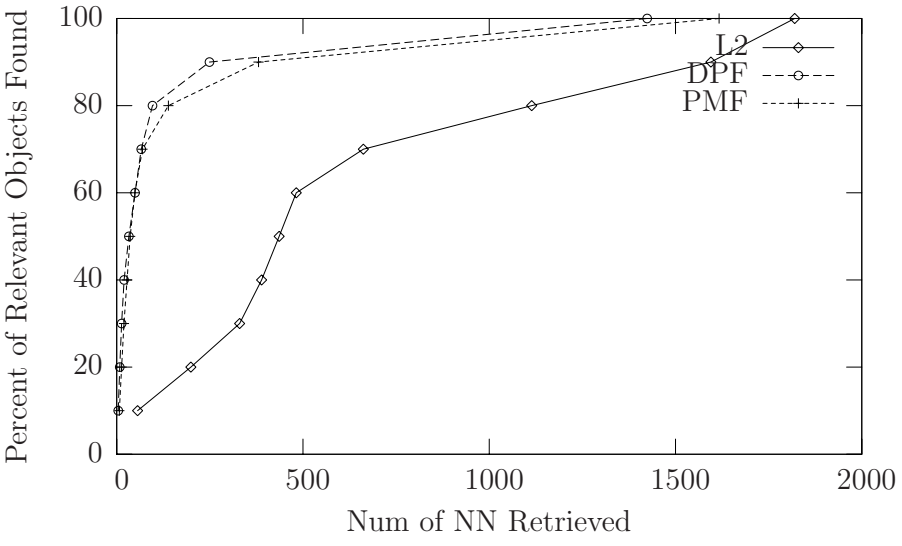


Fig. 2. PMF Effectiveness

it maintains separate sorted lists for each dimension. (In the case of the PMF index, no sorting or ordering of object representations is required.) In terms of space requirement, the size of the PMF index is given by $s \times D \times H$ bits.

Insertion and Deletion. The PMF index does not require the list of objects to be sorted or ordered in any fashion. This makes insertion and deletion straightforward. When an object is to be deleted from the database, its binary transformation can simply be removed from the PMF list, leaving an empty slot in its place. When an object is inserted, its binary transformation can be placed in an empty slot if available or appended to the end of the list, thus maintaining the sequentiality of the index.

In contrast, for FKNM [2], insertion and deletion is challenging, since indexing for FKNM is done by maintaining sequential sorted lists for each dimension. Maintaining sequentiality while inserting new objects would have to be done either by leaving empty slots in the list while creation, or by recreating the lists when new objects are inserted.

4 Empirical Study

In this section, we briefly present experimental results using PMF. To study the effectiveness of PMF, we used an *image replica* dataset based on the Corel image database The *image replica* dataset contains 2000 images divided equally into 50 classes. Each class consists of an image from the Corel image database and 39 near-replicas of the image, which were created as in [1]. The images are

represented by a 279-dimensional feature vector that includes texture and color features [1] extracted from the images.

The effectiveness of the PMF measure was studied and compared with DPF and L_2 (or Euclidean) distance. To measure effectiveness, we measured the average percentage of relevant objects (i.e. belonging to the same class) in the set of k retrieved NNs, for varying k. In the plots, first we notice that the performance of the L_2 distance is poor. We find that PMF performs significantly better than L_2, and as well as DPF. Additionally, PMF has the advantages of being metric, and easy to index.

5 Conclusion

In this paper, we proposed the PMF distance measure and demonstrated that it is a powerful method for measuring perceptual similarity. Experimental results showed that the PMF index can allow high retrieval performance. Besides achieving comparable performance as a distance measure, PMF has the following advantages over the competing methods for partial match.

1. *Metric property:* Since PMF works by measuring the Hamming distance in a binary vector space, the distances are metric. DPF and n-match (building block of Frequent k-n-match) do not fulfill the triangle inequality and are hence non-metric.
2. *Suitability for clustering and classification:* PMF is a metric distance measure that can be used for clustering and classification, besides retrieval. In contrast, Frequent k-n-match [2] is not a distance measure. It is a retrieval method that uses the n-match distance measure with varying n to retrieve a result set.
3. *Suitability to high-dimensional applications:* In emerging applications such as multimedia, dimensionality of objects can be very high, often in the hundreds. Since PMF uses a compression-based approach to indexing and does not evaluate any of the original vectors, it can scale to very high dimensions without degradation of performance.
4. *Efficiency:* The PMF index is much smaller than the database, which may make it possible to cache the entire index in memory. In a disk-based setting, sequential access is used, making retrieval efficient.

References

1. Qamra, A., Meng, Y., Chang, E.Y.: Enhanced perceptual distance functions and indexing for image replica recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 379–391 (2005)
2. Tung, A.K.H., Zhang, R., Koudas, N., Ooi, B.C.: Similarity searching: a matching based approach. In: VLDB (2006)
3. Goldstone, R.L.: Similarity, interactive activation, and mapping. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 3–28 (1994)

4. Goh, K., Li, B., Chang, E.Y.: Dyndex: a dynamic and non-metric space indexer. In: ACM International Conference on Multimedia (2002)
5. Weber, R., Schek, H.J., Blott, S.: A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In: VLDB (1998)
6. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional spaces. In: ICDT (2001)
7. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is nearest neighbors meaningful? In: ICDT (1999)
8. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: VLDB (2000)
9. Li, B., Chang, E., Wu, Y.: Discovery of a perceptual distance function for measuring image similarity. ACM Multimedia Journal Special Issue on Content-Based Image Retrieval 8(6), 512–522 (2003)

Searching One Billion Web Images by Content: Challenges and Opportunities

Zhiwei Li, Xing Xie, Lei Zhang, and Wei-Ying Ma

Web Search and Data Mining Group
Microsoft Research Asia
{zli, xingx, leizhang, wyma}@microsoft.com

Abstract. Although content-based image retrieval has been studied for decades, most commercial image search engines are still text-based. However, there is a growing demand for techniques to support content-based image search at Web scale. In this paper, we propose an ambitious goal to searching one billion Web images by content, and discuss the major challenges and opportunities. We also present several important applications that can be greatly benefited by techniques to enable Web-scale image search by content. These applications include image copyright infringement detection, street-side photo search, and search-based image annotation. We believe that the insights presented in the paper are enlightening to researchers in this field, and any breakthrough we make in this space will lead to many impactful applications in the future.

Keywords: Content based image retrieval, web search, multimedia mining.

1 Introduction

Content-based image retrieval (CBIR) has been studied for decades. Although researchers have obtained a lot of promising results, the use of CBIR technologies in real commercial system is still very limited. On the other hand, the number of images on the Web keeps growing rapidly. It is reported that there are more than one million new images uploaded to Flickr [1] every day, and most commercial search engines have indexed over several billions of Web images based on the surrounding text. It seems that keyword-based image search engine is already sufficient for serving most people's information need. However, we believe that if billions of online images can be well utilized, it is possible to develop a more powerful search engine and lead to new and promising applications.

Most existing CBIR systems suffer from scalability problem and cannot scale to billions of photos. It is difficult to build effective index for high dimensional image features. Motivated by the success of web search engines, many researchers have tried to map image retrieval problems to text retrieval problems, hoping that the proven effective indexing and ranking schemes can be used to handle the scale. The basic idea is to map image features to words. Typically images are first represented by local features, and then by clustering, each local feature is mapped to a discrete keyword. Such image representation is called "bag of features," similar to "bag of words" for

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 33 – 36, 2007.
© Springer-Verlag Berlin Heidelberg 2007

document representation. With this representation, comparing two images become matching words in them, and therefore, a text-based search engine can be utilized to reduce the computational and memory cost.

To develop this type of image search engines, there are still many technical challenges and problems that we need to address:

- **Vocabulary**: what kinds of image features should be used? How to map them to words? The most generally utilized method is clustering. Some researchers also adopted a hierarchical clustering method to generate a vocabulary tree. But it is clear that we need to develop some kinds of visual language models to solve the problem.
- **Long query**: The reason why text search engine is effective is because text queries usually only contain a few words. So, the query-document matching can be conducted efficiently by inverted index. Although images can be represented by "bag of features," the retrieval problem is still very different from text retrieval because query-by-example is actually equivalent to using a whole document as a query. So, the search is more like document-to-document matching. How can we deal with this kind of "long query" effectively?
- **Content quality**: Web search engine is effective because it can use link analysis to obtain quality and importance measurement (e.g. PageRank) for Web pages. For images, it is hard to obtain similar kind of measurement because the links are typically not directly associated with images. Without PageRank for images, we won't be able to take advantage of many top-k search techniques typically used in web search, and it also will lead to the lack of efficient cache of index.
- **Relevance ranking**: The similarity measure between two images is quite different from text. How image words are weighted in computing the relevance. And how to deal with "word proximity" in images?
- **Distributed computing for Web-scale multimedia analysis**: Because of the large volume of image data we need to process and index, the system has to be a distributed system, consisting of hundreds of powerful servers. It is inevitably to confront with the challenges as in text-based search engines, such as fault tolerance, data redundant backup, auto configuration, etc.

To be able to response to a query within one second, the system has to employ a very efficient indexing solution, which is probably similar to inverted lists used in text-based search engines. The use of the indexing solution will make many algorithms depending on sequential scan of the whole database impractical. Therefore, most existing CBIR algorithms need to be re-evaluated on a Web-scale content-based image search system.

Due to the restrictions imposed by the indexing solution, the user interface for relevance feedback needs to be restudied. For example, it is not trivial to refine the search result by query point movement or distance function modification based on relevance feedbacks in an inverted index-like system. How to leverage users' feedback either explicitly or implicitly will be an interesting research problem in a Web-scale content-based image search system.

It is not a trivial task to extract low level features from every image in a database containing one billion images. We need a flexible platform and infrastructure to provide large-scale data management and data processing capabilities. The infrastructure should facilitate the extraction and experimentation of various features and similarity measures for image search, so that it can help researchers and engineers to find a best practical solution by carefully evaluate the capabilities and limitations of different features and algorithms.

2 Applications

There are many important applications that can be benefited by the technology if we can effectively scale it up to search one billion image by content. In the following, we introduce a few related projects at Microsoft Research Asia.

2.1 Copyright Infringement Detection

Images are also one kind of intelligence property (IP) of their creators. However, with more and more user-created content on the Web, people may utilize other people's copyright images and videos without authorization. Thus, there is a need to develop a technique to help image owners to detect whether their images are used on the Web without their authorization or permission. Given a suspicious image, our task is to find the most similar image (or near-duplicate image) using our image search engine that indexes billions of Web images. Note that pirated images may be edited slightly, and the amount of images we have to match is very huge. Detecting such copyright infringement in a large image data collection is a non-trivial task.

2.2 Street-Side Photo Search

Mobile phones with embedded cameras are becoming popular nowadays and have huge growth potentials. Most current services for information acquisition on mobile devices are using text-based inputs. Nevertheless, sometimes it is difficult for users to describe their information needs in words. Instead of current flat query modes, camera phones can support much richer queries, not only text but also images. Therefore, it is important to develop a mobile search service that allows users to search for relevant information on the Web via the pictures taken on a mobile phone.

2.3 Search-Based Image Annotation

Although it has been studied for several years by computer vision and machine learning communities, image annotation is still far from practical. One reason is that it is still unclear how to model the semantic concepts effectively and efficiently. The other reason is the lack of training data, and hence the semantic gap cannot be effectively bridged.

With the growing number of images on the Web, it is possible to use the search technology to annotate unlabeled images. We can reformulate the image annotation problem as a novel two-step fashion: searching for semantically and visually similar images on the Web, and mining key phrases extracted from the descriptions of the

images [2],[3]. Intuitively, if a well annotated and unlimited-scale image database is available, then for any query image, we can find its near-duplicate in this database and simply propagate its annotation to the query image. In a more realistic case that the image database is of limited scale, we can still find a group of very similar images in terms of either global features or local features, extract salient phrases from their descriptions, and select the most salient ones to annotate the query image.

3 Summary

Searching one billion images by content is an ambitious and challenging direction. In this position paper, we have analyzed the major difficulties in indexing and searching a Web-scale image database. The challenges include how to construct a representative visual vocabulary, how to efficiently find relevant images for a long query, and how to compute "PageRank" for images for cache design and quality improvement. These challenges also motivated us to develop a flexible platform and infrastructure for "scale" experiment with various algorithms [4]. We also present several important applications related to Web-scale image search by content. We believe that the insights in this paper are enlightening to researchers in this field, and the large scale content-based image search will lead to many impactful applications in the future.

References

1. http://blog.flickr.com/flickrblog/2006/08/geotagging_one_.html Geotagging - one day later (August 29, 2006)
2. Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.Y.: Image Annotation by Large-Scale Content-based Image Retrieval, In: Proc. of ACM Int. Conf. on Multimedia, Santa Barbara, USA (2006)
3. Wang, X., Zhang, L., Jing, F., Ma, W.Y.: AnnoSearch: Image Auto-Annotation by Search. In: Proc of the International Conference on Computer Vision and Pattern Recognition (CVPR), New York (2006)
4. Wen, J.R., Ma, W.Y.: WebStudio - Building Infrastructure for Web Data Management, ACM SIGMOD (2007)

Challenges on Peer-to-Peer Live Media Streaming

Wen Gao and Longshe Huo

Institute of Digital Media,
Peking University, China
{wgao, lshuo}@pku.edu.cn
http://idm.pku.edu.cn

Abstract. Recently there has been significant interest in the use of peer-to-peer technologies for deploying large-scale live media streaming systems over the Internet. In this position paper, we first give a brief survey on the state-of-the-art of peer-to-peer streaming technologies, and then summarize and discuss some major challenges and opportunities in this research area for future directions.

Keywords: Peer-to-Peer, media streaming.

1 Introduction

During the past decade, the rapid development of the Internet has changed the conventional ways that people access and consume information. Besides traditional data services, various new multimedia contents would also be delivered over the same IP network, among which the live media streaming service will play a more important role. The typical applications of live media streaming include Internet television, distance education, sports events broadcasting, online games, and etc.

Recently, Peer-to-Peer (P2P) has emerged as a promising technique for deploying large-scale live media streaming systems over the Internet, which represents the paradigm shift from conventional networking applications. In a P2P system, peers communicate directly with each other for the sharing and exchanging of data as well as other resources such as storage and CPU capacity, each peer acts both as a client who consumes resources from other peers, and also as a server who provides service for others. Compared with traditional streaming techniques such IP multicast and CDN (content delivery networks), P2P based streaming system has the advantages of requiring no dedicated infrastructure and being able to self-scale as the resources of the network increase with the number of users.

A great number of systems for P2P based live media streaming have been proposed and developed in recent years. Merely in China, nowadays there are about more than a dozen of P2P streaming applications deployed in the Internet. In this position paper, we first give a brief survey on the state-of-the-art of P2P streaming technologies, and then summarize and indicate some major challenges and opportunities in this research area for future directions.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 37 – 41, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 State-of-the-Art of P2P Streaming Technologies

In typical P2P streaming implementations, media data are distributed along a application-layer logical overlay network constructed over the underlying physical IP network. To construct and maintain an efficient overlay network, mainly three questions should be answered. The first relates to the P2P network architecture, i.e., what topologies should the overlay network be constructed? The second concerns routing and scheduling of media data, i.e., once the overlay topology is determined, how to find and select appropriate upstream peers from which the current peer receives the needed media data? The third is membership management, i.e., how to manage and adapt the unpredictable behaviors of peer joining and departure?

Recently, several P2P streaming systems and algorithms have been proposed to address the above issues. From the view of network topology, current systems can be classified into three categories approximately: tree-based topology, forest-based topology, and mesh topology. Following we outline a brief survey on P2P streaming techniques according to this classification.

(1) Tree-based topology

PeerCast [1] is a typical example of tree-based P2P streaming systems. In PeerCast, peers are organized into multicast trees for delivering data, with each data packet being disseminated using the same structure. When a peer receives a data packet, it also forwards copies of the packet to each of its children. Since all data packets follow this structure, it becomes critical to ensure the structure is optimized to offer good performance to all receivers.

Generally, there exist four route selection strategies in tree-based topology: random selection, round-robin selection, smart selection according to physical placement, and smart selection according to bandwidth. To achieve a balanced multicast tree, custom routing policy should be chosen carefully for individual peer node.

(2) Forest-based topology

Conventional tree-based multicast is inherently not well matched to a cooperative environment. The reason is that in any multicast tree, the burden of duplicating and forwarding multicast traffic is carried by the small subset of the peers that are interior nodes in the tree. Most of the peers are leaf nodes and contribute no resources. This conflicts with the expectation that all peers should share the forwarding load.

To address this problem, forest-based architecture is beneficial, which constructs a forest of multicast trees that distributes the forwarding load subject to the bandwidth constraints of the participating nodes in a decentralized, scalable, efficient and self-organizing manner. A typical model of forest-based P2P streaming system is SplitStream [2]. The key idea of SplitStream is to split the original media data into several stripes, and multicast each stripe using a separate tree. Peers join as many trees as there are stripes they wish to receive, and they specify an upper bound on the number of stripes that they are willing to forward. The challenge is to construct this forest of multicast trees such that an interior node in one tree is a leaf node in all the remaining trees and the bandwidth constraints specified by the nodes are satisfied. This ensures that the forwarding load can be spread across all participating peers.

(3) Mesh topology

In conventional tree-based P2P streaming architectures, at the same time a peer can only receive data from a single upstream sender. Due to the dynamics and heterogeneity of network bandwidths, a single peer sender may not be able to contribute full streaming bandwidth to a peer receiver. This may cause serious performance problems for media decoding and rendering, since the received media frames in some end users may be incomplete.

In forest-based systems, each peer can join many different multicast trees, and receive data from different upstream senders. However, for a given stripe of a media stream, a peer can only receive the data of this stripe from a single sender, thus results in the same problem like the case of single tree.

Multi-sender scheme is more efficient to overcome these problems. In this scheme, at the same time a peer can select and receive data from a different set of senders, each contributing a portion of the streaming bandwidth. In addition, members of the sender set may change dynamically due to their unpredictable online/offline statuses. Since the data flow has not a fixed pattern, every peer can send and also receive data from each other, thus the topology of data plane likes mesh. The main challenges of mesh topology are how to select the proper set of senders and how to cooperate and schedule the data sending of different senders.

Recently, DONet [3] implemented a multi-sender model by introducing a simple and straightforward data-driven design, which does not maintain an even more complex structure. The core of DONet is the data-centric design of streaming overlay, and the Gossip-based message distribution algorithm. In DONet, a node always forwards data to others that are expecting the data. In other words, it is the availability of data but not a specific overlay structure that guides the flow directions. Such a design is suitable for overlay with high dynamic nodes. Experiments showed that, compared with a tree-based overlay, DONet can achieve much more continuous streaming with comparable delay.

3 Technical Challenges and Opportunities

Though some successes have been made in recent years, especially with the introducing of mesh-based approaches, there are still challenging problems and open issues need to be overcome in P2P live media streaming.

The main problem results from the heterogeneity of the underlying IP networks. There exist mainly two types of heterogeneities in the current Internet: heterogeneous receivers and asymmetric access bandwidths. In a P2P based live media streaming system, for each individual peer the receiving capability is decided by its downlink bandwidth, however for the whole system the total available bandwidth is decided by the sum of the uplink bandwidths of all the participated peers. Under this situation, same and perfect QoS is hard to be guaranteed for all of the participated peers. For example, if the access bandwidth of a peer is less than the average bit rate of the media stream it requires, or the sum of the uplink bandwidths of all upstream peers who provide data for this peer is less than the average bit rate, then random packet-losses may occur ether during the network or at the buffer of upstream peers. This may lead

to incorrect decoding at the client side even partial data have been received, which means not only the waste of bandwidth resources, but also degraded media reconstruction qualities.

The most hopeful solution to this problem is to provide self-adaptive QoS for each individual peer according to the current network conditions, at the same time the total available uplink bandwidths of all peers are utilized as full as possible. To satisfy this objective, three main issues should be addressed:

(1) Content aware media data organization

Current P2P streaming systems focus mainly on network topology and protocol design, but pay rare attention to the media contents carried over the network. In fact, since streaming media have their distinct characteristics from normal data file, good performance can be achieved only when both the characteristics of media coding and networking are considered together perfectly. While scalable coding techniques hold promise for providing network adaptive media transmission, they are yet to be deployed in today's mainstream media codec. A promising solution is to partition the current non-scalable coded media data based on content analysis, and reorganize them into another form with scalable capability to some extent, so that selective and priority-based schedule strategies can be used while transmission.

(2) Priority-based media data delivery mechanism

For the quasi-scalable media data prepared above, efficient transmission and control mechanisms should be invented to guarantee that the minimal decodable media units (for example, a video frame or slice) can be transmitted to the receiver in a restrict order based on their priorities. This implies that every data unit received by a peer at any time is intact and decodable independent of any still un-received media data. By this way, no waste of bandwidth is involved and free-error and fluent media experience can be obtained even in the case of worse network conditions.

(3) QoS adaptive multi-source and layered media data schedule algorithm

Based on the above content aware data organization and priority-based delivery mechanism, efficient data schedule algorithms are needed to retrieve data from multiple senders in order to maximize the overall bandwidth utilization of the whole network and minimize the average media reconstruction distortion of all users. Compared with conventional P2P streaming systems which simply partition a streaming media into a series of data blocks and schedule each data block as the minimal transmitting unit, the scheduling model of this system and its solutions are more complicate to establish and resolve.

In another paper of this workshop [4], we present a novel data organizing and delivery framework for P2P live media streaming, which takes into account both the characteristics of media coding and P2P networking, and aims to provide self-adaptive QoS for different users of the Internet. We hope that it can become a good start for inspiring more and more researches on beneficial solution to the above problems.

4 Conclusion

Recently, P2P streaming has attracted a lot of attentions from both academy and industry. Various P2P media streaming algorithms have been studied, and the systems have been developed. However, a number of key technical challenges still need to be

overcome in order to maximize the whole network resource utilizations. Opportunities are always along with challenges. We still have a long way to go in the research area of P2P based live media streaming systems.

References

1. Deshpande, H., Bawa, M., Garcia-Molina, H.: Streaming live media over a Peer-to-Peer network, Stanford database group technical report (2001-20) (August 2001)
2. Castro, M., Druschel, P., Kermarrec, A.-M., Nandi, A., Rowstron, A., Singh, A.: Split-Stream: High-bandwidth content distribution in a cooperative environment, In: Proc. the International Workshop on Peer-to-Peer Systems, Berkeley, CA (February 2003)
3. Zhang, X., Liu, J., Li, B., Yum, T.: Coolstreaming/DONet: A data-driven overlay network for live media streaming. Proc. IEEE INFOCOM '05 (2005)
4. Huo, L., Gao, W., Fu, Q., Guo, R., Chen, S.: QoS adaptive data organizing and delivery framework for p2p media streaming. In: Proc. MCAM'07 (June 2007)

Video Streaming to Mobile Handheld Devices: Challenges in Decoding, Adaptation, and Browsing

Chang Wen Chen[1] and Houqiang Li[2]

[1] Department of Electrical & Computer Eng., Florida Institute of Technology,
150 W. University Blvd., Melbourne, FL 32901 USA
[2] Dept. of EEIS, University of Science and Technology of China,
Hefei, 230027, P.R. China
cchen@fit.edu, lihq@ustc.edu.cn

Abstract. Growing popularity and richer functionality of contemporary mobile handheld devices such as PDAs and smart phones have enabled emerging video streaming applications to these devices via various wireless networks. However, these handheld devices are significantly constrained by limited resources available to them for some sophisticated operations. Among these limitations, the available processing power has limited these devices to perform full scale decoding of the video compressed with H.264 standard [1], [2]. The typically small display size of the handheld devices has been imposing significant barriers for the devices users to enjoy high-resolution videos. This paper presents some challenges in design decoding algorithm for smart phones, in video adaptation to suit heterogeneous devises, in browsing videos via handheld devices by diverse mobile device users. Several approaches will be proposed to meet these challenges and to offer the mobile device users an enhanced experience in these emerging mobile video applications.

Keywords: H.264, video streaming, mobile devices, handheld devices, video decoding, video adaptation, video browsing, visual attention modeling.

1 Introduction

Recent advances in wireless communication and networking have enabled an entire new array of applications in multimedia over wireless networks beyond traditional applications in voice and text. Among these real-time multimedia applications, video streaming to mobile handheld devices has become feasible under various wireless environments. In particular, with the growing popularity of mobile devices, users can enjoy videos in almost anyplace and at anytime. For video streaming applications in the wired networking environment, the videos stored in the video server are generally encoded at high-resolution (HR) and high bitrate streaming is generally able to guarantee users' access and browsing experiences.

There are several fundamental difficulties in enjoying these videos with mobile devices over wireless networks. The first critical constraint is the necessity of the low power operations for all the mobile handheld devices driven by the battery. These battery driven devices have limited processing power and memory size that generally

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 42 – 51, 2007.
© Springer-Verlag Berlin Heidelberg 2007

prevent the implementation of computationally complex algorithms. The second critical constraint is that the wireless links for mobile handheld devices to access the video data are usually band-limited and error prone. Transmission of compressed video over such links presents significant challenges. The third essential constraint is the limited display size of mobile devices, which often hinders the users to fully enjoy the video scene. It is very much desired for mobile users to access videos with current display size but with an enhanced viewing experience.

In this paper, we will present some challenges to overcome these fundamental constraints in order for the users of mobile handheld devices to enjoy a true universal multimedia access (UMA) experience under the pervasive media environment. UMA refers to the user experiences in which access and consumption of multimedia content over heterogeneous networks have been accomplished using diverse terminals in a seamless and transparent way [3]. We will then address the technical issues aiming at overcoming some fundamental constraints in video streaming to mobile devices. In particular, we will discuss the decoding of the compressed video on the smart phone, adaptation of video streams at the edge servers for heterogeneous mobile devices, and the browsing of streaming video by a diverse class of mobile device users.

The rest of this paper is organized as follows. Section 2 presents in some detail the challenges that stem from various physical constraints of the wireless networks or mobile devices. Section 3 introduces several key technologies that have recently been developed to address the challenges in order to provide enhanced UMA experiences to mobile device users. These technologies include video decoding, video adaptation, and video browsing for streaming to mobile devices. Section 4 concludes this paper with a summary and a discussion of future technology trends.

2 Challenges in Video Streaming to Mobile Handheld Devices

Mobile handheld devices are inherently resource constrained because they need to be portable and powered by battery. The first challenge for video streaming to the mobile devices is to design a suitable video encoding and decoding algorithm for mobile video applications. In general, portable, battery operated mobile devices pose the following particular constraints [4] that need to be carefully taken care of when implementing a video coding scheme:

1) Power (battery life)
2) Memory size
3) Cache size/memory bandwidth
4) CPU speed
5) Display Resolution
6) Bandwidth

As a result, we believe that the desirable attributes of a video coding scheme developed for mobile device applications should include:

1) High coding efficiency to meet the limited bandwidth constraints.
2) Balance of coding efficiency and complexity of implementation geared towards inexpensive consumer-level appliances.

3) Low cost implementation of decoder, considering the limited CPU power and the limited cache/memory.
4) Massive deployment in vast numbers of consumer-level mobile receivers
5) Coding tools included should provide significant coding gain relative to the cost for their decoder implementation.

Considering the current video streaming applications, the first step towards UMA through mobile devices is to enable seamless implementation of state-of-the-art video decoder on the mobile devices so as to allow the users to decode and display the received video bitstreams. This is particularly challenging when the cellular phone is used as the mobile device for video streaming and display. This is because the capacity left to implement video streaming is quite limited after the essential services of voice calling are implemented on the phone.

The second challenge for video streaming to the mobile devices is to design some video adaptation schemes in order to meet the requirements of seamless services for an array of heterogeneous mobile devices having diverse functionality and capability differences. In particular, the display size of these devices can be quite different and the display resolution of these devices is usually much smaller than the resolution of original video. Video adaptation with reduced spatial resolution is therefore necessary in order to stream the appropriate amount of bitstreams to the mobile devices whose link capacity is constrained by the bandwidth of wireless access networks.

In recent years, some video transcoding schemes [5], [6] have been developed to achieve the spatial resolution reduction. These traditional spatial transcoding schemes are usually based on spatial downsizing by the factor of two or power of two. Considering perceptual results, excessive resolution reduction will cause significant loss of desired information. More recently, some researchers proposed an automatic modeling approach to crop a most informative region within the original image and assign the cropped region a proper scaling ratio according to a certain display size [7]. Similarly, an attention based video adaptation solution [8] was developed for better perceptual results in small displays. However, these schemes adopt the pixel-domain approach that the original videos are decoded and then fully encoded after processing. This is often very costly in terms of computational complexity and is not suitable for mobile devices in practice.

Traditionally, the video adaptation operations are implemented at the proxy server linking between the wired networks and the mobile devices. One such example will be the access point for wireless local area networks (WLAN). When the access point receives the spatial resolution and other parameters from the mobile devices, these parameters will be utilized to design an appropriate adaptation scheme for this particular device. Every time a different request may be received from heterogeneous mobile devices diverse, the video adaptation scheme will needed re-designed. This creates significant burden at the proxy server such as access point in the WLAN which may become the bottleneck in implementing the UMA environments [3].

This technical issue constitutes the third challenge for video streaming to mobile devices. It is the best interest of any UMA service provider to shift the computational burden from proxy server to the original server which in general enjoys significantly more resources than the proxy servers. The shift of some key video adaptation operations from proxy server to original video server can be made possible as long as the

desired video adaptation for heterogeneous mobile devices can be appropriately parameterized and these parameters can be embedded into the compressed video bitstreams to be utilized by the proxy servers.

3 Key Technologies for Seamless Video Streaming to Mobiles

In the following, we shall discuss several key technologies that have been developed recently to address the technical issues in providing seamless video streaming services to mobile handheld devices. We will start with the most essential operations for such service in terms of reduced complexity video decoding on the mobile devices. We will then examine the video adaptation schemes that are necessary for streaming reduced spatial resolution video for small display size of the mobile devices. Finally, we will present an innovative approach that is able to shift the burden of extensive video adaptations at the proxy server to the original server to allow light operations at the proxy servers so as to enable a seamless video streaming from the server to proxy and to the mobile users.

3.1 Reduced Complexity Video Decoding for Mobile Devices

We consider implementing the video coding scheme on the mobiles to be compliance with the current H.264/AVC standard [1], [2]. H.264/AVC is the powerful and state-of-the-art video compression standard that has recently been developed by the ITU-T/ISO/IEC Joint Video Team consisting of experts from ITU-T's Video Coding Expert Group (VCEG) and ISO/IEC's Moving Picture Experts Group (MPEG) [10]. The H.264/AVC standard represents a delicate balance between coding gain, implementation complexity, and costs based on state of VLSI design technology.

To develop a reduced complexity video coding scheme based on H.264/AVC standard, we need to investigate each component of the video encoding and decoding in terms of its contribution to coding efficiency against its induced implementation complexity. Since this reduced complexity video decoding scheme is designed for local playback on mobile device, we focus on those coding modules that are adopted by both encoder and decoder. This is because these modules have greatest impact on both coding efficiency at the encoder and the implementation complexity at the decoder. However, when a module is included only at the encoder, we shall adopt a high coding efficiency coding technique even though the module is high in implementation complexity. Such coding module will not impact the mobile devices since these modules will not be implemented at the decoder.

The implementation has been focused on several key modules in video coding. One important common feature of these implementations is that the key components have been enforced to be compliance with the H.264/AVC standard. We shall use Nokia 6600 as an example to illustrate the principles behind the design. However, we the principles of implementation enable the potential widespread adoption of proposed scheme on other mobile devices from different manufacturers.

We have considered the following key modules for the potential reduction of implementation complexity:

1) Inter Mode
2) Intra Mode
3) Quarter-sample-accurate motion compensation
4) Rate Distortion Optimization
5) Transform and quantization [11]
6) Interpolation
7) Entropy coding [12]
8) In-the-loop deblocking filter [13]
9) Frame drop

For each of these modules, we investigate how the reduction of complexity can be accomplished with minimum loss of coding efficiency. When certain reduction causes substantial loss in coding efficiency, such reduction will not be carried out. The following table lists the modules we considered with a comparison between the H.264 Reference Software JM76 [14] and the proposed scheme. Notice that many modules remain unchanged because the gain in complexity reduction is out weighted by the loss in coding efficiency. On average, the loss of coding efficiency is under 0.5dB in PSNR caused by complexity reduction. However, the complexity reduced codec can now be implemented on the Nokia 6600 for decoding H.264 video bitstreams. The following table summarizes the results and implementation details can be found in [15].

Module	H.264 Ref. SW JM76	The Proposed Scheme
Inter Mode	seven	seven
Intra Mode	nine for Intra4x4 and four for macroblock	five for Intra4x4
Motion Vector	quarter-accuracy	quarter-accuracy
RDO	enabled	enabled
Transform	4x4 integer transform,	4x4 integer transform
Quantization	52 steps	52 steps
Interpolation	6-tap for half-pix and bilinear for quarter-pix	4-tap for half-pix and bilinear for quarter-pix
Entropy Coding	CAVLC	Exp-Golomb
Loop filter	enabled	enabled
Reference frame	two	two

3.2 Attention-Based Video Adaptation for Mobile Devices

As we indicated in Section 1, there have been numerous research activities in spatial resolution reduction for displaying video at mobile devices with small display size. Majority of traditional simple transcoding techniques have been on spatial downsizing by the factor of two or power of two [5], [6]. Considering the perceptual results, excessive resolution reduction will cause significant loss of detailed information. Therefore, for video streaming to mobile devices, a better way to solve this problem is to identify the most informative regions within the original video frame and crop an appropriate portion of the video frame containing the informative region for the small display at the mobile devices.

Once again, we consider the H.264/AVC standard as the default video coding scheme for these applications because of its efficient coding performance and its widespread adoption by the applications community. A typical approach for video adaptation would be to generate a new bitstream with the direct method. In this case, the scheme re-encodes each new frame of the sequence after full decoding of the received data. However, full-scale motion estimation and full modes decision will result in unaccepted complexity at the mobile devices. Therefore, we need to develop several techniques in order to reduce encoding complexity, including techniques in motion reuse and simple mode decision [16], [17].

One major issue needs to be resolved is how to identify the informative region within a video frame for the generation of reduced resolution video streaming and display at the mobile devices. The approach we developed is to determine the most informative region in the decoded frames by an attention based modeling. As shown in Definition 1, a set of information carriers – attention objects (AOs) [7] are defined in our method.

$$Definition1: \{AO_i\} = \{(SR_i, AV_i, MPS_i)\}, 1 \le i \le N \quad (1)$$

Each AO owns three attributes: SR, AV and MPS. SR is referred as a spatial region corresponding to an AO. The attention value (AV) indicates the weight of each AO in contribution to the information contained in the image. Since the delivery of information is significantly dependent on the dimension of presentation, minimal perceptible size (MPS) is introduced as an approximate threshold to avoid excessively subsampling during the reduction of display size.

Accordingly, three attributes of AOs will be measured by an automatic modeling method. We have taken into account four types of attention objects in the proposed model: motion objects, face objects, text objects and saliency objects. Since we study video streaming, the modeling is different from that of the static pictures [8] in that motion information with in a video frame carries significant information. For H.264 video format, the Motion Vector Field (MVF) of a frame is able to measure the motion information well.

After we extract the attention areas from the video frames, we are faced another technical challenge in generating the reduced resolution smooth video frames from the extracted attention areas. A straightforward approach is to directly present the focus regions to the user. However, this naïve approach will cause jitters in the video sequences since such frames will be discontinuous after cropping. We notice that in some tracking systems [18], virtual camera control is introduced to smooth the tracking process. Therefore, we assume that there is also a virtual video camera steered to pan and zoom in the video frames. Two types of focuses are introduced to implement virtual camera control: Camera Focus (CF) and Target Focus (TF). Camera Focus represents the focus displayed to the users and Target Focus represents the destination focus that can either be assigned manually or determined automatically. Corresponding display ratios of the two types of focus regions are defined as Camera Ratio (CR) and Target Ratio (TR), respectively. The Euclidean distance between CF and TF is denoted as Δd and the difference of CR and TR is defined as Δr. The direct focus shifting is substituted for a smooth following process from the current focus region to the target focus region, with a set of pan and zoom operations. As a result of a

sequence of fixation, following, and shifting operations, the output of adjusted cropped regions will be a smooth sequence. Such a rectangle region after adjustment will be noted as the attention area in video adaptation. The detailed implementation on attention information-based video adaptation for mobile devices can be found in [19]. Yet, there are still open issues in video adaptation [20].

3.3 Embedded Intelligent Video Adaptation for Browsing Via Mobiles

Although the attention area extraction-based video adaptation is able to generate reduced resolution video for mobile devices with small display, this region-of-interest (ROI) based transcoding at proxy has two critical shortcomings that need to be overcome to maximize mobile video browsing experiences. The first shortcoming of the ROI-based transcoding is the need to perform the detection of four types of attention areas separately in order to obtain a robust ROI within a given video. The computational operations to perform these detections and to combine the detection results shall become a significant burden for either edge server or proxy. The second shortcoming of the ROI-based transcoding is the need to perform ROI detection for heterogeneous users every time a user requests the video browsing service. Such repeated operations will sometimes overwhelm the proxy server.

These shortcomings can be overcome if the compressed video at the original server can be augmented with ROI information for proxy server to access. If we are able to embed the ROI information into the bitstream of compressed video to be extracted by the proxy server, the burden of the computational complexity can be shifted from the transcoding at proxy server to encoding at the original server. It is this analysis of shifting the operations from proxy server to original server that motivates us to design an attention-information-based video adaptation for browsing video via mobile devices.

This intelligent design of video spatial adaptation is based on the assumption that the videos stored in the original video server are usually off-line generated and the computational complexity is not an issue with off-line operations. Furthermore, we also assume that the attention objects in each video frame may remain the same even for different mobile users. This is because the attention model we adopted is quite generic for a wide variety of users. If we are able to shift the attention detection operations from the transcoding process to the encoding process, then, we will be able to shift the complexity from proxy to the video server.

Based on the above analysis, we have recently developed an intelligent spatial adaptation framework as shown in Figure 1. This framework has two parts: video content generation and video adaptation operation. During the generation of compressed video bitstreams, the attention information will be detected from the video simultaneously. Then the bitstreams and the attention information will be stored together in the video server. When the server offers services for mobile devices users, the adaptation system placed on a proxy or server will perform adaptation manipulation on the original high resolution video by making use of the attention information in the video to meet the display constraint of the mobile devices.

The separation of attention information detection and adaptation operation has two benefits. First, since the detection process need be performed only once and the detected attention information can be used for heterogeneous users, the workload of this new adaptation system will be reduced greatly while the system still remains

flexible. This intelligent adaptation system only needs to complete the function in transforming high resolution videos into low resolution videos for mobile devices. This will greatly facilitates the potential implementation of real-time adaptation process for mobile devices. The second benefit of the proposed scheme is that we can actually improve the video adaptation performance by fully utilizing the pre-detected attention information.

Since original sequences are unavailable in the transcoding or video adaptation process, reconstructed frames after decoding have to be used as the input video to transcoding. The better quality the reconstructed original video frames, the higher performance the transcoder will be able to achieve. In browsing video via mobile devices, the new video sequence generated by the transcoding or video adaptation suitable for the target mobile device will consist of mostly attention areas in each frame. Therefore, if we can adjust the bit allocation strategy in the original encoding for attention and non-attention areas, we shall be able to improve the quality of low resolution adapted videos. That is to say, if it is known to us that the majority of the clients are mobile device users, we can move some bits allocated for non-attention areas to attention areas when we encode the original video.

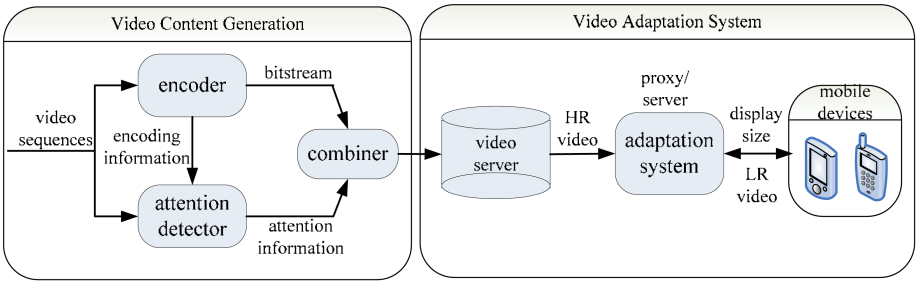


Fig. 1. The structure of attention-information-based spatial adaptation framework

We have implemented several strategies in motion estimation and mode decision to achieve a high performance transcoding [21], [22], [23]. We tested the proposed intelligent video adaptation with several standard video sequences at a variety of bitrates in which the original CIF bitstreams are transcoded into QCIF bitstreams containing attention areas. Comparing with the simple video adaptation scheme we presented in Section 3.2, this embedded intelligent video adaptation is able to obtain R-D performance improvement at all bitrates and the improvement can be as high as 0.5dB. The details of this intelligent video browsing scheme can be found in [24].

4 Summary and Future Technology Trends

In this paper, we have presented several fundamental constraints in order for the users of mobile handheld devices to enjoy true universal multimedia access experiences under the pervasive media environment. These fundamental constraints are originated from the inherent physical limitations of either wireless access links or mobile device

operations. We then addressed several technical issues aiming at overcoming some fundamental constraints in order to provide seamless video streaming to mobile devices. In particular, we discussed the decoding of the compressed video on the smart phone, adaptation of video streams at the edge servers for heterogeneous mobile devices, and the browsing of streaming video by a diverse class of mobile device users via the wireless access links.

One recent emerging mobile video application is the mobile Internet TV (IPTV) which has attracted significant attention from both research community and service providers. In this application, the consumers are expecting seamless streaming of broadcasting video contents to the mobile receiving devices. Although the content creators may produce some special programs targeting only mobile reception devices with low spatial resolution, many broadcasting video programs need to be adapted from high resolution originals produced for high end users to low resolution videos for reception by mobile devices. In this case, all three technical approaches presented in this paper are applicable to mobile IPTV applications.

Another future technology trend is in the application of video adaptation in peer-to-peer (P2P) video streaming to and from mobile devices. Video adaptation becomes necessary when the video data stored in your peers have significant resolution mismatch because simple streaming of original data cannot be implemented. In this case, we will need to integrate the video adaptation with protocol designs for the seamless P2P streaming of video data to and from mobile device peers.

Finally, the visual attention-based video adaptation has been based on the basic assumption that human attention tends to remain similar when viewing entertainment video. This assumption is not applicable to special circumstances in which we seek unique visual information within a video frame that cannot be easily characterized by generic human visual attention attributes. Special design is needed for these unique cases, including surveillance and tactical video for security and defense applications.

Acknowledgment. This work has been supported by a grant from the National Natural Science Foundation of China (No. 60572067). The authors wish to thank Mr. Yi Wang for his contributions to this project.

References

1. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Trans. Circuits Syst. Video Technol 13, 560–576 (2003)
2. Tamhankar, A., Rao, K.R.: An overview of H.264/MPEG-4 part 10., Communications and Signal Processing Track, Texas Systems Day, November 15, 2003, Dallas Hall, SMU (2003)
3. Kim, J.-G., Wang, Y., Chang, S.-F., Kim, H.-M.: An optimal framework of video adaptation and its application to rate adaptation transcoding. ETRI Journal 27(4)(2005)
4. Mobile Profile Proposal, ISO/IEC, JTC1/SC29/WG11 and ITU-T SG16 Q.6, document JVT-C161, 3nd meeting, Fairfax, VA, USA, 6-10 March 2002 (2002)
5. Vetro, A., Christopoulos, C., Sun, H.: An overview of video transcoding architectures and techniques. IEEE Signal Processing Magazine 20(2), 18–29 (2003)
6. Xin, J., Lin, C.-W., Sun, M.-T.: Digital Video Transcoding, In:Proceedings of the IEEE, 93(1) (January 2005)

7. Fan, X., Xie, X., Zhou, H.-Q., Ma, W.-Y.: Looking into video frames on small displays. In: Proceedings of the eleventh ACM international conference on Multimedia November 2003, Berkeley, CA, USA, pp. 247–250 (2003)
8. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. ACM Multimedia Systems Journal 9(4), 353–364 (2003)
9. Hua, X.-S., Chen, X.-R., Liu, W., Zhang, H.-J.: Automatic Location of Text in Video Frames, In: Proceedings of the third International Workshop on Multimedia Information Retrieval, Ottawa, Canada, October 2001, pp. 126–129 (2001)
10. Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14496-10 AVC, In: Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-GO50 (2003)
11. H.264/MPEG-4 Part 10: Transform & Quantization, H.264 tutorial http://www.vodex.com
12. Marpe, D., Schwarz, H., Wiegand, T.: Context-Based Adaptive Binary Arithmetic Coding in H. 264/AVC Video Compression Standard. In: IEEE Trans. Circuits Syst. Video Technol 13, 620–636 (2003)
13. List, P., Joch, A., Lainema, J., Bjontegaard, g., Karczewicz, M.: Adaptive deblocking filter. IEEE Trans. Circuits Syst. Video Technol 13, 614–619 (2003)
14. JVT reference software official version, Image Processing Homepage http://bs.hhi.de/ suehring/tml/
15. Wang, Y., Li, H., Chen, C.W.: A novel video coding scheme for mobile devices. Proc. Int. Conf. Mobile and Ubiquitous Multimedia October 2004, College Park, MD (2004)
16. Kuo, C.-H., Shen, M., Kuo, C.-C.J.: Fast Inter-prediction mode decision and motion search for H.264," Proceedings of IEEE International Conference on Multimedia and Expo 2004, Taipei, Taiwan, pp. 663–666 (June 2004)
17. Lee, J., Jeon, B.: Fast mode decision for H.264. In: Proceedings of IEEE International Conference on Multimedia and Expo 2004, Taipei, Taiwan June 2004, pp.1131–1134 (2004)
18. Sun, X., Foote, J., Kimber, D., Manjunath, B.S.: Panoramic video capturing and compressed domain virtual camera control. In: Proceedings of the ninth ACM international conference on Multimedia, Ottawa, Canada, September2001 pp. 229–238 (2001)
19. Wang, Y., Li, H.Q., Fan, X., Chen, C.W.: An attention based spatial adaptation scheme for H.264 videos over mobiles. Int. Journal Pattern Recognition & Artificial Intelligence special issue on Intelligent Mobile & Embedded Systems 20(4), 565–584 (2006)
20. Chang, S.-F., Vetro, A. : Video adaptation: Concepts, technologies, and open issues, In: Proc. IEEE, 93(1) 148–158 (2005)
21. Tourapis, A.M., Au, O.C., Liou, M.L.: Highly efficient predictive zonal algorithms for fast block-mathcing motion estimation. IEEE Trans. Circuits Syst. Video Technol. 12(10), 934–947 (2002)
22. Yin, P., Tourapis, H.-Y.C., Tourapis, A.M., Boyce, J.: Fast mode decision and motion estimation for JVT/H.264. In: Proceedings of International Conference on Image Processing 2003, Barcelona, Spain, September 2003 pp. 853–856 (2003)
23. Shen, B.: Submacroblock motion compensation for fast down-scale transcoding of compressed video. IEEE Trans. Circuits Syst. Video Technol. 15(10), 1291–1302 (2005)
24. Wang, Y., Li, H., Chen, C.W.: An attention-information-based spatial adaptation framework for browsing videos via mobile devices. EURASIP Journal of Advances in Signal Processing (Accepted for publication) (2007)

Fixed-Mobile Convergence, Streaming Multimedia Services, and Peer-to-Peer Communication

Jason J. Yao and Homer H. Chen

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
{jasonyao,homer}@cc.ee.ntu.edu.tw

Abstract. Peer to peer technology has dramatically transformed the landscape of the Internet traffic in recent years. In this position paper, we highlight how the technology relates to the convergence of fixed and mobile networks with features that work irrespective of location, access technology, or user-interactive devices.

Keywords: Multimedia streaming, P2P, fixed-mobile convergence.

1 FMC General Landscape

Mobile telephony is regarded as the sweet spot of today's telecom market. Its 1.7 billion user base still enjoys high growth rates of subscribers compared to other sectors of telecom. Unfortunately, mobile telephony is still based on traditional circuit-switching, which translates to higher costs and less flexibility to carry packet-based data traffic. Although serious efforts, both in publicity and in real work, have been made to migrate toward VoIP, the transformation has barely begun. In developed countries, mobile phone operators face the challenge of finding new growth prospects (for example, there are already more mobile phones than people in Sweden). While battles for market share help consumers, they are costly and provide limited growth for operators. To expand or retain their customer base, operators constantly need to offer new services. One attractive selling point is to offer a single handset working as your mobile phone, home phone as well as office phone. Another is to offer a set of services accessible through multiple devices, such as your office phone, your mobile handset, or even a speakerphone in a conference room. The key for these scenarios is the convergence of fixed and mobile networks that enables new features that work irrespective of location, access technology or user-interface device.

2 Mainstream (Incumbent) Approaches to FMC

As a first step to achieve Fixed Mobile Convergence or FMC [1], many equipment manufacturers now produce dual-mode mobile phones (SIP/WiFi & GSM). Motorola, Nokia and others have been offering such devices that connect to the Internet via a WiFi network whenever possible, or via a gateway of the traditional mobile network

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 52 – 56, 2007.
© Springer-Verlag Berlin Heidelberg 2007

when the phone is beyond the range of WiFi networks. BT Fusion, a service rolled out by British Telecom based on UMA (Unlicensed Mobile Access) is an example of FMC on the service provider side. UMA, a 3GPP global standard, developed by major mobile and fixed/mobile operators and their primary vendors to create a cellular/Wi-Fi convergence solution to support all existing mobile voice and data services, can integrate into existing mobile networks, support all WLAN environments, and can easily fit in the future network evolution plans, including IMS.

IMS (IP Multimedia Subsystem), based on the protocols and principles of IP telephony, is an evolving architecture for providing voice, video and other multimedia services to mobile and fixed phones. Compared to other IP-based protocols, IMS emphasizes central management and billing functions, thus allows operators to offer centrally administered VoIP and other multimedia services on a managed IP network. This is a critical feature to the operators as they are accustomed to provide services to their customers in a "captive" environment. IMS standards come from the 3GPP (3rd Generation Partnership Project), a consortium focused on evolving GSM networks to 3G W-CDMA, and the detail specs can be found at [2].

Today mobile operators have a head start in the FMC race. Their VoIP service can run over any fixed broadband network and their mobile voice users rely on the mobile operators' existing infrastructure or their captive managed network. This advantage will disappear in time as fixed and mobile broadband Internet access (WiFi, WiMAX and 3.5G/4G) becomes widely available from multiple competing providers. Since 3rd party VoIP services, like Vonage or Skype, can run over anyone's broadband connection, mobile operators lose their lock on mobile voice services through market competition. Neither Skype nor Vonage uses or needs IMS. For now, however, mobile operators can launch FMC services more readily and attract more customers.

3 P2P Communication: Disruptive Technology Again for FMC?

Peer-to-Peer or P2P technology has drastically transformed the landscape of the Internet traffic in recent years, but so far we have not yet mentioned its potentials in FMC. P2P file-sharing services like Napster, eMule and BitTorrent, have been widely used and sparked legal controversies. P2P voice streaming created a tremendous sensation when eBay bought Skype for $2.6 billions in 2005. Meanwhile P2P video streaming is also becoming popular. P2P architecture offers a solution to the scalability problem often encountered by streaming networks. As a node joins a P2P network, it not only consumes resources but also contributes its bandwidth or computation power. By relaying data over P2P networks, users receiving data also help its distribution. In addition, the nature of overlay routing in P2P networks makes path diversity possible and this mechanism greatly relieves the load for streaming servers and facilitates traffic load balancing. But some difficulties exist in the current structure of P2P networks. More and more users who join the network are using NAT (Network Address Translation) or behind firewalls, thus they cannot directly contribute their resources to relay data packets. This is especially true for WLAN users. Besides, routing traffic through end-users machines is not efficient with the overhead involved.

Recent news about an investment deal to a startup company, FON, from Skype, Google and premier venture capital firms may reveal the thoughts of forefront

proponents in the P2P camp about how P2P can cut into the FMC pie. As shown in Figure 1, the WLAN routers, while serving as a regular private wireless access point, are equipped with special software that would allow fellow users, called Foneros in FON's terminology, to share its bandwidth and gain wireless access to the Internet. Given enough bandwidth, the roaming Fonero can make voice calls, download data, and view streaming videos just like in their own home LAN. Likewise,the hosting Foneros enjoy the same benefit with other Foneros' services when they are away from home. The peers' resource sharing activities are pushed closer to the core network at the local wireless routers, thus wireless routing does not increase the load of end user workstations. This is similar to the sharing idea of P2P networks as the participants are offering and enjoying collectively Hotspot services. The traffic can also be routed in a P2P fashion while inside the Internet cloud shown in the figure.

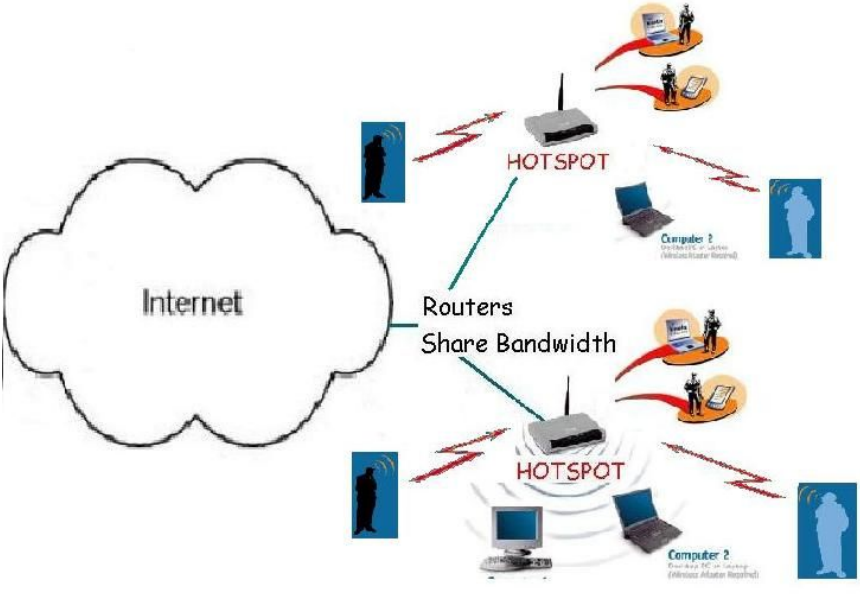


Fig. 1. WLAN Bandwidth Sharing

Despite the exciting prospect of such a scenario, which is referred to as P2P FMC for lack of a better term, many problems exist in the road ahead. Here is a list of them.

1. Pricing model: it would hardly seem fair if someone sets up a Hotspot in a remote area with little traffic and consumes a large amount of resources as he roams extensively. An accurate form of monitoring, accounting and pricing would be needed to make it attractive to participate. Without a sound business model, there will be limited capital investment for such operations to take off.
2. Range and mobility concerns: FON uses WiFi routers as its building blocks, taking advantage of its wide deployment. However, 802.11 was designed mainly for the range of homes and small offices, so it would be difficult for roaming Foneros to

receive a good signal unless they are relatively close to the access point. Furthermore, the protocol did not support connection hand-over, so even when a Fonero is in range with multiple access points, a roaming device and the system cannot dynamically connect with the access point emitting the best signal when the receiving condition changes due to movement. This is a serious disadvantage because mobile users are accustomed to the seamless handover from cell phone operators. These issues may be partially solved when WiMax matures and provides such capabilities, but then again, it may not deploy so widely as WiFi due to its initial cost.

3. Signal interference: Incumbent mobile phone operators purchased bandwidth from the government, so they can map the coverage of their base stations. In the disruptive scenario, private parties use unlicensed band to help each other. To attain better coverage, they can boost their signal strengths but interferences will occur. To mitigate the adverse effect, a negotiation protocol among peer hotspots is necessary to regulate their transmitting power. This is also being researched in WiMax License Exempt Standard Group.
4. Legal issues: At this time, some regulations are preventing a total VoIP takeover with issues such as emergency calls, government wiretapping and international tariffs. Fixed line operators providing unlimited access with fixed monthly fee may oppose the extra traffic generated by bandwidth sharing because they do not profit from such activities. New broadband networks like FTTH (Fiber-To-The-Home) are being deployed as we speak, so the task is upon advocates of FON ideas to convince established ISPs in working together with them.

4 New Information Age

As explained above, P2P-FMC may totally disrupt the incumbents' game plan for FMC. Although the concept of resource sharing is not new, its implementation for telecom did not take off until recent years, thanks to innovative technologies and ideas. Indeed many obstacles still lie ahead, but none of them seems to be fundamentally insurmountable. Traditionally, a wireless operator first has to purchase a licensed bandwidth from regulators and pay for a very costly infrastructure which needs to satisfy many regulations and reach a wide coverage to achieve economies of scale. Only after that can it reap the profit. Naturally incumbent operators try to keep a lock on its customers to protect their investment interest, and since they own the whole network, they can probably implement some new features more quickly to satisfy their customer needs and charge a high premium. In P2P-FMC, every peer can establish its own Hotspot as a part of the whole infrastructure, sharing the startup cost of the operation. Although the underlying network may be heterogeneous, the peers can work together as long as they can communicate at the access level and above. With this unconventional model, anyone can join the effort of building a large infrastructure without a deep pocket, just sharing the proper amount of cost and benefit. In the past, the telecom infrastructure was so costly that only very few entities can afford it, and oftentimes the government has to own and operate it. As always, it is hard to achieve efficiency in a large organization, especially in the public sector, but that seemed to be the necessary evil of the industry. P2P-FMC may finally break the old

rules and change the whole landscape. It has happened before in the commerce world. Just a few decades ago, the capital requirement is huge to set up store fronts for retail business and even more so if it involves a global operation. Now average consumers can buy globally at eBay, and the cost of setting up an online store is only a tiny fraction compared to the cost before. Meanwhile, such online deals also enable average consumers to resell their unwanted goods which, in the past, would most likely just stay in a storage room. More people participate, more transactions take place, and the market is more efficient.

Welcome to the New Information Age.

References

1. Heavy Reading, Fixed-Mobile Convergence Reality Check. vol. 2(26) (December 2004)
2. http://www.3gpp.org/specs/specs.htm

Blind Passive Media Forensics: Motivation and Opportunity

Shih-Fu Chang

Department of Electrical Engineering
Columbia University
New York, NY 10027
sfchang@ee.columbia.edu

Abstract. In this presentation, I will describe the motivating scenarios and applications for blind-passive media forensics (BPMF) technologies, contrasting them with conventional solutions that require adoption of end-to-end protocols or embedding of watermarks. BPMF methods verify the authenticity of the received content at the point of use by checking the integrity of features characteristic of the natural scenes or capturing devices, or the anomalies caused by the tampering operations. I will review the significant results achieved so far and discuss the direction for future research.

Keywords: media forensics, detection of image forgery and tampering, media security.

Information integrity is a fundamental requirement for the cyber space, in which users need to ensure received information is trustworthy, free from tampering or forgery. Audio-visual information (photo, video, and audio) is becoming increasingly important, with a wide range of applications like surveillance, news reporting, intelligence, insurance, and criminal investigation. However, with advanced editing technologies making manipulation easier and easier, the wisdom that "A picture is a fact" [1] is no longer valid. There is critical need in developing robust and flexible techniques for verifying the authenticity of audio-visual information and thereby restoring its trustworthiness.

The needs of trustworthy media can be easily found in many applications in the new era of Web 2.0, in which users are also becoming content creators, publishers, and commentators. In December 2006, the well-known news agency Reuter and Yahoo News announced a joint effort called *You Witness News* [2], in which public users may submit and share news worthy photos and video clips captured on their own cameras (even cameras on mobile phones). Though such application of "citizen journalism" is indeed compelling, there is also a great concern about submission of digitally altered or forged content. Actually, digital alteration has become so frequent that its association to reputable publishers is not uncommon [3, 4]. In March of 2003, the Los Angeles Times published a news photograph, which was later confirmed to be a photomontage (a paste-up of a soldier from another picture), in a report on the Iraq

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 57 – 59, 2007.
© Springer-Verlag Berlin Heidelberg 2007

war. Some popular web sites [5, 6] currently highlight the best photo manipulation effects or computer graphics photo-realistic quality, virtually undistinguishable from natural photos by human eyes. In addition, in applications of intelligence analysis, in order to verify the integrity of video sources, a common question asked is whether different video clips are indeed captured by the same camera at the same location, rather than results of digital mixing.

Faced with the real-world scenarios of content tampering like those discussed above, we are interested in answering the following questions. Has any area in an image or video been tampered with? Is the image/video actually captured by a natural camera, instead of being created by some digital synthesis tool? Is the image or video indeed captured by the camera claimed to be the source? Can we verify two videos are indeed taken from the same location of the same event? With the rapid progress in editing and computer graphics technologies, an automatic system with complete robustness against all varieties of forgery attacks is simply unrealistic. However, a comprehensive suite of methods and tools for detecting various types of tampering anomalies will be possible, making content tampering and forgery much harder.

Conventional techniques for content integrity verification require generation or embedding of additional data, such as content hashing or watermarking. Such solutions are often impractical in real-world situations because of the strict requirement of cooperative sources and compatible end-to-end protocols. In contrast, researchers have embarked on a new direction, *blind and passive media forensics* in recent years and demonstrated some promising results [7]. Such methods extract unique signatures from real-world signals and sensing devices and also detect anomalies caused by tampering. Only the media content at hand is needed in the verification process without requiring additional information.

Considering the myriad of various possible manipulations, the task of checking media content integrity without using any external data may seem to be infeasible. However, careful examinations reveal several promising directions. First, the devices used for capturing video or audio signals have unique *device signatures*, including sensor nonlinearity, interpolation filters, noise patterns, cut-off bandwidth, etc. Such device signatures, if adequately estimated, may be used to verify the source consistency among different parts of an image or video clip. A video showing inconsistent device signatures is likely to be faked. Secondly, digital alteration is not a completely transparent process; instead, each manipulation step may leave *tampering artifacts* that become detectable traces. For example, the frame structures used in compressed audio-visual streams and the effect of re-compression can be used to check whether editing has taken place. Finally, for videos taken at the same location for a real-world event, their associated audio-visual content should have strong correlations as a common visual scene is likely to be covered by multiple cameras and salient environmental sounds will be recorded by separate audio recording devices. Lacking of such *contextual consistency* between multiple audio-video streams captured under the same context may be used as a basis for raising alerts of unverifiable sources.

In this presentation, I will discuss significant results achieved so far in these areas and promising direction for future research.

References

[1] Mitchell, W.J.: The Reconfigured Eye: Visual Truth in the Post-Photographic Era. MIT Press, Cambridge, Mass (1992)
[2] YouWitnessNews (2006) http://news.yahoo.com/you-witness-news
[3] Gavard, S.: Photo-graft: A critical analysis of image manipulation. MA Thesis, McGill University, Montreal, quebec (1999)
[4] Baker, F.: Is Seeing Believing? A Resource For Educators. (2004) http://www.med.sc.edu:1081/ isb.htm
[5] FakeorFoto. http://www.autodesk.com/eng/etc/fakeorfoto/quiz.html
[6] Worth1000 Image Editing Contest Site. http://www.worth1000.com
[7] Ng, T.T., Chang, S.F., Lin, C.Y., Sun, Q.: Passive-blind Image Forensics, in Multimedia Security Technologies for Digital Rights, Elsvier (2006)

Security Models of Digital Watermarking

Qiming Li and Nasir Memon

Polytechnic University

Abstract. Digital watermarking, traditionally modeled as communication with side information, is generally considered to have important potential applications in various scenarios such as digital rights managements. However, the current literature mainly focuses on robustness, capacity and imperceptibility. There lacks systematic formal approach in tackling secure issues of watermarking. One one hand, the threat models in many previous works are not sufficiently established, which result in somewhat superficial or even flawed security analysis. On the other hand, there lacks a rigorous model for watermarking in general that allows useful analysis in practice. There has been some efforts in clearing the threat models and formulate rigorous watermarking models. However, there are also many other cases where security issues are lightly or incorrectly treated. In this paper, we survey various security notions and models in previous work, and discuss possible future research directions.

1 Introduction

A digital watermarking scheme, in general, is a set of algorithms that allow us to embed some information (i.e., *watermarks*) into some host signal (or *cover objects*) in such a way that these watermarks can later be extracted or detected, even if the cover objects are corrupted by a small amount of permissible noise.

A watermarking scheme usually consists of three major components. A *watermark generator* generates desired watermarks for a particular application, which are optionally dependent on some keys. An *embedder* embeds the watermark into the cover object, sometimes based on an embedding key. A *detector* is responsible for detecting the existence of some predefined watermark in a cover object, and sometimes it is desirable to extract an message from the watermarked cover object.

Watermarking schemes are potentially for a number of applications. For example, authentication and tamper detection, ownership and copyright protection, and fingerprinting and traitor tracing. Clearly, it is very important to address the security issues in these applications. However, many of the previous works on digital watermarking mainly focus on the robustness, where the attacker model is often over simplified. In particular, it is often assumed that the attacker only adds some sort of random noise to the cover objects, hence robustness against random noise would be sufficient to withstand watermark removal attempts. However, in this case nothing can be concluded for noise types that are not considered, or for smart attackers with carefully designed modifications, especially

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 60–64, 2007.
© Springer-Verlag Berlin Heidelberg 2007

when a watermark detector is publicly available. Zero-knowledge watermark detection is proposed to limit the information leakage due to public watermark detectors [3]. However, it has been increasingly evident that a capable attacker with the access to the detection oracle would be able to remove watermarks without much difficulty even when the watermark embedding algorithm is not known, as shown by the results of the *Break Our Watermarking System* (BOWS) contest [5].

Furthermore, even if a watermarking scheme is perfectly robust, it is still possible to launch protocol attacks in particular applications. For instance, when watermarks are used as evidence of ownership, it is important that an attacker cannot hinder the claim of ownership from legitimate owners, which can be achieved without estimating or removing the watermark. These attacks are often referred to as *invertibility attacks* [6] or, in a more general form, *ambiguity attacks* [2]. In these attacks, the attacker finds a watermark that is already detectable in a given work, and invert the watermark embedding to obtain a fake original, so as to make a forged claim of ownership that may be difficult to distinguish from that of the real owner. Later work regarding these attacks include [11, 12, 9, 10, 13].

Another type of protocol attacks is called *copy attacks*, where the attacker attempts to copy the watermark in a given cover object to a dissimilar object without introducing much distortion to it [8]. There is some work addressing this type of attacks, including [4, 1, 7].

To address watermarking security issues in a formal and rigorous manner, one can take one of the following two approaches. First, we can examine each particular application scenario, consider a relatively narrow range of attacks, and formulate the security problems for each type of the attacks. Alternatively, we can start by giving a formal definition and model for watermarking itself, and see if we can prove general results that would apply to a wide range of scenarios. Most previous formal approaches belong to the first category, such as [2, 9], and only recently some attempts has been made using the second approach [7].

In this paper, we are going to discuss in detail some of the security notions, and explore possible future research directions.

2 Security Against Protocol Attacks

Craver et al. propose a method to combat invertibility attacks by generating the watermark in a one-way manner [6]. In particular, given a cover object, a one-way hash function is applied to obtain a hash value, and then use a pseudo-random number generator to expand the hash value into a watermark, which is then embedded into the object. It is conjectured that an attacker, after finding a detectable watermark in a cover object first, would have to invert the one-way hash function to make a ownership claim.

Ramkumar et al. [12] point out that such a conjecture may not hold, since an attacker could always perturb the watermarked object and apply the hash function to obtain a random watermark, and then check if the watermark is detectable. Hence, if the false-positive of the underlying watermarking scheme is high, an

attacker could succeed with high probability. An improved scheme is proposed in [12], where the original is required during the ownership proof, and the attacker has to make sure that the true owner's watermark cannot be detected in his fake original. This technique would make a random attacker infeasible.

Observing that it is difficult to design non-invertible watermarking schemes, a scheme based on a trusted third party (TTP) is proposed in [2]. Although provably secure, it may be difficult to find such TTP in practice.

It is later shown that it is possible to build non-inverible schemes without a trusted third party [9]. Their scheme involves a cryptographically secure pseudo-random number generator in the watermarking generation process, which is similar to the use of hash function in [6]. An important difference is that the need to distinguish *valid* and *invalid* watermarks is highlighted, and by limiting the number of valid watermarks to be only a negligible fraction of the total number of possible watermarks, it is then possible to prove the security of the scheme, together with the assumption that the false-positive rate is negligible. A zero-knowledge version of the scheme appears in [10].

Noting that low false-positive rate is essential in the security of non-invertible watermarking schemes, Sencar and Memon [13] propose the use of multiple watermarks instead of a single watermark in spread-spectrum based watermarking scheme to bring down the false-positive rate.

3 Formal Watermarking Model

As we mentioned earlier, there are two approaches dealing with watermarking security. Basically, we can either establish rigorous security for a particular application, or try to give a general formal model for watermarking and try to prove general results.

While the first approach may give sound and provably schemes in particular application scenarios, the implications of such techniques are often limited in the sense that new analysis and new proofs are often required to deal with different application scenarios and/or different attackers.

For example, although the security prove in [9] is sound, the scheme is based on a spread-spectrum watermarking scheme, and it is not straightforward to see how to adapt the proof for other types of watermarking schemes (e.g., QIM). Also, the use of a cryptographically secure pseudo-random number generator may not be suitable in some applications. However, from the proof itself it may be hard to see how the security is affected if the generator is replaced by other one-way functions.

The second approach ([7]) basically defines an *ideal* watermarking scheme, where the cover objects are in a well-defined metric space, and a point in the space is declared as watermarked if and only if it is near a watermarked object. Hence, a secure watermarking scheme can be defined as one that behaves very closely as the ideal counterpart. If such watermarking scheme can be built, it is clear that it is robust and resistant to copy attacks (which requires a distant object to be watermarked).

This approach may seem comprehensive and more formal at first, it however remains somewhat too theoretical and gives little insight or guideline regarding how to design such schemes in practice. For example, in real applications the most difficult part is actually the definition of the similarity measure (or distance function), which has to capture the perceptual characteristics of the data, and also take into consideration all possible permissible noise. Unfortunately, such formal model would not be able to tell a developer how to define such a metric space properly, let alone a secure scheme on top of it.

Furthermore, the model in [7] does give some positive results on how to construct secure watermarking schemes in some arbitrary metric space, but the main idea seems to make the scheme secure by making it difficult to sample in the neighbors of a watermarked object. Although this leads to provable security, it is hard to see how such difficulty in sampling can be imposed in practical scenarios.

4 Conclusions

To formulate the security problems and to design techniques to tackle them is a tricky business in digital watermarking. Many previous approaches are more or less heuristic, which often lead to schemes that are later proved insecure. Hence, we are going to need more rigorous approaches to better assess the security.

There are currently two main categories of approaches. One is application and attack specific, but easier to be applied. The other is more high level and theoretical, but allows more general results to be proved.

To achieve security in watermarking in general, we need a blend of these two approaches, where the results are rigorous and general, yet they allow detailed analysis of security in practice.

References

[1] Adelsbach, A., Katzenbeiser, S., Veith, H.: Watermarking schemes provably secure against copy and ambiguity attacks. In: Proceedings of the 2003 ACM Workshop on Digital Rights Management, pp. 111–119 (2003)

[2] Adelsbach, A., Katzenbeisser, S., Sadeghi, A.: On the insecurity of non-invertible watermarking schemes for dispute resolving. In: Kalker, T., Cox, I., Ro, Y.M. (eds.) IWDW 2003. LNCS, vol. 2939, Springer, Heidelberg (2004)

[3] Adelsbach, A., Sadeghi, A.: Zero-knowledge watermark detection and proof of ownership. In: Moskowitz, I.S. (ed.) Information Hiding. LNCS, vol. 2137, pp. 273–288. Springer, Heidelberg (2001)

[4] Barr, J., Bradley, B., Hanningan, B.T.: Using digital watermarks with image signatures to mitigate the threat of the copy attack. In: ICASSP, pp. 69–72 (2003)

[5] Break our watermarking system contest. `http://lci.det.unifi.it/BOWS/`

[6] Craver, S., Memon, N., Yeo, B.L., Yeung, M.M.: Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. IEEE Journal on Selected Areas in Communications 16(4), 573–586 (1998)

[7] Hopper, N., Molnar, D., Wagner, D.: From weak to strong watermarking. In: Theory of Cryptography Conference (2007) http://eprint.iacr.org/2006/430.
[8] Kutter, M., Voloshynovskiy, S., Herrigel, A.: The watermark copy attack. In: Electronic Imaging 2000, Security and Watermarking of Multimedia Content II, vol. 3971 (2000)
[9] Li, Q., Chang, E.-C.: On the possibility of non-invertible watermarking schemes. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 13–24. Springer, Heidelberg (2004)
[10] Li, Q., Chang, E.-C.: Zero-knowledge watermark detection resistant to ambiguity attacks. In: ACM Multimedia Security Workshop (2006)
[11] Qiao, L., Nahrstedt, K.: Watermarking schemes and protocols for protecting rightful ownership and customer's rights. Journal of Visual Communication and Image Representation 9(3), 194–210 (1998)
[12] Ramkumar, M., Akansu, A.: Image watermarks and counterfeit attacks: Some problems and solutions. In: Symposium on Content Security and Data Hiding in Digital Media, pp. 102–112 (1999)
[13] Sencar, H.T., Memon, N.: Combatting ambiguity attacks via selected detection of embedded watermarks. IEEE Transactions on Information Forensics and Security

Evolvement of DRM Schema: From Encryption to Interoperability and Monitoring

Tiejun Huang

Institute for Digital Media, Peking University,
100571 Beijing, China
tjhuang@jdl.ac.cn

Abstract. By reviewing DRMs up to now and two typical examples – AVS DRM and DMP IDP, the paper tries to find out the fundamental challenge of content protection approach from technical and social viewpoints. Not only it is difficult to deploy and update content encryption and security infrastructure, but also the content diffusion is limited and Fair Use is affected. The new schema for DRM should be content monitoring system in public space that prevents illegal diffusion of content in copyright but permits content being used freely in private space or for social liberty. The traditional rights in analog times will fluently move to digital space under the proposed schema.

Keywords: DRM, Protection, Interoperability, Authentication, Monitoring.

1 Introduction

Although it isn't a normal way to start a technical paper by a literary quotation, I want to start the discussion on digital rights management (DRM) by the famous gnome from Charles Dickens -

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to heaven, we were all going direct the other way - in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

(From A Tale of Two Cities)

The DRM story came into the view of public and mass media from the transition to the new millennium when more and more audio and video contents in copyright or out of copyright were digitized and bestrewed into Internet to be shared by public. The PC with higher performance and the broadband widely deployed – made it easier and easier to record, store, process, deliver and exchange digital content with low cost.

When someone cheer for the best of times in digital, some others despair for the lost of control on their content in the worst of times for copyright.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 65 – 75, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The fundamental reason for this is media content in digital can exist and be spread without limitation of the physical medium. In analog times of last millennium and before that, physical medium shapes the content (e.g. characters and figures on bones, stone, metal) and become the measure to control it's copyright (e.g. paper book, audio and video tape and movie on plastic DVD). DMCA (The Digital Millennium Copyright Act) seem to be signed by US President at the end of last millennium to prove specially the forecast of Charles Dickens. After DMCA, the MP3 on Internet strongly crashed the record industry and MP4 and others online digital video technology began to strike video industry.

Digital rights management was expected to save the content industry. Various DRM technologies, solutions, standards were proposed in last several years and partly deployed but effect is limited. From Conditional Access System for DTV to HDCP (High Bandwidth Digital Content Protection) that protects data from STB to TV monitor, more and more separated efforts appeared to protection in consumer electronics field. CSS (Content Scramble System) from DVD Content Control Association, Secure Digital Music Initiative for MP3 player, DTCP(Digital Transmission Content Protection) for home network CPRM (Content Protection for Removable Media) for recording on removable media, Secure Video Processor (SVP) for decoder and AACS (Advanced Access Content System) for high definition optical media. The Coral Consortium launched in the end of 2005 and the Marlin Joint Development Association started in the beginning of 2005 announced to develop standards ensuring that copy-protected content can be played on any kind of consumer electronics device by converting from one DRM format to another by accessing a conversion service.

IT industry is playing a more and more important role in providing end-to-end DRM solution. At first, Intertrust, ContentGuard and other small companies proposed respective DRM solution. Then Apple DRM system entered into the market with the success of iPod and iTunes. Microsoft provided its DRM solution with Microsoft Media. As a different approach, Sun Microsystems announced DReaM initiative to develop a DRM solution focusing on open-standards-based-solutions that will be open source and royalty-free.

Fairplay – the DRM system from Apple - is the most successful one in the market. Facing the complaint about a private solutions like Fairplay may be used to monopolize market in last year, Apple is appealed to open it's DRM so that music purchased from iTunes can be played on digital devices purchased from other companies, and protected music purchased from other online music stores can play on iPods. Steve Jobs fights back by "thoughts on music" [2] with the following key points:

(1) Only 22 out of 1000 songs, or under 3% of the music on the average iPod, is purchased from the iTunes store and protected with DRM. In other words, Fairplay and his congener from Microsoft and Sony etc doesn't divide the market into several unconnected garden, 97% world of the music world is open without DRM.

(2) There is no theory of protecting content other than keeping secrets...one must still "hide" the keys which unlock the music on the user's computer or portable music player. If Apple doesn't keep the secret itself but license it to its competitor, the security of DRM will come down.

(3) DRMs haven't worked, and may never work, to halt music piracy. To abolish DRMs entirely so that every online store sells DRM-free music encoded in open

licensable formats. This is clearly the best alternative for consumers, and Apple would embrace it in a heartbeat.

Is Jobs right? For thought of music, maybe right. For thought of DRM, partly wrong. But anyway, Job's thought provide an exact point to review the short history of DRM and foresee its long future.

This paper try to find an answer by analyzing and comparing the three DRM schemas – protection by cipher technology, interoperable DRM infrastructure and content-based media authentication and monitoring in public space.

Firstly, this paper will review the AVS DRM from the Audio and Video coding standard of China for which the author is a major designer. The sample DRM shows that it is possible to design a DRM standard by which all the players in a market can focus on respective DRM products at different chains and inter-operate without comedown of security. In fact, there are many other successful cases outside of DRM field that doesn't control by one company. For example, e-business and its security infrastructure are open and secure in practical sense.

For the more, it is also possible to construct a worldwide interoperable DRM infrastructure also. Starting from the Digital Media Manifesto[6] published in Sep.,2003, Leonardo Chiariglione lead the Digital Media Project (DMP)[7] with the mission to "promote continuing successful development, deployment and use of Digital Media that respect the rights of creators and rights holders to exploit their works, the wish of end users to fully enjoy the benefits of Digital Media and the interests of various value-chain players to provide products and services". DMP released the "Interoperable DRM Platform" Phase I specification (for Portable Audio and Video Devices) in 2005 and Phase II specification (for Stationary Audio and Video Devices) in 2006. Phase III specification that tries to be media infrastructure will be release in 2007.

The fact that under 3% of the music is protected with DRM Fairplay pricks a bubble that protection approach is an effective measure for rights management. In the last section, by analyzing the disadvantages of current DRM, this paper will clarify the simple target for digital rights management is to prevent the diffusion of content in copyright in the public space like Internet and allow the Fair Use in the private space of consumer and commonweal space like library and classroom. To meet this target, content-based media authentication and monitoring for plain content should replace cipher-based media encryption and watermarking. And as a result, media usage in analog times will be back. The analog holes in current DRM will be blocked also.

2 AVS DRM: Simple But Typical DRM

Among dozens of DRM solution today, AVS DRM is a sample but typical one. Alike the video and audio coding standard from the Audio and Video coding Standard Working Group of China tries to provide an uniform AV format for the country to avoid possible bulwark between different AV systems such as DTV broadcasting systems, IPTV and storage AV player, the DRM part of the national standard tries to avoid possible rampart generated by different cipher technology in DRM. Uniform AV coding and DRM standards will decrease the total cost for media industry and consumers by boosting competition at products level.

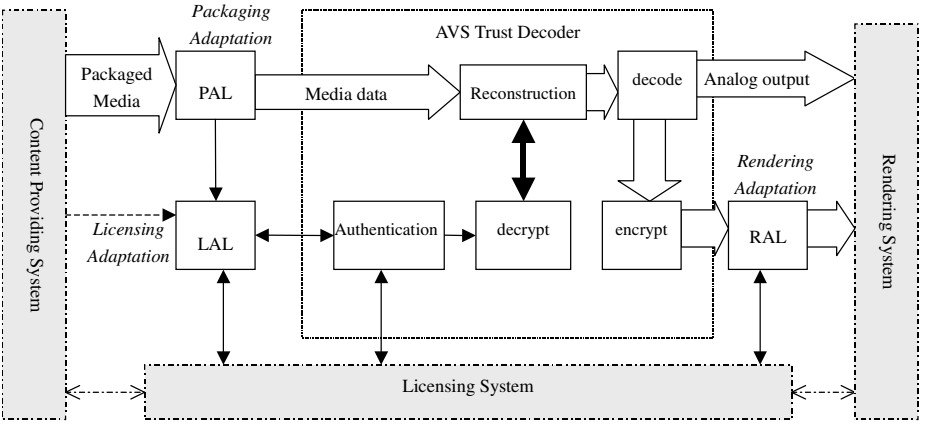


Fig. 1. Architecture of AVS DRM Core profile

Fig.1 shows the Architecture of AVS DRM Core profile which consists of the AVS Trusted Decoder (ATD), the Adaptation Layers and the peripheral systems.

AVS Trusted Decoder is an extension of the traditional decoder with the addition of an authentication module, a decryption module, a data reconstruction module and an output encryption module. AVS DRM peripheral systems consist of the content providing system, the licensing system and the rendering system. They are related but not essential components of AVS DRM. The adaptation layers between the ATD and the peripheral systems are responsible for the interaction and negotiation among them. There are three adaptation layers –Packaging Adaptation Layer (PAL), Licensing Adaptation Layer (LAL) and Rendering Adaptation Layer (RAL).

The ATD should be strongly protected as a whole, for example, as an ASIC chip with physical anti-track shell. Licensing system can communicate with it through an security channel setup by authentication model in ATD and counterpoint in licensing system such as smartcard for end user or AAA (authentication, authorization, and accounting) server at head-end system.

AVS DRM is an open solution that any company can provide decoder with ATD and any content provider can publish its content through any transmission system as long as they obey the same AV coding and DRM standard. Although ATD contains "secret" components that should be touched only by authorized entity or only by itself (e.g. private key), it doesn't mean the system must be kept "secret". All the players in a market know there is a secret but at the same time they can produce respective DRM products at different chains and inter-operate without comedown of security.

Just like Fairplay is workable in Apple world, AVS DRM with uniform cipher technology is workable in a country like China. But for worldwide, we must look for super solution for co-existence of different cipher technologies, content format and so on.

3 DMP Interoperable DRM Platform

The Digital Media Project (DMP) is a non-profit Association registered in Geneva, Switzerland. Its mission is "to promote the successful development, deployment and use of digital media that respect the rights of creators and rights holders to exploit their works, the wish of end users to fully enjoy the benefits of digital media and the interests of value-chain players to provide products and services, according to the principles laid down in the Digital Media Manifesto" [6] [7].

The DMP is the first attempt at achieving an end to end Interoperable DRM Platform (E2E IDP). Its objective is to achieve consensus between representatives of all traditional value chain players (including end users) using standard technologies. It is E2E because it represents content throughout the chain including representation of all entities digital or not that are relevant to the value chain and to which rights are attributed for. This initiative presents an opportunity for minimizing differences between value chain players focusing on use cases that support the interests of all.

DMP approaches the problem of DRM Interoperability by specifying technologies – that DMP calls Tools – required to implement what DMP calls "Primitive Functions". These are "smaller" functions obtained when the functions value-chain users perform when they do business between themselves are broken down into more atomic elements. DMP provides specifications of Tools enabling Primitive Functions along with examples of how Value-Chains serving specific goals can be set up using the standard Tools. The ensemble of all standardized DRM Tools is called "Interoperable DRM Platform (IDP)".

A Value-Chain is a group of interacting Users, connecting (and including) Creators to End-Users with the purpose of Delivering Content to End-Users. In general Users require Devices to perform the Functions proper of their role in the Value-Chain. Fig. 2 shows the Devices mentioned above in a generic Value-Chain and identifies their principal relationships.

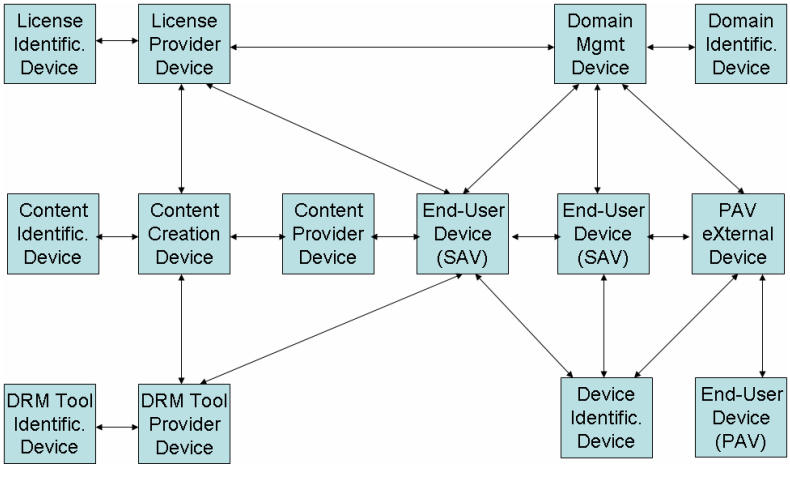


Fig. 2. Devices in a Value-Chain defined by DMP

4 Challenges from Technical Perspective

There are three approaches used in DRM to protection content up to now: Cipher-based model, Watermarking and Digital fingerprinting.

Cipher-based approach employs encryption tools to protect content and key management infrastructure to delivery encryption key and other sensitive information. Encryption is the process for controlling access to confidential data, known as plaintext, by scrambling the data into an unintelligible form [8]. The key used by encryption and other sensitive information like license which stores usage rules about specialized content for specialized end user(s) can be transferred between value chain participants one by one through a trusted key management and delivery infrastructure which need to authenticate all of the participants or their devices.

Watermarking [9][10] has been proposed as a means for content protection evenafter data has been decrypted. A watermark is a signal that is embedded into an original content to produce the watermarked content. Content owner related information can be inserted to identify the ownership of the copyright. Usage rules can be embedded to direct the watermark-enabled device how to use the content.

Digital fingerprinting was proposed by [11] more than twenty years ago. Fingerprinting is the process of embedding a distinct set of marks into a given host signal to produce a set of fingerprinted signals that each "appear" identical for use, but have a slightly different bit representation from one another. These differences can be used to track of a particular copy of the fingerprinted signal. The marks, also called the fingerprint *payload*, are usually embedded through the process of robust digital watermarking.

The above technology was studied for many years. It is expected that more strong or robust protection to appear in the future. But no matter what the protection technology is or would be, there are more fundamental problems that must be faced. Here are some most important that were pointed out in last several years (partly from [8]):

(1) Cipher-based approach encryption techniques do not offer any protection once the encrypted data has been decrypted. As an extreme but popular case, analog hole always exist. It is always possible to make analog copy from the end device through the use of "legacy" devices that are not DRM compliant. Once the content flows from DRM world into analog world, it can be converted into digital without DRM.
(2) Despite the considerable effort that has been spent on developing robust watermarks for digital video, the robustness and security of current watermarking techniques may not be sufficient for some DRM applications. Removal attacks and spatial and temporal synchronization attacks remain challenging for watermark detection. Security is also an issue.
(3) Digital fingerprint faces most problems of watermarking. For the more, privacy infringing is challenge when end user related information is embedded in the content.
(4) For DRM systems are required and must govern all access, use, and copying of protected content and do so securely, a enormous infrastructure is needed to construct. Who should be charged to do this? Content owners, providers, operators or

consumers? Although it is always consumers to pay for everything, it is impossible for DRM to be afforded when it limits Fair Use space or infringes the privacy of the consumers.

(5) Beside the cost for constructing a DRM system, renew an attacked system means more cost and more complex tasks. Some challenging questions remain for using device revocation also.

(6) Protection technology cumbers Fair Use of content. Encryption makes fair share in family, library and classroom impossible or very difficult.

5 Challenges from Social Perspective

Before technical guys can find more sophisticated protection artifices, although his iPod and Fairplay gained amazing success, Steve Jobs already has been beyond endurance on DRM and announced that "DRMs haven't worked, and may never work, to halt music piracy."

Another thought from librarian [12] mentioned that DRM is perceived as a barrier for it does build in obsolescence, end equitable access, invades user privacy and enable censorship.

DMP thought DRM by Digital Media Manifesto [6] which is the foundation for it interoperable infrastructure. After analysis on disadvantages of current DRM, DMP proposed the following steps to remove the hurdles to a fuller digital media experience. Although DMP approach is ahead of current protection approaches and standards, it isn't difficult to notice that DMP solution is influenced by protection methodology also. By protection approach, it is not possible or affordable to Map Traditional Rights and Usages (TRUs) to the Digital Space [13].

DMP TRUs should be listed in the most important contributions to media. It is the start point for next generation rights management. The 88 TRU picked out by DMP are classified into five categories as following:

(1) Already-established legislative TRUs of content creators (21)
(2) Already-established legislative TRUs belonging to end-users (7)
(3) Commercial and remuneration TRUs of direct economic significance (22)
(4) TRUs related to general social liberties (13)
(5) Fundamental TRUs from historical practice and interaction with analogue media (14)
(6) Consumer-choice TRUs relevant to the high-tech environment (12)

You can find the protection approach satisfies few of them but make most of them impossible or excessive complex to implement in digital space.

Content protection should not forget the primary goal of content itself. Technology should accelerate the diffusion of knowledge and culture but not limit it. Let's image a scene many years later, when our descendants excitedly discover the digital content piece in our age, they find it was locked in black box by encryption but key is not there because the complex key management system designed by us disappeared or crashed like dinosaur, we should be ashamed for our short see on rights management to make our descendants disappointed.

6 Monitoring Rights in Public Digital Space by Contentprint

Although amount of effort was taken to protect digital content, the real situation is most of the content in copyright still go around everywhere and the protection is revolted by many consumers and organizations. Contrasting with stronger and stronger protection technologies, more and more contents in copyright without DRM are exchanged in Internet. Now Apple and Microsoft decided to provide DRM-free music on their online store. It is the time to consider new approach for rights management.

The origin of content protection is digital content in copyright was released into cyberspace. The involuntary reaction of media industry is to protect the content by encryption, prove the ownership by watermarking and track the user by fingerprinting. An important fact ignored by most people is the cyberspace. The cyberspace is a kind of public space. It is illegal to spread content in public space without permission of the content owner. So the real question is how to prevent illegal spreading activities.

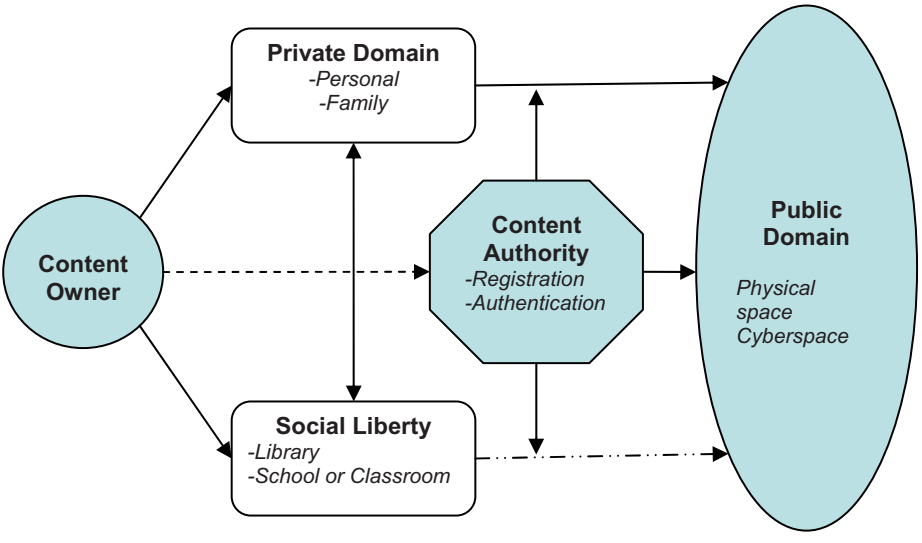


Fig. 3. Content diffusion

Fig.3 shows the question in detail. After a content is created, it should be registered on a authority organization. A person buys the content from the rights owner or middle-man and uses it by himself/herself or shares it in his/her private space (e.g. family). Library can collect and loan it to its readers for social liberty. After the expiration of the copyright, the content enters the public space as a part of human knowledge or culture. In analog ages, pirate is controlled by legal measure. In digital space, pirate can be controlled by legal measure also. For the technical community, the most important that should be invented is not the tool for content owner to protect their content but the tool for authority or police to trace and prevent the pirate activities in digital space.

The shift of DRM schema from content protecting by owner to pirate monitoring by public authority follows typical social evolvement. Not like previous DRM that modifies the content by scrambling or marking, the content authentication-based DRM doesn't change anything. For the more, once a content is registered in authority database, its pirate copy (even though with artificial or other change) will be monitoring in public space. If someone attempts to spread a content in copyright to others, he or she will be found and the activities will be stopped. The monitoring only operates in public space, the Fair Use of content in private space or sharing for social liberty purpose is outside of the monitoring.

The new schema is better than cipher-based model and watermarking model. The advantages are following:

(1) Content without change: maximum freedom for the content diffusion not only for today but also for the future.
(2) Consumers friendly: anyone can get a copy from the content owner. The consumer only pays for the content. No DRM-enable device is needed. All traditional right usages are reserved.
(3) No analog hole: The monitoring system is designed to block analog hole. Although anyone can get a copy but he or she can't spread it for the monitoring system in the public space will stop any illegal transmission.
(4) Flexible right management: content in copyright that is registered will be monitored in public space. Once expiration of the copyright, it will be out of monitoring and be released to public space automatically.
(5) Cost can be controlled: monitoring system need cost also. But the cost is from content owner who wants to protect the content that can bring him revenue. If the content owner believes his/her content is necessary to be monitored, the means he or she wants to pay for it. That is to say, there are enough money to construct a monitoring system and maintains it.

To monitor digital content in digital space, a content-related identifier is needed. The identifier is called contentprint. Contentprint is a new word which is created to identify a content just like fingerprint, voiceprint to identify a person. It is abstracted from a media content. Content should be robust enough - constant to various variations (compression, analog to digital or contrariwise, zoom etc.) and should be unique – different content should generate different contentprint. When special feature is regarded, contentprint can be reified as visualprint, auralprint and so on.

There are some similar concepts for contentprint. The first is perceptual hashing[14]. Hash mean satisfies the unique characteristic but doesn't satisfy the constant characteristic. Another is video or audio fingerprint. For fingerprinting is used to describe embedding user-related information into original content, we prefer to create the new word to describe abstracting features from the content.

There are many works on content authentication [15~18] for different purpose. For need to authenticate content that has experienced legitimate editing in addition to potential tampering attacks, [19] develop one formulation based on a strict notion of security, and characterize and interpret the associated information-theoretic performance limits.

Although there are many algorithms for contentprint abstracting, research on robust abstracting algorithm for ultra amount content dataset is needed. Benchmark for

contentprint abstracting is needed also. For the more, normative contentprint is needed to setup open contentprint management and monitoring system.

7 Conclusion

By reviewing state of art DRMs including two DRMs that the author involves in, the paper checked the challenges of DRM from technical and social perspectives. The conclusion is that protection approach solved few problems for rights management but lead to more new problems (e.g. block the Fair Use). The fact that under 3% of the music is protected with Apple Fairplay is an auspice for replacing it with new schema.

This paper points out a simple fact that the origin and final target for rights management is to prevent illegal spread in public space. Content-based media authentication and monitoring will play more important role in new generation DRM. To construct such a new rights management, more efforts on contentprint abstracting, authentication, identification and high performance contentprint monitoring are neeed.

References

1. The U.S. Copyright Office. The Digital Millennium Copyright Act. (1998) http://www.copyright.gov/legislation/dmca.pdf
2. Fernando, G., Jacobs, T., et al.: Project DreaM: An Architectural Overview. Sun Microsystems White paper (September 2005)
3. Jobs, S.: Thoughts on Music. (February 6, 2007) http://www.apple.com/hotnews/thoughtsonmusic/
4. China national standard GB/T 20090.2. Advanced Audio and Video coding Part 2: Video (2006)
5. Tie-jun, H., Yongliang, L.: Basic Considerations on AVS DRM Architecture. J. of Computer Science and Technology 21.3, 366–369 (2006)
6. The Digital Media Manifesto (2003/09/30). http://www.dmpf.org/manifesto/dmm.htm
7. Digital Media Project. http://www.dmpf.org/
8. Eugene, T.L., Ahmet, M.E., Rgginald, L.L., Edward, J.D.: Advances in Digital Video Content Protection. In: Proceedings of the IEEE, vol. 93, pp. 171–183 (2005)
9. Doërr, G., Dugelay, J.-L.: A guide tour of video watermarking. Signal Process. Image Commun. 18(4), 263–282 (2003)
10. Cox, I., Miller, M., Bloom, J.: Digital Watermarking. Morgan Kaufmann, San Francisco, CA (2002)
11. Wagner, N.R.: Fingerprinting. In: Proceedings of IEEE Symp. Security and Privacy, pp. 18–22 (1983)
12. Hill Slowinski, F.: What Consumers Want in Digital Rights Management (DRM): Making Content as Widely Available as Possible in Ways that Satisfy Consumer Preferences. AAP/ALA White Paper (March 2003) http://dx.doi.org/10.1003/whitepaper
13. DMP0270. Collection of TRU templates (November 11, 2004). http://www.dmpf.org/open/dmp0270.zip
14. Kalker, T., Haitsma, J.A., Oostveen, J.: Issues with digital watermarking and perceptual hashing. In: Proceedings of SPIE 4518, Multimedia Systems and Applications IV (2001)

15. Du, R., Fridrich, J.: Lossless Authentication of MPEG-2 Video. In: Proceedings of 2002 International Conference on Image Processing, vol. 2, pp. 893–896 (2002)
16. Pröfrock, D., Richter, H. Schlauweg, M., et al.: H.264/AVC video authentication using skipped macroblocks for an erasable watermark. In: Proceedings of Vision Communication and Image Processing, pp. 1480–1490 (2005)
17. Queluz, M.P.: Towards robust, content based techniques for image authentication. In: Proceedings of the 2nd Workshop on Multimedia Signal Processing, pp. 297–302 (1998)
18. Lin, C.-Y., Chang, S.-F.: Issues and solutions for authenticating MPEG video. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases, San Jose, CA, USA (1999)
19. Martinian, E., Wornell, G.W., Chen, B.: Authentication with Distortion Criteria. IEEE Transactions on Information Theory 52(7), 2523–2542 (2005)

Shape Retrieval Based on the Relativity of Chain Codes

Junding Sun[1,2,3] and Xiaosheng Wu[1,2]

[1] School of Computer Science and technology, Henan Polytechnic University, Jiaozuo 454003, China
[2] Provincial Key Laboratory of Computer Information Processing Technology, Suzhou University, 215006, China
[3] Complex Networks Lab, Hena Henan Polytechnic University, Jiaozuo 454003, China
{sunjd,wuxs}@hpu.edu.cn

Abstract. The direction chain code has been widely used in image retrieval for its simplicity and low storage requirement. One problem is that the traditional methods did not consider the relativity among the chain codes, which limited its use in image retrieval. In this paper a novel shape feature called chain code relativity entropy (CCRE) is proposed for shape classification and retrieval. The direction chain code is firstly mapped to different state of a Markov chain and a new transition probability matrix is introduced. Then, relativity histogram is defined, which includes the transition probability of one state to the others and the others to this one. Based on relativity histogram and information theory, we give the definition of CCRE. After that, the characters of CCRE and an improved method are discussed. Comparisons are conducted between the CCRE and several other feature descriptors. The results show that the CCRE is efficient and it provides noticeable improvement to the performance of shape retrieval.

Keywords: direction chain code, shape retrieval, transition probability matrix, chain code relativity entropy (CCRE).

1 Introduction

Retrieval of images based on the shape of objects in images is an important part of Content-Based Image Retrieval (CBIR). There exist several methods for the shape analysis of objects. These methods can be divided into two categories, contour-based and region-based [1]. The main concern of this paper is based on the Freeman chain code. The Freeman chain code [2] is a compact way to represent the contour of an object. However, it is very sensitive to noise as the errors cumulative. The starting point, the orientation and the scaling of a contour also affect the chain code greatly. In order to solve such problems, two statistical methods, chain code histogram (CCH) and Minimize Sum Statistical Direction Code (MSSDC) were introduced in [3] and [4]. These two methods are translation and scale invariant shape descriptor. Furthermore, MSSDC is invariant to rotation. However, these two methods cannot preserve information on the exact shape of a contour, because they only show the probabilities of the different directions present in the contour. That is to say, the mentioned

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 76–84, 2007.
© Springer-Verlag Berlin Heidelberg 2007

methods didn't take the direction distribution in a chain code into account. In [5], we proposed using Chain Code Coherence Vector (CCCV) and Chain Code Distribution Vector (CCDV) as contour descriptors, which take the direction distribution into account and gain higher performance than the traditional methods. It is a pity that the relativity of one direction with the other directions was ignored.

In this paper, a novel method, called Chain Code Relativity Entropy (CCRE) is presented. The proposed method emphasizes the relativity of one direction of the chain code with the other directions, which also has the advantages of being invariant to translation, rotation, and scaling. Experiment results demonstrate the efficiency of our method.

In the next section we describe the related works correlative to the chain code. In Section 3, we put forward a new strategy (CCRE) to describe the relativity between the chain codes. Similarity measure is given in Section 4. Experimental Setup and Results are given in Section 5. We provide the conclusion in Section 6.

2 Related Works

The traditional direction chain code is defined as follows: for one pixel on the boundary of an object, it has n neighbors, numbered from 0 to n-1, which are called direction codes. According to different connectivity definition, there exist 4-directions chain code and 8-directions chain code illustrated in Fig. 1. Evidently, the traditional boundary chain code varies with the scaling and rotation. However, shape representation is normally required to be invariant to translation, rotation, and scaling.

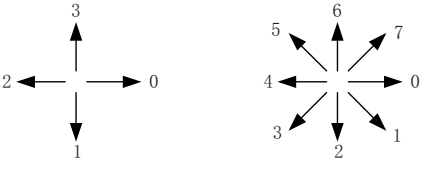


Fig. 1. 4- and 8-directions chain code

In order to solve such problems, CCH was proposed in [3], which is defined as

$$q_i = n_k / n . \tag{1}$$

where n_k is the number of chain code values k in a chain code, and n is the number of links in a chain code. The CCH shows the probabilities for different directions present in a contour.

In [4], statistical direction code (SDC) was defined to describe the feature of a chain code as

$$X = (x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7) . \tag{2}$$

where $x_i(i=0,\cdots,7)$ counts the quantity of pixels on the boundary with direction number i. Because SDC is not invariant to rotation, [4] also gave a code named Minimize Sum Statistical Direction Code (MSSDC).

However, CCH and MSSDC have the same drawback. They didn't take the direction distribution in a chain code into account. In [5], we proposed two methods, CCCV and CCDV, solving such problem. For CCCV, the directions were classified as either coherent or incoherent. Coherent directions are a part of some sizable contiguous sequence while incoherent directions are not. For 8-directions chain code, CCCV was defined as

$$<(\alpha_0,\beta_0),(\alpha_1,\beta_1),\cdots,(\alpha_7,\beta_7)>. \tag{3}$$

where α_i denotes the coherent directions an β_i the incoherent directions. The CCCV could prevent coherent directions in one chain code from matching incoherent directions in another. On the other hand, CCDV was defined as

$$<(h_0,\sigma_0),(h_1,\sigma_1),\cdots,(h_7,\sigma_7)>. \tag{4}$$

where $h_i(i=0,1,\cdots,7)$ denotes CCH of direction i, and σ_i denotes the distribution of direction i. The definition of σ_i was discussed in detail in [5].

It is clear that the variation of a contour is mainly reflected by the variation of the direction in the chain codes. However, such variation was not considered in CCH and MSSDC. In CCCV and CCDV, only the chain codes with the same direction were taken into consideration. In order to describe the relativity of different directions of a contour, CCRE, a novel descriptor is introduced in section 3.

3 Chain Code Relativity Entropy

Let {10103322} be a 4-directions chain code whose contour is depicted in Fig. 2. It is evident that the main feature of the contour is reflected by the variation of $1\to0, 0\to1, 1\to0, 0\to3, 3\to2$ and $2\to1$ .For the methods mentioned above, such feature was not discussed.

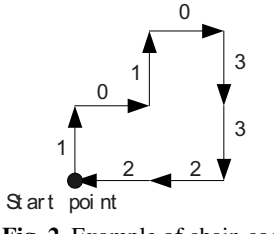


Fig. 2. Example of chain codes

3.1 Relativity of Chain Codes

It is known that the chain codes describe the directions of a contour. If we regard the different direction as different state, then, the Freeman chain code can be thought of as a chain state. Therefore, the variation of the directions can be represented by the variation of the states in the chain state. It is apparent that this problem can be solved by Markov chain. A Markov chain is a chain that can be in one of several states, and can pass from one state to another each according to fixed probabilities. By this way, the chain code of a contour can map to a Markov chain. Accordingly, the number of states is 4 for 4-directions chain code and 8 for 8-directions code. The state variation of a Markov chain can be denoted by its transition-probability matrix, so the

transition-probability matrix also means the variation of the directions in a Freeman chain code. In order to gain some characters, we define a new transition-probability matrix.

Let k_{ij} denote the number of transition from *i*-state to *j*-state in a Markov chain, n be its whole states. A novel transition-probability matrix is defined as

$$\boldsymbol{P} = \begin{bmatrix} 0 & p_{01} & \cdots & p_{0(n-1)} \\ p_{10} & 0 & \cdots & p_{1(n-1)} \\ \vdots & \vdots & \vdots & \vdots \\ p_{(n-1)1} & p_{(n-1)2} & \cdots & 0 \end{bmatrix}. \tag{5}$$

where p_{ij} denotes the transition probability from *i*-state to *j*-state, $p_{ij} = \dfrac{k_{ij}}{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} k_{ij}}$ and $\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} p_{ij} = 1$ for $i \neq j$ and $p_{ij} = 0$ for $i = j$. It is clear that the definition is different from the transition-probability matrix educed from Markov chain for whose $p_{ij} = \dfrac{k_{ij}}{\sum_{j=0}^{n-1} k_{ij}}$ and $\sum_{j=0}^{n-1} p_{ij} = 1$. On the other hand, it is worth the whistle that we do not consider the variation of one state with itself. The reason is given thereinafter.

3.2 Chain Code Relativity Entropy

The transition-probability matrix $\boldsymbol{P}$ reflects the variation of the states in a chain, so it can be thought of as the relativity between the chain codes. For *i*-state, a relativity histogram is introduced and defined as

$$\boldsymbol{H}_i = \{p_{i0}, p_{i1}, \cdots, p_{ij}, \cdots, p_{i(n-1)}, p_{0i}, p_{1i}, \cdots, p_{ji}, \cdots, p_{(n-1)i}\}, i \neq j. \tag{6}$$

The relativity histogram of *i*-state consists of two parts, the relativity of *i*-state with the other states and the other states with *i*-state. For simplification, the relativity histogram can be replaced by the following equation.

$$\boldsymbol{H}_i = \{h_1, h_2, \cdots, h_k, \cdots, h_{2n-2}\}. \tag{7}$$

where $h_1 = p_{i0}, h_2 = p_{i1}, \cdots,$ and $h_{2n-2} = p_{(n-1)i}$.

Though $\boldsymbol{H}_i$ can be considered as the relativity feature of *i*-state with the other states, the dimension is correspondingly high. If the number of state is n, the dimension is $(2n-2)$. In order to reduce the dimension, we proposed using entropy which was developed by Shannon in [6] to represent the relativity between the states.

Based on the relativity histogram and the definition of entropy, we propose a new descriptor, CCRE, describing the relativity. The CCRE of *i*-state is defined as

$$E_i = -\sum_{k=1}^{2n-2} h_k \log_2(h_k). \tag{8}$$

E_i gives the relativity between *i*-state and the other states, that is to say, gives the relativity between *i*-direction and the other directions in a Freeman chain code. Large E_i means the relativity between the states is large, otherwise the relativity is low.

3.3 Characters of CCRE

Now, let's analyze the character of E_i. It is obvious that E_i is translation invariant because it reflects the statistical feature. Fig. 3 gives an example contour which has different scaling with Fig. 2. The chain code of Fig. 3 is {1100110033332222}. The transition-number matrixes for Fig.2 and Fig. 3 are shown in Fig. 4 (a) and (b).

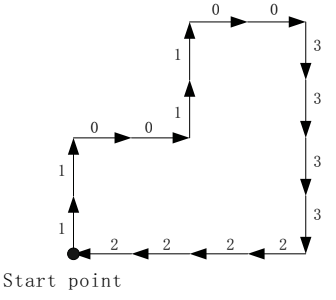


Fig. 3. Example of a different scaling contour with Fig.2

State	0	1	2	3
0	0	1	0	1
1	2	0	0	0
2	0	0	1	0
3	0	0	1	1

(a)

State	0	1	2	3
0	2	1	0	1
1	2	2	0	0
2	0	0	3	0
3	0	0	1	3

(b)

State	0	1	2	3
0	0	1	0	1
1	2	0	0	0
2	0	0	0	0
3	0	0	1	0

(c)

Fig. 4. The transition-number matrixes. (a) Transition-number matrix of Fig.2, (b) Transition-number matrix of Fig.3, (c) Transition-number matrix of different states of Fig.2 and Fig.3.

Fig. 4 (a) and (b) show that Fig. 2 and Fig. 3 have different transition-number matrix. The reason is that their scaling is different. However, if we do not consider the transition of a state to itself, that is to say, if the diagonal transition numbers in the transition-number matrix are zero, Fig. 2 and Fig. 3 have the same transition-number matrix. The result is shown in Fig. 4 (c). That's the reason why we suppose $p_{ij} = 0$ when $i = j$ in section 3.1. On the other hand, the feature of a contour is mainly reflected by the variation of different directions. That's the second reason why we don't

discuss the relativity of the same direction in this paper. By this method, E_i is scale invariant.

When using chain code, the location of the start point is always pivotal. The same is to transition-probability matrix. This problem can be solved if we think of the chain code as an annular chain. Using Fig. 2 as an example, we give its annular chain in Fig.5. It is obvious that the transition numbers are irrelative with the location of the start point if we consider the variation of the last state with the first one.

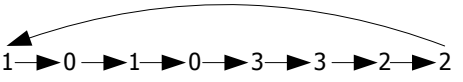


Fig. 5. Annular chain codes

For rotation invariance, we can recur to MSSDC (see [4]).

3.4 Improved CCRE

Because of the symmetrical property, dissimilar relativity histogram may have the same entropy. Three improved methods were given to avoid such influence in [7]. Histogram area is adopted in this paper.

The area of *i*-state relativity histogram $\boldsymbol{H}_i$ is defined as

$$A_{\boldsymbol{H}_i} = \sum_{k=1}^{2n-2} (h_k \times k) \,. \tag{9}$$

It is clear that $A_{\boldsymbol{H}_i}$ get maximum $2n-2$ when $h_{2n-2} = 1$. In generic, histograms with similar probability distribution have dissimilar area if their dimension orders are different. By the area, a weight function was given as

$$f(\boldsymbol{H}) = 1 + A_{\boldsymbol{H}_i} \big/ (2n-2) \,. \tag{10}$$

As discussed in [7], the entry of *i*-state is $(q_i, E_i, A_{\boldsymbol{H}_i})$. Then the index of a contour can be written as

$$((q_0, E_0, A_{\boldsymbol{H}_0}), (q_1, E_1, A_{\boldsymbol{H}_1}), \cdots, (q_i, E_i, A_{\boldsymbol{H}_i}), \cdots, (q_{n-1}, E_{n-1}, A_{\boldsymbol{H}_{n-1}})) \,. \tag{11}$$

where q_i is the CCH of *i*-direction, E_i is its CCRE , $A_{\boldsymbol{H}_i}$ is its area and n is the number of directions.

4 Similarity Measurement

Suppose $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be two contours, $E_i^{\boldsymbol{c}_1}$ and $E_i^{\boldsymbol{c}_2}$ be their CCRE of *i*-direction, $(E_i^{\boldsymbol{c}_1})'$ ($(E_i^{\boldsymbol{c}_1})' = f(\boldsymbol{H}_i^{\boldsymbol{c}_1}) \times E_i^{\boldsymbol{c}_1}$) and $(E_i^{\boldsymbol{c}_2})'$ ($(E_i^{\boldsymbol{c}_2})' = f(\boldsymbol{H}_i^{\boldsymbol{c}_2}) \times E_i^{\boldsymbol{c}_2}$) be the improved

entropy by histogram area. With reference to the definition of similarity metric Formula (5) in [8], the distance of these two contours can be noted as

$$S(\boldsymbol{c}_1,\boldsymbol{c}_2)=\begin{cases}\sum_{i=0}^{n-1}\min(q_i^{\boldsymbol{c}_1},q_i^{\boldsymbol{c}_2})\times\dfrac{\min((E_i^{\boldsymbol{c}_1})',(E_i^{\boldsymbol{c}_2})')}{\max((E_i^{\boldsymbol{c}_1})',(E_i^{\boldsymbol{c}_2})')}, & if\ \left|E_i^{\boldsymbol{c}_1}-E_i^{\boldsymbol{c}_2}\right|<\alpha\\ \sum_{i=0}^{n-1}\min(q_i^{\boldsymbol{c}_1},q_i^{\boldsymbol{c}_2})\times\dfrac{\min(E_i^{\boldsymbol{c}_1},E_i^{\boldsymbol{c}_2})}{\max(E_i^{\boldsymbol{c}_1},E_i^{\boldsymbol{c}_2})}, & \text{others}\end{cases}. \tag{12}$$

where α is a threshold. If $\left|E_i^{\boldsymbol{c}_1}-E_i^{\boldsymbol{c}_2}\right|<\alpha$, it means that their entropy is similar. However, their relativity histograms may be dissimilar because of the symmetrical property of entropy. In this occasion, CCRE is replaced by the improved CCRE.

With the same meaning as the formula (5) defined in [8], the similarity metric is made up of two parts: the first part, $\min(q_i^{\boldsymbol{c}_1},q_i^{\boldsymbol{c}_2})$, is histogram intersection which measures the similarity between $q_i^{\boldsymbol{c}_1}$ and $q_i^{\boldsymbol{c}_2}$, while the second part $\dfrac{\min(E_i^{\boldsymbol{c}_1},E_i^{\boldsymbol{c}_2})}{\max(E_i^{\boldsymbol{c}_1},E_i^{\boldsymbol{c}_2})}$ or $\dfrac{\min((E_i^{\boldsymbol{c}_1})',(E_i^{\boldsymbol{c}_2})')}{\max((E_i^{\boldsymbol{c}_1})',(E_i^{\boldsymbol{c}_2})')}$ measures the similarity of the relativity of a direction.

5 Experiments

The SQUID database [9] is used to perform experiments and test the proposed method. This database consists of 1100 fish shapes and has been extensively used by researchers for testing. Indexing and retrieval experiments are performed on the database for the CCH, MSSDC, CCDV and CCRE.

The retrieval accuracy was measured in terms of recall, precision and *ANMRR*. The precision rate and recall rate are defined as follows

$$\text{precision}=\frac{\text{number of relevent images selected}}{\text{total number of retrieved images}}. \tag{13}$$

$$\text{recall}=\frac{\text{number of relevent images selected}}{\text{total number of similar images}}. \tag{14}$$

To measure the effectiveness of the methods, we also use the MPEG-7 retrieval metric the Average Normalized Modified Retrieval Rank (ANMRR). The ANMRR indicates not only how many of the correct items are retrieved, but also how highly they are ranked among the retrieved items. It is defined in [10] as

$$ANMRR=\frac{1}{NQ}\sum_{q=1}^{NQ}\frac{\dfrac{1}{NG(q)}\sum_{k=1}^{NG(q)}Rank^{*}(k)-0.5\times[1+NG(q)]}{1.25K(q)-0.5\times[1+NG(q)]}. \tag{15}$$

where $Rank^*(k) = \begin{cases} Rank(k), & \text{if } Rank(k) \le K(q) \\ 1.25K, & \text{if } Rank(k) > K(q) \end{cases}$. $NG(q)$ is the size of the ground truth set for a query image q, $Rank(k)$ is the ranking of the ground truth images by the retrieval algorithm, NQ is number of query images, and $K(q)$ specifies the "relevant ranks" for each query. As the size of the ground truth set is normally unequal, a suitable $K(q)$ is determined by

$$K(q) = \min(4 * NG(q), 2 * GTM) . \tag{16}$$

where GTM is the maximum of $NG(q)$ for all queries.

In order to evaluate the performance of the proposed method, CCRE were compared with CCH, CCDV and MSSDC. Fifty independent query images which belong to ten types were chosen from the database. Fig. 6 presents the precision-recall graph of the results averaged over 50 queries. Table 1 gives a comparison of the four methods by *ANMRR*. The retrieval results show that the proposed method, CCRE gives better performance in image retrieval than CCH, CCDV and MSSDC.

Table 1. Results of ANMRR for the four methods

methods	ANMRR
CCRE	0.218
CCDV	0.246
MSSDC	0.312
CCH	0.368

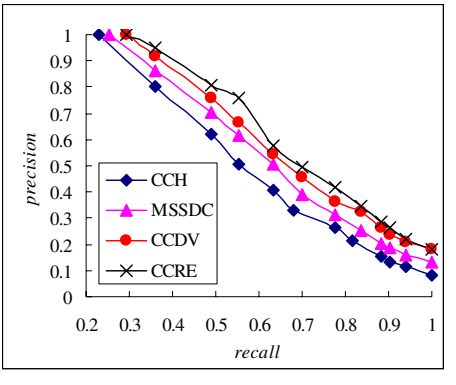


Fig. 6. Average precision-recall of the 4 methods, CCH, MSSDC, CCDV and CCRE

6 Conclusions

In this paper a novel method, CCRE, was presented based on Freeman chain code to describe the relativity between the chain codes. The CCRE is robust to translation,

scaling and rotation. It takes not only the statistical feature but also the relativity between the chain codes into account. The retrieval performance was tested and compared with several other feature descriptors using SQUID image databases. The results in all cases show the CCRE to be quite efficient.

Acknowledgements. The authors wish to thank the image database of SQUID, Henan Polytechnic University for the Dr. Grant B050901 and the support of Provincial Key Laboratory of Computer Information Processing Technology of Suzhou University.

References

1. Zhang, D.S., Lu, G.J.: Review of shape representation and description techniques. Pattern Recognition 1, 1–19 (2004)
2. Freeman, H.: On the encoding of arbitrary geometric configurations. IRE Trans. on Electronic Computers, pp. 260–268 (1961)
3. Iivarinen, J., Visa, A.: Shape recognition of irregular objects. In: Intelligent Robots and Computer Vision XV: Algorithms, Techniques, Active Vision, and Materials Handling, SPIE, pp. 25–32 (1996)
4. Wang, X.L., Xie, K.L.: A novel direction chain code-based image retrieval. In: Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04) pp. 190–193 (2004)
5. Sun, J.D., Wu, X.S.: Chain code distribution-based image retrieval. In: Intelligent Information Hiding and Multimedia Signal Processing, pp. 139–142 (2006)
6. Shannon, C.E.: A Mathematical Theory of Communication. Bell Systems Technical Journal 3, 379–423 (1948)
7. Sun, J.D.: Image retrieval based on improved entropy and moments. Intelligent Information Hiding and Multimedia Signal Processing, 347–350 (2006)
8. Lim, S., Lu, G.J.: Spatial statistics for content based image retrieval. In: International Conference on Information Technology: Computers and Communications, pp. 28–30 (2003)
9. SQUID: http://www.ee.surrey.ac.uk/Research/VSSP/imagedb
10. Manjunath, B.S., Ohm, J.R., Vasudvan, V., Yamada, A.: Color and texture descriptors. IEEE Transactions on Circuits and Systems for Video Technique 6, 703–715 (2001)

3D Motion Segmentation from Straight-Line Optical Flow

Jing Zhang[1], Fanhuai Shi[2], Jianhua Wang[1], and Yuncai Liu[1]

[1] Inst. Image Processing and Pattern Recognition,
[2] School of Materials Science and Engineering,
Shanghai Jiao Tong University,
Shanghai 200240, P.R. China
{zhjseraph,fhshi,whomliu}@sjtu.edu.cn

Abstract. We present a closed form solution to the problem of segmenting multiple 3D motion models from straight-line optical flow. We introduce the multibody line optical flow constraint(MLOFC), a polynomial equation relating motion models and line parameters. We show that the motion models can be obtained analytically as the derivative of the MLOFC at the corresponding line measurement, without knowing the motion model associated with that line. Experiments on real and synthetic sequences are also presented.

1 Introduction

Motion segmentation is a fundamental problem in many applications in computer vision, such as traffic surveillance, recognition of human gaits, etc.And recently, there has been an increasing interest on the algebraic, geometrical and statistical models to resolve this problem.

Existing approaches on 3-D motion segmentation from two views include different variations of the Expectation Maximization (EM) algorithm [1,2,3]. But EM is very sensitive to initialization [4].

Algebraic approaches, which are based on polynomial and tensor factorization, have been proposed [5]. In [6],Vidal and Ma viewed the estimation of multiple motion models as the estimation of a single multibody motion model and proposed a unified algebraic approach to 2-D and 3-D motion segmentation from two-view correspondences or optical flow, which applies to most of the two-view motion models adopted in computer vision. In [7,8], they introduce the multibody brightness constancy constraint(MBCC) and presented a closed form solution to direct motion segmentation.

Other related works include [9], which considers the problem of modeling a scene containing multiple dynamic textures undergoing multiple rigid-body motions. And in [10], X.Fan and R.Vidal study the rank and geometry of the multibody fundamental matrix.

In addition, efforts for the robustness of multibody motion segmentation include those works that uses RANSAC with priors to do articulated motion segmentation [11]. L. Hajder et.al. [12] consider robust 3D segmentation of multiple moving objects under weak perspective.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 85–94, 2007.
© Springer-Verlag Berlin Heidelberg 2007

However, all the previous methods dealing with 3-D motion segmentation use point correspondences. Generally, the choice of types of features depends on their availability in the images and the reliability of their measurement. When points are not available in large quantities, all the previous methods will have problems. Image line is also a common feature used in computer vision and was successfully used for 3-D motion estimation from 1980s to the latest. So, it is significant to consider the problem of 3-D motion segmentation from line correspondences. In [13], Shi et.al proposed a method to do motion segmentation using line correspondences. But they only dealt with translating objects. In [14],they proposed a general method for 3-D motion using line correspondences, but this method requires too many features when the number of motion is large.

1.1 Contributions of This Paper

In this paper, we proposed a novel method to segment multiple rigid-body 3D motions from the optical flow of line correspondences in two views. The remainder of this paper decomposes as follows. In Section 2, we introduce the concept of line optical flow. In Section 3, we show now to segment the 3D motion using the line optical flow and the method of GPCA [5]. Finally, Section 4 validates our algorithm by experiments with simulated data and real scenes.

2 Line Optical Flow [15]

A line l in the image plane is represented by a vector $m = (m_x, m_y, m_z)^T$ giving its equation:

$$m_x x + m_y y + m_z f = 0 \tag{1}$$

where f is the focal length. The interpretation of m is that it is the normal to the plane defined by the 2D line and the optical center of the camera(see Figure 1); note that this plane contain also the corresponding 3D line, L. We use two relationships between the representation of a 3D line and its image in the image plane. The first relation is that the direction v of the 3D line is perpendicular to the normal m to the plane it defines with the optical center.

$$m^T v = 0 \tag{2}$$

Then we take its time derivative and obtain

$$g^T v + m^T \partial v / \partial t = 0 \tag{3}$$

where $g = (\partial m_x/\partial t, \partial m_y/\partial t, \partial m_z/\partial t)^T$. We define g as the *line optical flow*. We assume that the 3D line under consideration is attached to a rigid body whose motion is described by its instantaneous angular velocity $W = (\omega_1, \omega_2, \omega_3)^T$ and linear velocity $T = (t_1, t_2, t_3)^T$, its kinematic screw is at the origin o.Then

$$\partial v / \partial t = W \times v \tag{4}$$

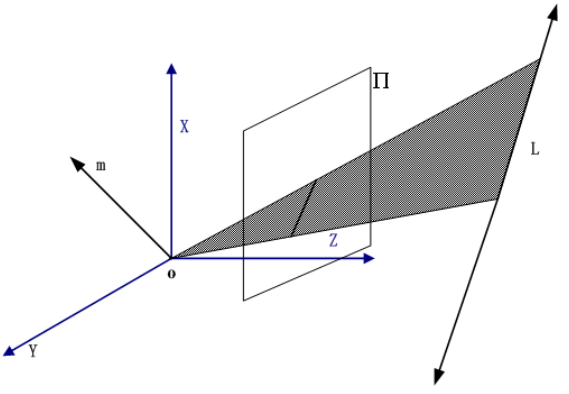


Fig. 1. The vector m

$$A = \begin{vmatrix} m_x & m_y & m_z \\ \omega_3 m_y - \omega_2 m_z + \Delta m_x & \omega_1 m_z - \omega_3 m_x + \Delta m_y & \omega_2 m_x - \omega_1 m_y + \Delta m_z \end{vmatrix} \tag{7}$$

replacing $\partial v/\partial t$ by this value, (3)give us the second relationship between the representation of a 3D line and its image

$$g^T v + m^T (W \times v) = 0 \tag{5}$$

From equation (5) and (2), one can get the linear constraints about v

$$\begin{cases} g^T v + m^T (W \times v) = 0 \\ m^T v = 0 \end{cases} \tag{6}$$

Let A is the coefficient matrix of the two equations(A is defined in equation(7)), then $R(A) = 1$ [16]. So,

$$\frac{m_x}{\omega_3 m_y - \omega_2 m_z + \Delta m_x} = \frac{m_y}{\omega_1 m_z - \omega_3 m_x + \Delta m_y} = \frac{m_z}{\omega_2 m_x - \omega_1 m_y + \Delta m_z} \tag{8}$$

Let $k = -\frac{m_x}{m_y}$ is the slope of image line and $b = -\frac{m_z f}{m_y}$is the intercept. Then one can get

$$\begin{cases} kb\omega_3 - b^2\omega_1/f - k\omega_2 \cdot f - \omega_1 \cdot f = -\Delta b \\ (1+k^2)\omega_3 + b\omega_2/f - kb\omega_1/f = -\Delta k \end{cases} \tag{9}$$

Let $\omega_1/f = \omega_{1f}, \omega_2/f = \omega_{2f}, \omega_1 \cdot f = \omega_{1F}, \omega_2 \cdot f = \omega_{2F}$,then

$$\begin{cases} kb\omega_3 - b^2\omega_{1f} - k\omega_{2F} - \omega_{1F} = -\Delta b \\ (1+k^2)\omega_3 + b\omega_{2f} - kb\omega_{1f} = -\Delta k \end{cases} \tag{10}$$

Let

$$y(k,b) = \begin{bmatrix} kb & -b^2 & 0 & -1 & -k & \Delta b \\ (1+k^2) & -kb & b & 0 & 0 & \Delta k \end{bmatrix}^T \tag{11}$$

and

$$u = (\omega_3, \omega_{1f}, \omega_{2f}, \omega_{1F}, \omega_{2F}, 1)^T \tag{12}$$

Equation (10) can be expressed as

$$y(k,b)^T u = \mathbf{0} \tag{13}$$

As each line-line correspondence gives 2 independent equations, one may estimate u linearly from 3 correspondences for one motion.

3 Direct Motion Segmentation Using MLOFC

3.1 Multibody Line Optical Flow Constraint(MLOFC)

Now consider a scene with n rigid-body motions (n is known) with associated motion component $u_i \in \mathbb{R}^6$, where u_i is the motion component associated with the motion of the i^{th} object.

Therefore, we define the *multibody line optical flow constraint*(MLOFC)

$$MLOFC = \prod_{i=1}^{n} (y(k,b)^T u_i) = \mathbf{0} \tag{14}$$

MLOFC is a homogeneous polynomial of degree n in y, which can be written as a linear combination of the monomials $y_1^{n_1} y_2^{n_2} \cdots y_6^{n_6}$ with $n_1 + n_2 + \cdots + n_6 = n$. If we stack these $M_n^6 = \binom{n+6-1}{6-1}$ independent monomials into a vector $\nu_n(y) \in \mathbb{R}^{M_n^6}$,we get

$$MLOFC = \nu_n(y)^T \mathcal{U} = \sum \mathcal{U}_{n_1,n_2,\cdots,n_6} y_1^{n_1} y_2^{n_2} \cdots y_6^{n_6} \tag{15}$$

The vector $\mathcal{U} \in \mathbb{R}^{M_n^6}$is called *multibody line optical flow*, and $\nu_n : \mathbb{R}^6 \rightarrowtail \mathbb{R}^{M_n^6}$ is called the *Veronese* map of degree n.

In the following subsections, we will demonstrate that in the case of 3-D rigid motion models, the MLOFC can be expressed linearly in terms of a multibody motion model $\mathcal{W}$. By exploiting the algebraic properties of $\mathcal{W}$, we will derive an algebraic closed form solution to the following problem [8]:

Problem 1 (Direct multiple-motion segmentation from line optical flow). *Given the slopes and intercepts of line correspondences of a motion sequence generated from n 3-D rigid motion models, estimate the model parameters* $\{\mathcal{W}_i\}_{i=1}^n$*, without knowing which image measurements correspond to which model.*

3.2 Computing the Multibody Motion Model

Let $l(k,b) \leftrightarrow l'(k',b')$ be an arbitrary line-line correspondence associated with any of the n motions in two consecutive image frames. We may estimate $\mathcal{U}$ linearly from $M_n^6/2$ correspondences, using equations (15).

But note that the entries (1,3),(2,4) and (2,5)of each y_i are zero, $y_1^{n_1} y_2^{n_2} \cdots y_6^{n_6} = 0$ in both equations of (15) when $(n_3 \neq 0 \quad and \quad n_4 \neq 0)$ or$(n_3 \neq 0 \quad and \quad n_5 \neq 0)$. After enforcing these equations we obtain

$$\sum \mathcal{Y}_n^T \widetilde{\mathcal{U}} = \mathbf{0} \tag{16}$$

where $\mathcal{Y} \in \mathbb{R}^{M_n^6 - Z_n}, \widetilde{\mathcal{U}} \in \mathbb{R}^{M_n^6 - Z_n}$ is equal to $\nu_n(y), \mathcal{U}$ without the both zero entries in two equations of (15).When $n = 2$, then $Z_n = 2$; when $n > 2$, then $Z_n = 3M_{n-2}^3 + \sum_{i=1}^{n-1} M_{n-2-i}^3 \cdot M_i^3 + \sum_{i=1}^{n-1} M_{n-2-i}^3 \cdot M_i^2 + M_{n-1}^2$.

As one line-line correspondence gives 2 independent equations(16), we may estimate $\widetilde{\mathcal{U}}$ linearly from $(M_n^6 - Z_n)/2$ correspondence.

3.3 Estimate f

After the estimation of $\widetilde{\mathcal{U}}$, we compute focal length f

$$f^{2n} = \frac{\prod_{i=1}^n w_{1F}^i}{\prod_{i=1}^n w_{1f}^i} = \frac{\prod_{i=1}^n w_{2F}^i}{\prod_{i=1}^n w_{2f}^i} \tag{17}$$

$$f^2 = \frac{\sum_{i=1}^n w_{1F}^i}{\sum_{i=1}^n w_{1f}^i} = \frac{\sum_{i=1}^n w_{2F}^i}{\sum_{i=1}^n w_{2f}^i} \tag{18}$$

where w^i is the parameter of the ith motion models. And

$$\prod_{i=1}^n w_{1F}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + M_{n-1}^4 + 1}, \quad \prod_{i=1}^n w_{1f}^i = \mathcal{U}_{M_{n-1}^6 + 1} \tag{19}$$

$$\prod_{i=1}^n w_{1F}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + M_{n-1}^4 + M_{n-1}^3 + 1}, \quad \prod_{i=1}^n w_{2f}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + 1} \tag{20}$$

$$\sum_{i=1}^n w_{1F}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + M_{n-1}^4 + M_{n-1}^3}, \quad \sum_{i=1}^n w_{1f}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5} \tag{21}$$

$$\sum_{i=1}^n w_{2F}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + M_{n-1}^4 + M_{n-1}^3 + M_{n-1}^2}, \quad \sum_{i=1}^n w_{2f}^i = \mathcal{U}_{M_{n-1}^6 + M_{n-1}^5 + M_{n-1}^4} \tag{22}$$

In our algorithm, we estimate f as the average value of (17)(18).

3.4 Refined MLOFC

After the estimation of f, we let

$$y'(k,b) = \begin{bmatrix} -(\frac{b^2}{f} + f) & -kf & kb & \triangle b \\ -\frac{kb}{f} & \frac{b}{f} & (1+k^2) & \triangle k \end{bmatrix}^T \tag{23}$$

$$u' = [w_1, w_2, w_3, 1]^T \tag{24}$$

MLOFC(15) can be refined as

$$MLOFC' = \nu_n(y')^T \mathcal{U}' = \sum \mathcal{U}'_{n_1,n_2,n_3,n_4} {y'_1}^{n_1} {y'_2}^{n_2} {y'_3}^{n_3} {y'_4}^{n_4} \tag{25}$$

Then, we can solve for the multibody motion model $\mathcal{W} = \mathcal{U}'$ uniquely from (25).

3.5 Segmenting the Multibody Model

Given $\mathcal{U}'$, we can use the algorithm in [6] to compute the parameters u'_i of each individual motion model from the derivatives of $MLOFC'$, i.e.,

$$D_{MLOFC'} = \frac{\partial MLOFC'}{\partial y'} = \sum_{i=1}^{n} \prod_{l \neq i} ({y'}^T u'_l)) u'_i \tag{26}$$

If we evaluate $D_{MLOFC'}$ at a line $y = z_i$ that correspondence to the *i*th motion, i.e.if z_i is such that $z_i^T u'_i = 0$, then we have $D_{MLOFC'} \sim u'_i$. Therefore, given $\mathcal{U}'$ we can obtain the motion parameters as

$$u'_i = \frac{D_{MLOFC'}}{e_K^T D_{MLOFC'}}|_{y=z_i} \tag{27}$$

where $e_K = [0, \ldots, 0, 1]^T \in \mathbb{C}^K$ and $z_i \in \mathbb{C}^K$ is a nonzero vector that $z_i^T u'_i = 0$.

Once the right individual motion parameters have been computed, one may compute the Sampson error by assigning each feature to $\{u'_i\}_{i=1}^n$, and cluster the correspondence to the one that minimize the error.

We summarize the whole procedure in Algorithm 1.

Algorithm 1. (3D Motion segmentation using Straight-line optical flow)
Given $N > (M_n^6 - Z_n)/2$ line correspondence $\mathcal{L} = \{l_i(k,b) \leftrightarrow l'_i(k',b')\}_{i=1}^N$ in two consecutive image frames:

1. Construct Multibody Line Optical Flow Constraint $MLOFC = \prod_{i=1}^n (y(k,b)^T u_i) = \mathbf{0}$;
2. Compute $\widetilde{\mathcal{U}}$ using equation (16);
3. Estimate f using equations (17)(18);
4. Construct refined Multibody Line Optical Flow Constraint $MLOFC' = \nu_n(y')^T \mathcal{U}'$;
5. for $i = n : 1$

$$z_i = \arg\min_{z \in \mathcal{L}} \frac{\frac{|MLOFC'|}{\|\Pi D_{MLOFC'}\|} + \delta}{\frac{|z^T u'_{i+1}| \cdots |z^T u'_n|}{\|\Pi u'_{i+1}\| \cdots \|\Pi u'_n\|} + \delta};$$

where $\delta > 0$ is a small positive number,$\Pi = [I_{K-1} \quad 0] \in \mathbb{R}^{(K-1)\times K}$

$$u'_i = \frac{D_{MLOFC'}}{e_K^T D_{MLOFC'}},$$

end

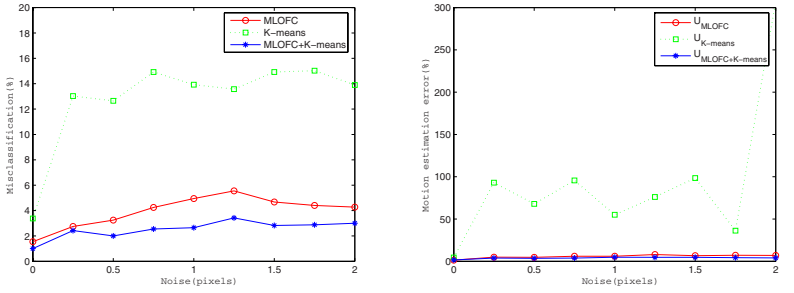


Fig. 2. Correctness of motion segmentation and motion estimation ($\{u_i'\}_{i=1}^2$) as a function of noise. The number of line correspondences of each object is set to be 20.

4 Experiments on Synthetic and Real Images

In this section, we evaluate our motion segmentation algorithms on both real and synthetic data. We compare our results with K-means methods and use our algorithms to initialize iterative technique. We consider the following algorithms:

1. MLOFC: Using Algorithms 1 to cluster the line correspondences.
2. K-means [17]: this algorithm alternates between computing(linearly) the parameters of different motion classes and clustering the line correspondences using the Sampons-distance error to the different motions.

4.1 Synthetic Experiments

First, we conduct some simulated experiments to check the correctness of the proposed algorithm with respect to the amount of noise in the image measurements for $n = 2$ and $n = 3$. We randomly generated n groups of $N = n \times m$ 3-D lines, each group has m lines. Then the n groups of 3-D lines are projected in two views. In the presence of noise, we added zero-mean Gaussian noise with standard deviation between 0 and 2 pixels in an image size of 1000×1000.

To show the general sensitivity to noise, the average percentage of misclassified correspondences, i.e., the percent of wrong classified line number in whole set and the average relative errors in the estimation of $\{u_i'\}_{i=1}^n$ are recorded. We first run K-means algorithm starting from random classification. The K-means algorithm converges to a local minimum due to bad initialization. Running the K-means algorithm starting from the clustering produced by the MLOFC algorithm result in a better result. The results when $n = 2$ are showed in Fig.2 and the results when $n = 3$ are showed in Fig.5.

During the procedure of segmentation using MLOFC, we didn't consider the influence of linear velocity $T = (t1, t2, t3)^T$. We treat it as noise. In the second simulated experiment, we evaluate the performance of the algorithms with respect to linear velocity. The linear velocity of the two motion objects are t times as large as the angular

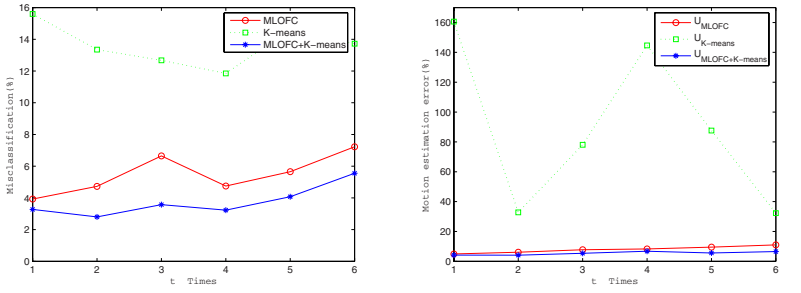


Fig. 3. Correctness of motion segmentation and motion estimation ($\{u_i'\}_{i=1}^2$) as a function of linear velocity. The image noise is set to be 1.0 pixel. The number of line correspondences of each object is set to be 20.

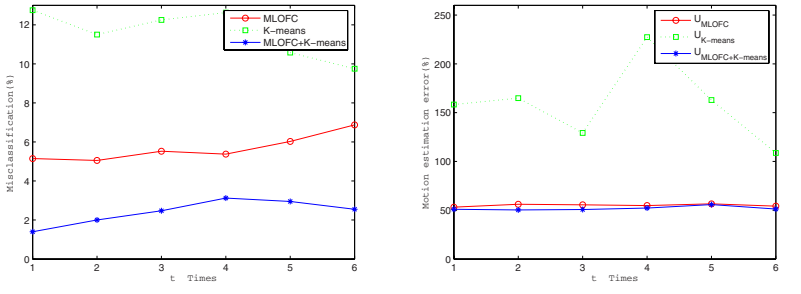


Fig. 4. Correctness of motion segmentation and motion estimation ($\{u_i'\}_{i=1}^2$) as a function of linear velocity, when the second object take pure translating motion. The image noise is set to be 1.0 pixel. The number of line correspondences of each object is set to be 20.

velocity. The result is showed in Fig.3. One can see that with the increasing of linear velocity, the correctness of motion segmentation is decrease. But the result is still good.

Then, we conduct another experiment to test the performance to pure translating motion. In this experiment, we set the angular velocity of second object as zero, i.e., this object take a pure translating motion. The linear velocity of the two objects are same. It is t times as large as the first object's angular velocity. The result is showed in Fig.4. One can see that the performance is good, although we did not compute the linear velocity.

4.2 Real Experiment

Figure 6(a) shows one frame of a 640×560 sequence taken by a static camera observing two moving boxes using different 3D rigid motions. Figure 6(b) shows the lines that detected by *Edge linking and line segment fitting* algorithm that developed by *Peter*

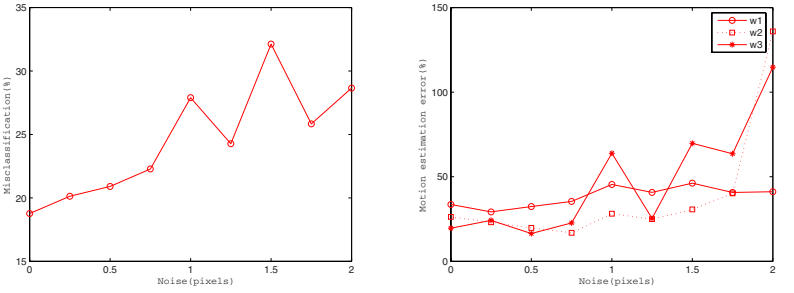
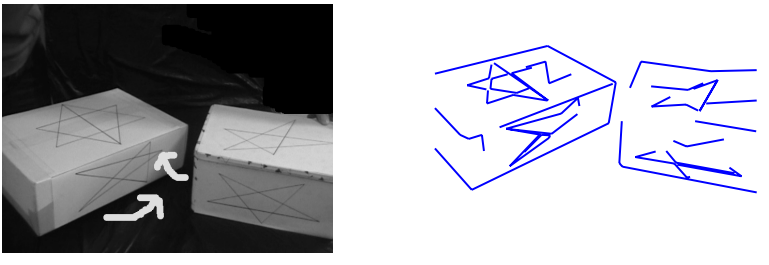


Fig. 5. Correctness of motion segmentation and motion parameters ($\{w_i\}_{i=1}^{3}$) as a function of noise. The number of line correspondences of each object is set to be 30.



(a) (b)

Fig. 6. Line segmentations

Kovesi. We manually select $m = 10$ line correspondences in two consecutive frames. And the average misclassification error using MLOFC is 17.197% for 10 pairs of consecutive frames. This result is not very satisfied and may due to the camera distortion and the insufficient of line correspondences. But may be improved by other optimization method, such as EM [17].

5 Conclusion

We present a closed form solution to the problem of segmenting multiple 3D motion models from straight-line optical flow. The algebraic method of motion classification involves computation of MLOFC and individual angular velocity. Our approach has the advantage that it provides a global, non-iterative solution, which was able to provide an initial classification for other optimization segmentation method. Experiments with simulated data and real scenes validate our algorithm. But we have to admit that this method can not deal with motions that all take the same angular velocity. So our future work is to develop a general algorithm to all kinds of 3D rigid motions.

References

1. Rittscher, J., Tu, P.H., Krahnstoever, N.: Simultaneous estimation of segmentation and shape. In: IEEE Conference on Computr Vision and Pattern Recognition (2005)
2. Torr, P.H.S.: Geometric motion segmentation and model selection Phil. Trans. Royal Society of London A (1998)
3. Feng, X., Perona, P.: Scene segmentation from 3d motion (1998)
4. Torr, P., Szeliski, R., Anandan, P.: An integrated bayesian approach to layer extraction from image sequences. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 297–303 (2001)
5. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (gpca). In: IEEE Conference on Computr Vision and Pattern Recognition, pp. 621–628 (2003)
6. Vidal, R., Ma, Y.: A unified algebraic approach to 2-d and 3-d motion segmentation. In: European Conference on Computer Vision, pp. 1–15 (2004)
7. Singaraju, D., Vidal, R.: A bottom up algebraic approach to motion segmentation. In: Asian Conference on Computer Vision, pp. 286–296 (2006)
8. Vidal, R., Singaraju, D.: A closed-form solution to direct motion segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 510–515 (2005)
9. Vidal, R., Ravichandran, A.: Optical flow estimation and segmentation of multiple moving dynamic textures. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
10. Fan, X., Vidal, R.: The space of multibody fundamental matrices: Rank, geometry and projection. In: IEEE International Conference on Computer Vision (2005)
11. Yan, J., Pollefeys, M.: Articulated motion segmentation using ransac with priors. In: IEEE International Conference on Computer Vision (2005)
12. Hajder, L., Chetverikov, D.: Robust 3d segmentation of multiple moving objects underweak perspective. In: IEEE International Conference on Computer Vision (2005)
13. Shi, F., Wang, J., Zhang, J., Liu, Y.: Motion segmentation of multiple translating objects using line correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 315–320 (2005)
14. Zhang, J., Shi, F., Liu, Y.: Motion segmentation by multibody trifocal tensor using line correspondence. In: International Conference on Pattern Recognition (2006)
15. Faugeras, O., Navab, N., Deriche, R.: On the information contained in the motion field of lines and the cooperation between motion and stereo. International Journal of Imaging Systems and Technology (1991)
16. Jiang, S.F., Chen, Z., Yang, S.H.: Reconstructing 3d rotation motion parameters from 2d image straight-line optical flow. Journal of Nanchang Institute of Aeronautical Technology(Natural Science) (2005)
17. Hartley, R., Vidal, R.: The multibody trifocal tensor: Motion segmentation from 3 perspective views. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)

A Fast Adaptive ME Algorithm Based on H.264

Zhisheng Liu and Yuhua Peng

School of information science and engineering, Shandong University, Jinan,250100
liuzhisheng@mail.sdu.edu.cn

Abstract. This paper proposes a new fast adaptive motion estimation algorithm based on H.264. It utilizes the statistics of video sequence and the correlations of SADs in the space to early terminate the search, determines the search pattern according to the block's local motion activity, and proposes a new synthetic diamond search pattern. Compared to fast full search, the PSNR is just 0.017db lower and bitrates is 0.45% higher while the speed is 15 times up.

Keywords: fast adaptive motion estimation, block matching algorithm, stationary block.

I Introduction

The H.264/AVC standard is the latest coding standard developed by the Joint Video Team, formed by MPEG and VCEG. It is designed for a wide range of applications including low bit rate wireless video applications, SDTV&HDTV and video streaming over the Internet. The H.264/AVC standard adopts a lot of new techniques such as variable block size motion compensation, quarter-pixel accuracy motion compensation, and multiple reference frame motion compensation. In terms of compression, it provides more than 50% bit rate savings with equivalent video quality compared to the performance of MPEG4 baseline and H.263++. Because of its new techniques, the encoder's computational complexity is much increased.

Motion estimation is the primary important component in the video encoder, directly affecting the encoding speed and image quality. In H.264/AVC standard, when all of the block sizes are considered and the number of reference frame is 5 with the search window 32 using full search pattern, motion estimation consumes more than 90% of the entire encoding time. Various fast algorithms were proposed to reduce the computational complexity by reducing the num of candidate motion vectors while keeping the similar image quality to full search. The two typical methods are three-step search [1] and 2D-LOG search [2]. These algorithms limit search steps, employ rectangular search pattern and improve speed search, but the image quality is not very good. Recently, some novel algorithms were proposed such as diamond search[3],[4], cross-diamond search[5], hexagon-based search[6] and a lot of transformations of them. The UMHexagonS[7] algorithm can keep all most the same image quality with full search and is used as a reference algorithm in the JM software. Also there are many methods mainly pay attention to multireference motion estimation[8][9].

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 95 – 101, 2007.
© Springer-Verlag Berlin Heidelberg 2007

[10] proposed a FAME algorithm based on MPEG4.It utilize the technologies such as local motion activity and early terminate the search and works well on various video sequences, but it may give rise to very large motion activities.

In this article, we extend it to H.264 and propose a new algorithm based on diamond search. We use the stationary block detection and early termination to improve the speed. The LMA(local motion activity)denotes the motion activity of the current block. When the LMA is low, we use the small diamond, cross diamond and large diamond search. If the LMA is high, we propose a synthetic diamond search pattern which is better than hexagon-based search.

The rest of the paper is organized as follows. Section II describes the fast adaptive algorithm. In section III, we give the experiment results. Comparison with other algorithms is also shown in this section. We discuss about some parameters in our method in section IV. Finally, conclusion is made in section V.

2 Fast Adaptive Motion Estimation Algorithm

2.1 The Initial Search Center

There exist two selections of the initial search center. The first is just choosing the position (0, 0) as the search center in the reference frame. Although it is simple, this method is easy to be trapped in local optimization. When the global optimization is not (0, 0) and the first search step is long, it can lead to search the point that is far away from the center. Another is to predict the initial search center. There are strong correlations between adjacent frames and adjacent blocks, so many algorithms utilize this to predict the initial search center.

In this article, we choose the second one also consider the position (0, 0).

2.2 Stationary Block Detection

Experiment shows that about 98% blocks has their SAD at position (0, 0) less than 512 for MB size of 16×16[11], [12]. If the block's SAD is less than 512 at (0, 0),we can just set its MV as(0,0) and skip the motion search. Previous methods use a fixed threshold to detect stationary blocks. [10] proposed an adaptive threshold, named threshold for stationary block(TSB), which makes the detection faster and robust in the sense that it can resist the influence of noise. TSB is determined as follows:

denote by MVCL the MV candidate list which contains MVs of the upper, upper-right and left MBs;

if all adjacent MBs in MVCL have MVs at (0,0),the algorithm uses the maximum of their SADs as threshold TSB;

if one of the adjacent MBs in MVCL has MV unequal to (0,0),we use the minimum of SADs of adjacent MBs as TSB.

But when comes to H.264, it uses seven types of block motion compensation, so we to shift the TSB095. The operation is as follows:

$$TSB = \begin{cases} TSB >> 1 & \text{blocktype=2,3} \\ TSB >> 2 & \text{blocktype=4} \\ TSB >> 3 & \text{blocktype=5,6} \\ TSB >> 4 & \text{blocktype=7} \end{cases} \tag{2.1}$$

2.3 Local Motion Activity[13]

Video objects in a frame often occupy a region covered by several macroblocks, so the adjacent motion vectors are highly correlated. We compute the LMA at a macroblock position based on the adjacent macroblocks' motion vectors. The definition of LMA is as follows:

$$l_i = | x_i - \bar{x} | + | y_i - \bar{y} | \tag{2.2}$$

$$L = Max\{l_i\} \tag{2.3}$$

Where L is local motion activity measurement factor, $(\bar{x}, \bar{y})$ is the average of MVs of the upper,upper-right and left MBs, and (x_i, y_i) is their MVs.

Let LMA denotes the activity of the current MB. We categorize LMA into three classes as follows:

$$LMA = \begin{cases} Low, & \text{if } L \le 1 \\ Median & \text{if } 1 < L \le 6 \\ High & \text{if } L > 6 \end{cases} \tag{2.4}$$

If the LMA is low, it means that its adjacent MBs have the similar motion activity. The current MB and its adjacent MBs may be inside the same moving object, and may have the same MV.

2.4 Threshold for Half Stop (THS)

In the searching process, when the result is good enough, we can terminate the search. In our algorithm, we use the SADs of upper, upper-right and left MBs of the current MB as a threshold for half stop. If the LMA is low or median, we use the maximum of the SADs, else we choose the mean of them.

The reason is when the LMA is low or median, it means that the current MB and adjacent MBs are high possibly belong to one object, so the minimum SADs will be same. Otherwise, the minimum SADs will be different.

2.5 Search Pattern

There are a lot of search pattern and the diamond search pattern performs quite well among them. When the real motion vector is small, SDP, LDP and CDP are suitable. SDP is used to refine a predicted MV while EDP and LDP are used for fast wide range search in diagonal direction and in horizontal and vertical direction respectively (see Fig 1). But if the real motion vector is large, these diamond search patterns do not work well. They are easy to get trapped in local minimization. Therefore, we propose the synthetic diamond search pattern(STDP). We all know that the movement in the horizontal direction is much heavier than that in the vertical direction for natural picture sequences, the optimum motion vector can be nearly accurately predicted by an unsymmetrical synthetic diamond search in the search window, see Fig 2.

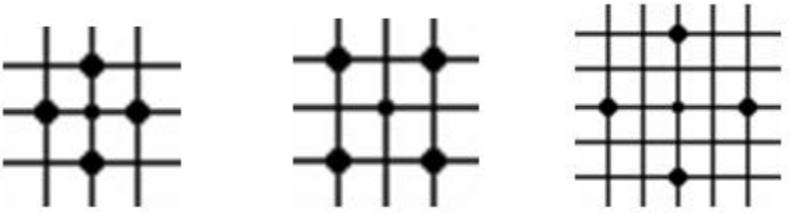


Fig. 1. SDP EDP and LDP

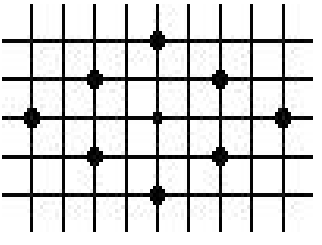


Fig. 2. STDP

2.6 Algorithm

The algorithm is described as follows:

(1). Calculate the LMA of the current MB as equations(2.2), (2.3), and get its predicted initial search point;

(2). If the LMA is low or median, the search starts from SDP, if the minimum SAD is located at the center of the diamond, the center represents the MV and the search can be terminated; else converts to EDP then to LDP. Once the minimum SAD is located at the center of the large diamond or the elastic diamond, the search goes to SDP;

(3). If the LMA is high, the search starts from STDP, if the minimum SAD is located at center, goes to SDP ; else converts to LDP.

In every step above, half stop has been used to avoid being trapped in unnecessary search.

3 Experiment Results and Comparisons

Our experiment environment is based on H.264 reference encoder of JM10, which was developed by JVT[14]. The encoder parameters are shown in table 1. All tests in the experiment are run on the Intel Pentium 4 3.6G with 512M RAM and the OS is Microsoft Windows XP. We compare our results with the FME used in JM10 and fast full search. As shown in table 2, we can see the proposed algorithm can obtain the same image quality as FME and is about 36% faster than it in various video sequences.

Table 1. Encoding Parameters

Blocktypes	QP	Num. of Reference Frame	Search Range	Encoding Pattern
7	28	5	32	IPPP

Table 2. Experiment Results

	FFS			FME			Proposed		
	PSNR (db)	BR (kbits/s)	MET (s)	PSNR (db)	BR (kbits/s)	MET (s)	PSNR (db)	BR (kbits/s)	MET (s)
news	36.77	75.35	529.790	36.72	74.97	43.619	36.72	75.30	29.543
carphone	37.29	94.00	539.710	37.28	93.65	55.772	37.27	94.37	36.594
silent	35.88	81.72	516.210	35.83	81.72	52.079	35.84	82.06	36.341
salesman	35.58	56.80	536.385	35.57	57.24	44.567	35.59	57.22	31.679
akiyo	38.19	29.36	568.642	38.21	29.50	40.104	38.19	29.46	24.803
Grandma	36.51	34.79	580.624	36.50	34.48	44.761	36.51	34.89	28.213
mobile	33.36	423.27	564.664	33.33	423.66	80.067	33.34	424.82	41.326
foreman	35.46	134.09	2463.056	35.43	133.63	300.118	35.43	135.13	188.803
stefan	35.58	1077.59	2034.869	35.56	1079.68	266.224	35.54	1088.90	181.662
tempete	34.72	1060.95	2245.258	34.70	1061.31	291.525	34.70	1062.61	175.324
sample	35.48	2077.13	2256.124	35.47	2076.66	505.208	35.48	2093.23	317.042

Table 3. The Average of Coding Efficiency

	FME	Proposed
BDPSNR(db)	-0.020	-0.017
BDBR (%)	-0.051	0.45
ME Speed Up	9.53	15.01

The average of BDPSNR, BDBR and ME speed up compared to Fast Full Search (FFS) are shown in table 3. With our algorithm, the PSNR is just 0.017db lower and bitrates is 0.45% high while the speed is 15 times up.

4 Discussion

In section II, we refer to TSB (threshold for stationary block) and THS (threshold for half stop). The effects of them were already discussed in [10]. An increase in the

upper bound of TSB and THS leads to small number of check points involved. However, the PSNR decreases while the Bit Rate increases. There is a tradeoff between video quality and encoding time. Now we give the experiment results in H.264, see table 4.

Table 4. The Effect of Upper-bounds of TSB and THS

TSB and THS's value	512			768			1024		
	PSNR	BR	MET	PSNR	BR	MET	PSNR	BR	MET
carphone	37.27	94.37	36.594	37.22	94.78	28.522	37.09	97.08	22.933
news	36.72	75.30	29.543	36.71	75.36	19.621	36.68	75.78	15.904
foreman	35.43	135.13	188.803	35.40	135.60	172.103	35.35	136.09	150.209
silent	35.84	82.06	36.341	35.84	82.44	32.395	35.83	83.19	23.569
salesman	35.59	57.22	31.679	35.59	57.29	26.155	35.57	57.30	18.447
sample	35.48	2093.23	317.042	35.47	2093.00	313.531	35.47	2098.43	285.852
stefan	35.54	1088.90	181.662	35.54	1090.59	157.572	35.53	1094.78	150.602
akiyo	38.19	29.46	24.803	38.18	29.45	15.850	38.11	30.46	12.655
grandma	36.51	34.89	28.213	36.50	34.66	23.377	36.48	34.73	16.407
tempete	34.70	1062.61	175.324	34.69	1063.40	152.746	34.66	1066.20	137.850
mobile	33.34	424.82	41.326	33.35	424.55	39.768	33.33	425.51	37.364

5 Conclusions

In video encoding software, the motion estimation is very important because it consumes most of the encoding time. So we propose a fast adaptive motion estimation algorithm. From the experiment results we know that it can save 36% ME time compared with FME while keeping the same image quality and similar coding efficiency. Our algorithm performs well both on large dynamic motion variation sequence and simple uniform motion video.

Reference

1. Koga, T., Ilinuma, K., Hirano, A., Iijima, Y., Ishiguro, T.: Motion compensated interframe coding for video conference. In: Proc. Nat. Telecommun. Conf., November 1981, New Orleans, LA, (1981)
2. Jain, J.R., Jain, A.K.: Displacement measurement and its application in interface image coding, IEEE Trans. Commun., COM-29(12) (December 1981)
3. Zhu, S., Ma, K.K.: A new diamond search algorithm for fast block-matching motion estimation. In: Proc, Int. Conf. Information, Communication and Signal Processing, September 9-12, 1997, vol. 1, pp. 292–296 (1997)
4. Tham, J.Y., Ranganath, S., Ranganath, M., Kassim, A.A.: A novel unrestricted center-biased diamond search algorithm for block motion estimation. IEEE Trans. Circuits Syst. Video Technol (August 1998)
5. Cheung, C.H., Po, L.M.: A novel cross-diamond search algorithm for fast block motion estimation. IEEE Trans. Circuits Syst. Video Technol, 12 (December 2000)

6. Zhu, C., Lin, X., Chau, L.P.: Hexagon-based search pattern for fast block motion estimation. IEEE Trans. Circuits Syst. Video Technol, 12 (May 2002)
7. Chen, Z., Xu, J.F., Zhou, P., He, Y.: Hybrid unsymmetrical-cross multi-hexagon-grid search strategy for integer pel motion estimation in H.264. In: Proceedings of PCS, San Malo (2003)
8. Kim, S.E., Han, J.K., Kim, J.G.: An efficient scheme for motion estimation using multi-reference frames in H.264/AVC. In: IEEE Trans. Multimedia, vol. 8(3), (June 2006)
9. Huang, Y.W., Hsieh, B.Y., Chien, S.Y., Ma, S.Y., Chen, L.G.: Analysis and complexity reduction of multiple reference frames motion estimation in H.264/AVC. IEEE Trans. Circuits Syst. Video Technol, 16(4) (April 2006)
10. Ahmad, I., Zheng, W., Luo, J., Liou, M.: A fast adaptive motion estimation algorithm. IEEE Trans. Circuits and systems for video tech. 16(3) (March 2006)
11. Hosur, P.I., Ma, K.K.: Motion vector field adaptive fast motion estimation. the Second Int. Conf. Inf, Commun., Signal Process., Singapore (December 1999)
12. Optimization Model Version 1.0, ISO/IEC JTC1/SC29/WG11 N3324 (March 2000)
13. Hosur, P.I.: Motion adaptive search for fast motion estimation. IEEE Trans. Consumer Electronics, 49(4) (November 2003)
14. JM10,http://iphome.hhi.de/suehring/tml/download/old_jm/jm10.zip
15. Wiegand, T.: JVT-K051, Joint Video Team (JVT) of ISO/IEC MPEG &ITU TVCEG ISO/IEC JTC1/SC29/WG11 and ITU T SG16 Q.6 , In: 12th Meeting, 17-23 July 2004: Redmond, WA, USA (2004)

Shared Latent Dynamical Model for Human Tracking from Videos

Minglei Tong and Yuncai Liu

Institute of Image Processing and Pattern Recognition
Shanghai Jiao Tong University, P.R.China, 200240
{tongminglei,whomliu}@sjtu.edu.cn

Abstract. Of different learning-based methods in human tracking, many state-of-the-art approaches have been dedicated to reduce the dimensionality of the pose state space in order to avoid complex searching in a high dimensional state space. Seldom research on human tracking refers shared latent model. In this paper, We propose a method of shared latent dynamical model (SLDM) for human tracking from monocular images. The shared latent variables can be determined easily if state vectors and observation vectors are statistically independent.With a SLDM prior over state space and observation space, our approach can be integrated into a Bayesian tracking framework of Condensation, and further a scheme of variance feedback is designed to avoid mis-tracking. Experiments using simulations and real images demonstrate this human tracking method is very efficient and promising.

Keywords: Shared Latent Dynamical Model,Human Tracking, Condensation, Variance Feedback.

1 Introduction

Tracking and reconstructing 3D articulated human motion from monocular video sequences is of interest in a broad set of applications as visual surveillance, human-computer interface, motion capture, etc. Due to the huge variability of human shapes, large number of DOF (Degree of Freedom), many local minima of solutions, occlusion in images and the ambiguous locations of joint points[3,4,7], human tracking is still a challenging research topic in human motion analysis. In general, model-based approaches for human tracking presuppose an explicitly known parametric body model [3,4,10,18,19,20,22] and then are typically formulated deterministically as a problem of nonlinear optimization or probabilistically a problem of maximum likelihood. By contrast, learning-based approaches[5,6,7,8,13,14,21,23] try to directly infer 3D pose of human from visual input or indirectly reconstruct human motions after tracking low dimensional variables learned from the data of prior motions or observations.

The approach we present in this paper is inline with the learning-based approach for human tracking. We propose a shared latent dynamical model (SLDM) derived from GPLVM [11] and GPDM [12]. GPLVM and GPDM can only learn a map from the latent space to state space. The works [25,26] present the original idea of shared latent structure, but they have not the characteristics of dynamics. Unlike the GPLVM

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 102–111, 2007.
© Springer-Verlag Berlin Heidelberg 2007

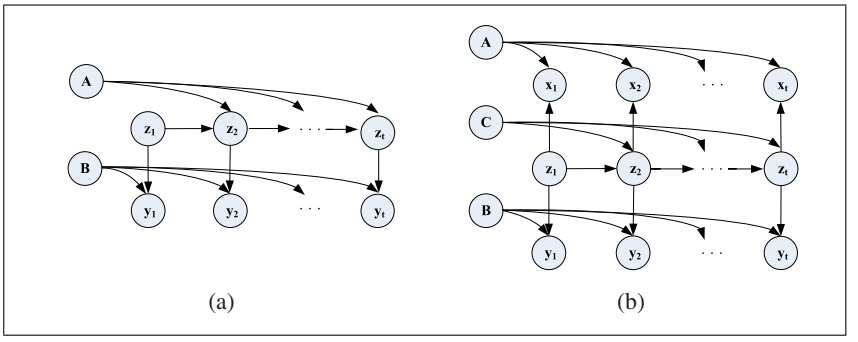


Fig. 1. The graphical model of SLDM and GPDM (a):The graphical model of GPDM; (b)The graphical model of SLDM

and GPDM, our SLDM, as shown in Fig.1, can provide a latent space simultaneously mapping to both state space and observable space with associated dynamics. The shared latent space can be determined conveniently if the state vectors and observation vectors are statistically independent. With a SLDM prior over state space and observation space, we advocate a tracking method of Condensation to estimate the latent positions and recover the human pose from reconstruction mapping. At the same time, we introduce variance feedback of mean prediction to avoid mis-tracking. Robust results are obtained when tracking human walking.

Previous Work: We focus on our survey on learning-based approaches. Agarwal and Triggs [7] learned a mapping from silhouettes to poses using relevance vector machines and tried to recover 3D poses directly from image observations. In their work, human model was synthesized by POSER with 54 degrees of freedom. A mean tracking error of only 6-7 (in degree) was obtained. However, latent space or low dimensional embedding was not applied in their work. Sminchisescu and Jepson [5] presented a learning algorithm restricting motion space to a low dimensional embedding, and provided a model of latent density with Gaussian mixture as well as recovered the pose from latent positions using a RBF mapping. Elgamma and Lee [8]learned a low dimensional embedding using Isomap or LLE from visual input, rather than from the data of pose motion, and gave a GRBF model mapping latent space to pose motion space. Rahimi et al [21] trained an embedding through a nonlinear RBF regression with an AR dynamical model to encourage smoothness in the latent space. Urtasun et al [14] advocated GPDM to learn a latent space with dynamics and the tracking problem was formulated as a MAP estimator on short pose sequences in a sliding temporal window. Tangkuampien and Suter [25] proposed a optimal pose manifold from human motion and a synthetic silhouette manifold with KPCA. They recovered human poses by mapping the silhouette manifold to pose manifold with LLE. In conclusion, there is a good deal of prior work on learning a latent space or a low dimensional embedding from

either motion data or visual input in human pose tracking. However relatively little work has been done in learning a shared latent space mapping to both of them.

Organization: This paper is organized as following: In Sec.2 , the derivation and existence term of SLDM are introduced. In Sec.3 we describes how to track human pose with the prior learned SLDM and experiments using simulated images and real image sequences are carried out.

2 Shared Latent Dynamical Model

2.1 Basic Mappings

In state space, we have continuous temporal states $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_t}]^{\mathbf{T}}$ with discrete time index $t \in \{1, 2, \ldots, N\}$and $\mathbf{x_t} \in \mathbb{R}^{\mathbf{D_1}}$. In observation space, we have $\mathbf{Y} = [\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_t}]^{\mathbf{T}}$ and $\mathbf{y_t} \in \mathbb{R}^{\mathbf{D_2}}$. In shared latent space, we have $\mathbf{Z} = [\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_t}]^{\mathbf{T}}$ and $\mathbf{z_t} \in \mathbb{R}^{\mathbf{d}}$. Consider the following three basic mappings:

$$\mathbf{x}_t = \boldsymbol{f}(\mathbf{z}_t; \mathbf{A}) + \mathbf{n}_{x,t} \tag{1}$$

$$\mathbf{y}_t = \boldsymbol{g}(\mathbf{z}_t; \mathbf{B}) + \mathbf{n}_{y,t} \tag{2}$$

$$\mathbf{z}_t = \boldsymbol{h}(\mathbf{z}_{t-1}; \mathbf{C}) + \mathbf{n}_{z,t} \tag{3}$$

where, $\mathbf{n}_{x,t}$, $\mathbf{n}_{y,t}$ and $\mathbf{n}_{z,t}$ are zero-mean, white Gaussian noise; $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{h}$ are mappings parameterized by $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively. Fig.1 depicts the graphical model of SLDM. In nonlinear case, $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{h}$ are linear combination of basis functions:

$$\boldsymbol{f}(\mathbf{z}; \mathbf{A}) = \sum_i \mathbf{a}_i \phi_i(\mathbf{z}) \tag{4}$$

$$\boldsymbol{g}(\mathbf{z}; \mathbf{B}) = \sum_j \mathbf{b}_j \psi_j(\mathbf{z}) \tag{5}$$

$$\boldsymbol{h}(\mathbf{z}; \mathbf{C}) = \sum_k \mathbf{c}_k \varphi_k(\mathbf{z}) \tag{6}$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots]$, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots]$ and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots]$ are weights ; ϕ_i, ψ_j and φ_k are basis functions.

2.2 Parameter Marginalization

$\mathbf{A}$,$\mathbf{B}$ and $\mathbf{C}$ can be marginalized out in closed form [11,12]. The weight $\mathbf{A}$,$\mathbf{B}$and$\mathbf{C}$ with an isotropic Gaussian prior on the columns is marginalized out:

$$\begin{aligned} \boldsymbol{p}(\mathbf{X}|\mathbf{Z}, \bar{\alpha}) = & \frac{|\mathbf{W}_1|^N}{\sqrt{(2\pi)^{ND_1}|\mathbf{K}_X|^{D_1}}} \\ & \cdot \exp(-\frac{1}{2} tr(\mathbf{K}_X^{-1} \mathbf{X} \mathbf{W_1}^2 \mathbf{X}^T)) \end{aligned} \tag{7}$$

$$\begin{aligned}\boldsymbol{p}(\mathbf{Y}|\mathbf{Z},\bar{\beta}) =& \frac{|\mathbf{W}_2|^N}{\sqrt{(2\pi)^{ND_2}|\mathbf{K}_Y|^{D_2}}} \\ & \cdot \exp(-\frac{1}{2}tr(\mathbf{K}_Y^{-1}\mathbf{Y}{\mathbf{W_2}}^2\mathbf{Y}^T)),\end{aligned} \tag{8}$$

$$\begin{aligned}\boldsymbol{p}(\mathbf{Z}|\bar{\gamma}) =& \boldsymbol{p}(z_1)\frac{1}{\sqrt{(2\pi)^{(N-1)d}|\mathbf{K}_Z|^d}} \\ & \cdot \exp(-\frac{1}{2}tr(\mathbf{K}_Z^{-1}\mathbf{Z}_{out}\mathbf{Z}_{out}^T)),\end{aligned} \tag{9}$$

2.3 Term of the SLDM's Existence

In this section, we derive the existence term of the SLDM with the prior density of state space and observation space.

Given prior density $p(\mathbf{X})$ and $p(\mathbf{Y})$, consider the joint probability density of latent variables and hyperparameters:

$$\begin{aligned}& p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}|\mathbf{X},\mathbf{Y}) \\ \propto & p(\mathbf{X},\mathbf{Y},\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}) \\ = & p(\mathbf{X},\mathbf{Y}|\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma})p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma})\end{aligned} \tag{10}$$

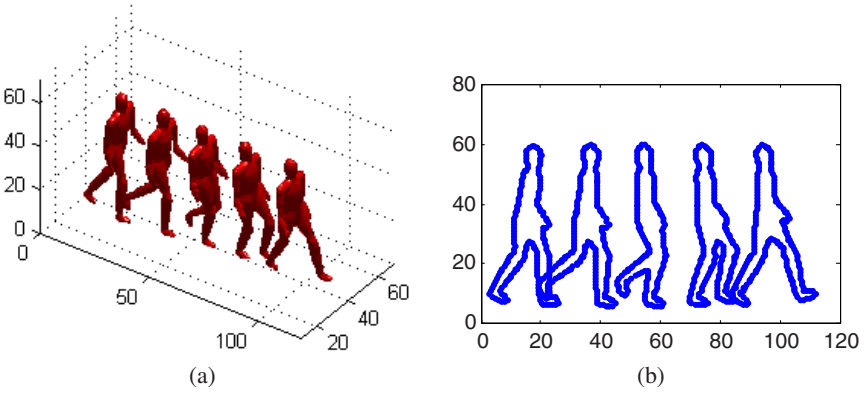
If $\mathbf{X}$ and $\mathbf{Y}$ are statistically independent,

$$\begin{aligned}& p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}|\mathbf{X},\mathbf{Y}) \\ \propto & p(\mathbf{X}|\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma})p(\mathbf{Y}|\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma})p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}) \\ = & p(\mathbf{X}|\mathbf{Z},\bar{\alpha})p(\mathbf{Y}|\mathbf{Z},\bar{\beta})p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}) \\ = & p(\mathbf{X}|\mathbf{Z},\bar{\alpha})p(\mathbf{Y}|\mathbf{Z},\bar{\beta})p(\mathbf{Z}|\bar{\gamma})p(\bar{\alpha})p(\bar{\beta})p(\bar{\gamma})\end{aligned} \tag{11}$$

Learning the SLDM from the state space and observation space entails minimizing the negative log-posterior:

$$\begin{aligned}\mathcal{L} =& -\ln p(\mathbf{Z},\bar{\alpha},\bar{\beta},\bar{\gamma}|\mathbf{X},\mathbf{Y}) \\ =& \frac{d}{2}\ln|\mathbf{K}_Z| + \frac{1}{2}tr(\mathbf{K}_Z^{-1}\mathbf{Z}_{out}\mathbf{Z}_{out}^T) \\ & - N\ln\mathbf{W_1} + \frac{D_1}{2}\ln\mathbf{K}_X + \frac{1}{2}tr(\mathbf{K}_X^{-1}\mathbf{X}{\mathbf{W_1}}^2\mathbf{X}^T) \\ & - N\ln\mathbf{W_2} + \frac{D_2}{2}\ln\mathbf{K}_Y + \frac{1}{2}tr(\mathbf{K}_Y^{-1}\mathbf{Y}{\mathbf{W_2}}^2\mathbf{Y}^T) \\ & + \sum_i \ln\alpha_i + \sum_j \ln\beta_j + \sum_k \ln\gamma_k\end{aligned} \tag{12}$$

In Equ.12, the first two terms come from the log dynamics density in Equ.9, the next three terms come from the reconstruction density in Equ.7 and the following three terms are derived from observation density in Equ.8.



(a) (b)

Fig. 2. Human model and contour (a) human body modelled by convolution surface (b) human contour modelled by convolution curve

2.4 Model Results

Image observation vector: Of many different image descriptors used for human pose estimation, we choose the shape context distribution[7] as observation vectors. Firstly, we input motion capture data into a human model of convolution surface [20], as shown in Fig.2(a), thus the human contour can be analytically computed from any viewpoint, as shown in Fig.2(b). Secondly, human contours are encoded into a 60-D shape contexts space. Finally, we reduce the distributions of all points on every silhouette to 100-D histograms by vector-quantizing the shape context space. The image observation vectors is represented as $\mathbf{y_t} \in \mathbb{R}^{\mathbf{D_2}}$ and $D_2 = 100$. **Pose state space:** The poses of our human model (Not including global motion) are represented by real vectors with discrete time t and $\mathbf{x_t} \in \mathbb{R}^{\mathbf{D_1}}$, $D_1 = 49$. The motion capture data is downloaded from `http://mocap.cs.cmu.edu` [1]. In our case, we neglect some trivial joint angles such as toes and thumbs.

SLDM initialization: In SLDM, the initial value of latent variables is set by the weighted PCA of observation vectors and state vectors. The weights can be considered as parameters to be optimized, but here we choose: $initvalue = 0.5 * PCA_{statevector} + 0.5 * PCA_{observvector}$.The hyperparameters are initially set to one and the negative log posterior $\mathcal{L}$ is minimized using Scaled Conjugate Gradient. A smoother curve of latent variables can be obtained if there are more iterations in the SCG operation.

3 Experiment in Human Tracking

With a SLDM prior over pose and observation, our tracking framework consists of three parts: dynamic equation of shared latent state, observation equation and reconstruction equation. We can infer the tracking formulation with mean prediction in [9,12]:

$$\boldsymbol{z}_t = \mu_Z(\boldsymbol{z}_{t-1}) + n_{z,t} \tag{13}$$

$$\boldsymbol{y}_t = \mu_Y(\boldsymbol{z}_t) + n_{y,t} \tag{14}$$

where Equ.13 and Equ.14 are respectively state equation and observation equation. The pose tracking problem is formulated as a low dimensional latent state tracking framework and further a reconstruction from latent space with the reconstruction equation:

$$\begin{aligned} \boldsymbol{x}_t =& \mu_X(\boldsymbol{z}_t) \\ =& \mathbf{X}^T \mathbf{K}_X^{-1} \mathbf{k}_X(\boldsymbol{z}); \end{aligned} \tag{15}$$

and $\mathbf{k}_X(\boldsymbol{z})$ is the vector with elements $k_X(\mathbf{z}, \mathbf{z}_j)$ for all other latent positions $\mathbf{z}_j$ in the model. The mean prediction $\mu_Z(\boldsymbol{z}_{t-1})$and $\mu_Y(\boldsymbol{z}_t)$ in Equ.13 and Equ.14 is very similar to the form of Equ.15, which is learned during off-line SLDM training. The SLDM approach can be integrated into a Condensation style Bayesian tracking [10,17] framework. In fact a generative model is built for observation space with the Equ.14. The likelihood at time $\boldsymbol{t}$ is:

$$\boldsymbol{p}(y_t|z_t) = \exp((y_t - \mu_Y(z_t)^T R^{-1}(y_t - \mu_Y(z_t))). \tag{16}$$

Condensation initialization: It is very important for us to get a proper initial state when condensation tracking is performed. Here we compute the Euclidean distances between the initial observation vector and the trained observation vectors. The shared latent variable with index of the least distance will be chosen as the initial state.

Resample: When the variance of mean prediction becomes larger, mis-tracking will happen. We design a simple feedback scheme for this problem. When the variance of mean prediction in state equation is larger than a threshold about 0.0013∼0.003, we resample the particles from the nearest latent state.

3.1 Tracking from Simulated Images

In this section, we will give a simulation of human tracking. First, we consider two sequences of simulated walking images modelled by convolution curves, which are generated from human motion data with 150 frames respectively in each sequence,shown

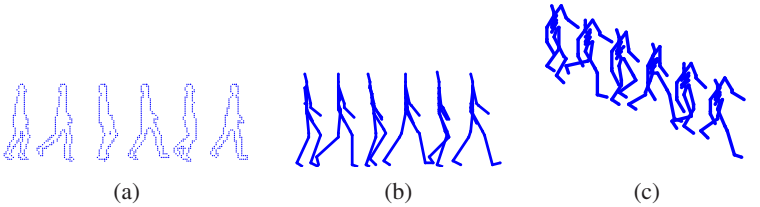


Fig. 3. Simulated experiment on a walking sequence (a) Human contour modelled by convolution curve (b)Side view of the 3D reconstruction (C) Oblique view of the 3D reconstruction

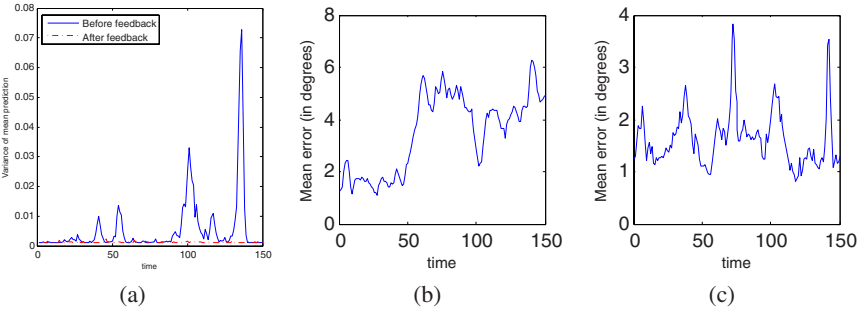


Fig. 4. The mean tracking error (a)Variance of mean prediction. (b) Mean error before the use of feedback (c)Mean error after the use of feedback

in Fig.3(a). The ground truth is provided by motion capture data downloaded from[1]. One sequence is trained with SLDM and the other is for testing. The reconstruction of 3D skeleton is shown in Fig.3(b) and Fig.3(c). Large variance of mean prediction will lead to serious mis-tracking, as shown in Fig.4(a) . The solid line in Fig.4(b) represents the variance of mean prediction before we use the feedback scheme and the dashed line represents the variance of mean prediction after the use of feedback scheme. In state equation, the one-step forecast is performed by mean prediction. If the variance of mean prediction becomes large, the tracking error will increase accordingly. The tracking

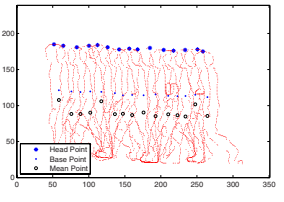


Fig. 5. Detected key points and mean point

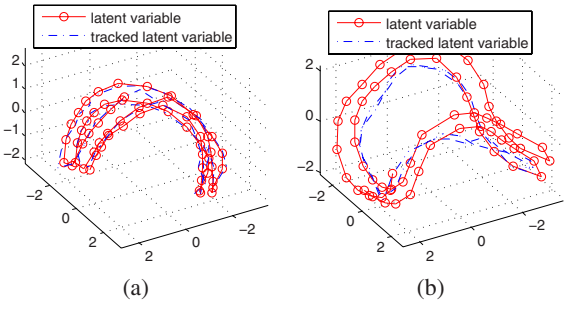


Fig. 6. Tracked latent positions (a) Tracking a man walking to the right (b)Tracking a woman walking to the left

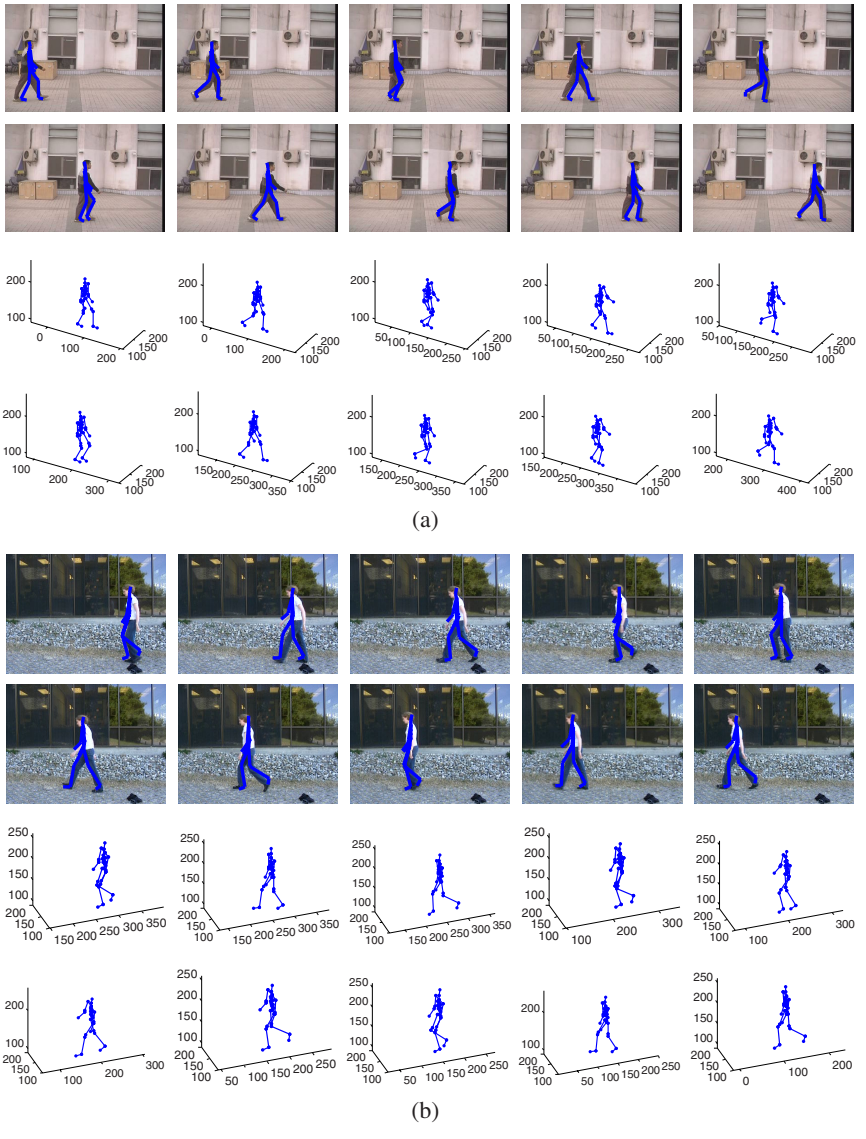


Fig. 7. Tracking two sequences of video images. The first two rows of (a) and (b) show that the skeleton of recovered model is projected into images. The last two rows of (a)and (b) are results of 3D reconstruction.

accuracy is improved when we use feedback from variance, as shown in Fig.4(c). The process noise will also greatly influence on the tracking accuracy and we can change the process noise in order to adapt to different speeds.

3.2 Tracking from Video Sequences

In this subsection, we test our method with two sequences of walking video images. Because of seriously cluttered background and badly noisy silhouette in video images, it is difficult for us to give accurate initial state automatically when tracking real images, therefore we set the initial value manually to tackle this problem. Another problem is the speed variation between the test images and the training data. For example, the test data has speed of 2 cycles per 60 frames while training data has about 2 cycles per 150 frames. We sample the training data uniformly to adapt to the speed of the test data. Experiment results show our method is very efficient.

We take the mean value of the top 10 points of human silhouette as the key point at the head. The key point on the base is estimated with the height of the human silhouette multiplied by an average coefficient calculated from the human model. In Fig.5, the circled points represent centroid of the silhouette. We can see the detected base points is more robust and stable than centroid. Fig.6 shows curves of the latent variables. The line with circled dots represents the latent variables trained by SLDM and the dashed line represents the estimated variables. The processes in Fig.6(a) and Fig.6(b) respectively correspond to human tracking in Fig.7(a) and Fig.7(b). We test our algorithm with walking sequences of 60 frames each comeing from [2] and [24] respectively, as shown in Fig.7. The experiment demonstrates that our method works robustly regardless of the cluttered background.

4 Conclusions

Conclusions. The SLDM method is introduced in this paper to map a shared latent space to both observation space and state space by learning. With a SLDM prior over pose and observation, our approach can be integrated into a Condensation style Bayesian tracking framework. Variance feedback mechanism is designed against mistracking. Experimental results show that our method is promising even working with the cluttered background.

Acknowledgments. This research is suportted by NSFC 60675017 and 973 Program 2006CB303103.

References

1. CMU Human Motion Capture DataBase. Available online at `http://mocap.cs.cmu.edu`
2. CASIA Gait Database. Available online at `http://www.sinobiometrics.com/Gait%20Database.htm`
3. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: proc. CVPR, vol. 2, pp. 126–133, Hilton Head Island (2000)
4. Sminchisescu, C., Triggs, B.: Kinematic jump processes for monocular 3d human tracking. In: proc. CVPR (June 2003)
5. Sminchisescu, C., Jepson, A.: Generative modeling for continuous non-linearly embedded visual inference. In: proc. ICML (2004)

6. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Learning Joint Top-Down and Bottom-up Processes for 3D Visual Inference. In: proc. CVPR (2006)
7. Agarwal, A., Triggs, B.: Recovering 3D Human Pose from Monocular Images. IEEE Trans. PAMI (January 2006)
8. Elgammal, A., Lee, C.: Inferring 3D Body Pose from Silhouettes using Activity Manifold Learning. In: proc. CVPR (2004)
9. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge (2003)
10. Sidenbladh, H., Black, M.J., Sigal, L.: Learning the Statistics of People in Images and Video. In: IJCV (2003)
11. Lawrence, N.D.: Gaussian process latent variable models for visulisation of high dimensional data. In: proc. NIPS 16, MIT Press, Cambridge (2004)
12. Wang, J., Fleet, D.J., Hertzmann, A.: Gaussian Process dynamical models. In: proc. NIPS 18, MIT Press, Cambridge (2005)
13. Urtasun, R., Fleet, D.J., Hertzmann, A., Fua, P.: Priors for People Tracking from Small Training Sets. In: proc. ICCV (2005)
14. Urtasun, R., Fleet, D.J., Fua, P.: 3D People Tracking with Gaussian Process Dynamical Models. In: proc. CVPR (2006)
15. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. In Science 290(5500), 2319–2323 (2000)
16. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
17. Isard, M., Blake, A.: Condensatioon-Conditional density propagation for visual tracking. In: IJCV (1998)
18. Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M.: Tracking Loose-limbed People. In: proc. CVPR (2004)
19. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. In: proc. ECCV (2002)
20. Tong, M., Liu, Y., Huang, T.S.: Recover human pose from monocular image under weak perspective projection.In: Workshop ICCV-HCI (2005)
21. Rahimi, A., Recht, B., Darrell, T.: Learing appearance manifolds from video. In: proc. CVPR (2005)
22. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: proc. CVPR (1998)
23. Li, R., Yang, M., Sclaroff, S., Tian, T.: Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers. In: proc. ECCV (2006)
24. `http://www.nada.kth.se/hedvig/data.html`
25. Tangkuampien, T., Suter, D.: Real-Time Human Pose Inference using Kernel Principal Component Pre-image Approximations. In: proc. BMVC (2006)
26. Shon, A.P., Grochow, K., Hertzmann, A., Rao, R.P.N.: Learning Shared Latent Structure for Image Synthesis and Robotic Imitation. In: NIPS (2005)
27. Ham, J., Lee, D., Saul, L.: Learning High Dimensional Correspondence from Low Dimensional Manifolds, in/it Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, ICML03

Panoramic Video Coding Using Affine Motion Compensated Prediction

Zheng Jiali, Zhang Yongdong, Shen Yanfei, and Ni Guangnan

Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
{zhengjiali, zhyd, syf}@ict.ac.cn, ngn@public.bta.net.cn

Abstract. This paper proposes an affine motion compensated prediction (AMCP) method to predict the complex changes between the successive frames in panoramic video coding. A panoramic video is an image-based rendering (IBR) technique [1] which provides users with a large field of view (e.g. 360 degree) on surrounding dynamic scenes. It includes not only the translational motions but also the non-translational motions, such as zooming and rotation etc. However, the traditional motion compensated prediction is a translational motion compensated prediction (TMCP) which cannot predict non-translational changes between panoramic images accurately. The AMCP can model the non-translational motion effects of panoramic video accurately by using six motion coefficients which are estimated by Gauss Newton iterative minimization algorithm [2]. Simulated results show that the gain of coding performance is up to about 1.3 dB when using AMCP compared with TMCP in panoramic video coding.

Keywords: Affine Motion Compensated Prediction (AMCP), Translational Motion Compensated Prediction (TMCP), panoramic video.

1 Introduction

With the increased demand for users' better experience in interactive applications such as virtual walkthrough, tele-presence and autonomous vehicles, virtual reality techniques are becoming popular. Image-based modeling and image-based rendering techniques have been received more attention as powerful alternatives to traditional geometry-based techniques for image synthesis. They rely on the characteristics of the plenoptic function [3] and use images rather than geometry as primitiveness for rendering novel views.

A panoramic video is the representation of such techniques that visualize the photorealistic and interactive environment. A panoramic mosaic is a high-resolution image obtained by projecting a series of images (registering and stitching), taken on a plane when a camera is moving along a given axis, on a cylindrical or spherical surface. [4]-[10] As the following equations [4] show, a cylindrical panorama can be obtained by mapping the world coordinates $P = (x, y, z)$ into the cylindrical screen coordinates (θ, v):

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 112–121, 2007.
© Springer-Verlag Berlin Heidelberg 2007

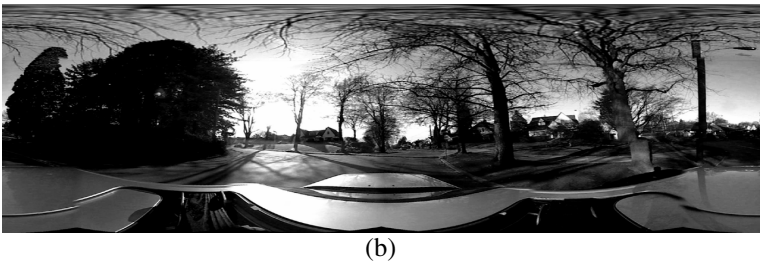
$$\begin{aligned} \theta &= \tan^{-1}(x/z) \\ v &= y/(\sqrt{x^2+z^2}) \end{aligned} . \tag{1}$$

where θ is the panning angle and v is the scan-line. Similarly, a spherical panorama can be obtained by mapping the world coordinates into the spherical screen coordinates (θ, Φ) using the equations [4] below:

$$\begin{aligned} \theta &= \tan^{-1}(x/z) \\ \Phi &= \tan^{-1}(y/\sqrt{x^2+z^2}) \end{aligned} . \tag{2}$$

Fig 1 shows the samples of the cylindrical panorama and the spherical panorama.

(a)



(b)

Fig. 1. Part of a panoramic mosaic mapped into (a) the cylindrical screen coordinates and (b) the spherical screen coordinates

Because panorama is obtained by stitching several images together, its resolution is usually very high (e.g. 1920×352). Considering such a panoramic video sequence with 25 frame/sec would occupy about 48M/s of transmission bandwidth, it is unacceptable in the interactive applications. How to compress the huge amount of panoramic video data efficiently while maintain the satisfying image quality, is a key point to improve the users' smooth interactive experience.

Moreover, a serious problem existing in panoramic video coding is that the encoder has to deal with a large amount of non-translational motions. As discussed above, to display a large field of view (e.g. 360 degree) and reduce the overlapped redundancy of the stitched images, a panoramic mosaic is mapped into a cylindrical or spherical surface. Correspondingly, a portion of horizontal straight lines in

cylindrical panorama or spherical panorama would become curved according to (1) and (2), which results in the warping changes between panoramas. Meanwhile, the panoramic images are digitally warped on-the-fly to simulate camera panning and zooming [5]. This is another important factor contributing to the increasing of the complex motions occurring in panoramic video.

To compress the huge amount of motion-video data effectively, a well known Inter-Frame predictive coding scheme named motion compensated prediction which exploiting the similarities between frames of a video signal, is widely used in existing video compression standards, such as ITU-T Recommendations H.261, H.263, H.264 and ISO MPEG-1, MPEG-2, MPEG-4. However, this motion compensated prediction scheme is based on a translational motion model which assumes the object motions as the rigid motions without any warping. That is to say, this translation motion compensated prediction (TMCP) cannot model accurately the warping motions mentioned above in panoramic video coding.

To overcome this problem, an affine motion compensated prediction (AMCP) is proposed in this paper. The AMCP is based on the principle of the affine motion vector field [11]-[17]. Here, we propose a six-parameter affine motion model to describe the motion effects of translation, zooming and rotation correspondingly in panoramic video coding. By using Gauss Newton iterative minimization algorithm to minimize the mean square prediction error, the low prediction error can be achieved and the values of motion coefficients for each block in current frame can be estimated.

The rest of this paper is organized as following: section 2 describes affine motion compensated prediction scheme in detail. The simulated results for evaluating the quality of the prediction frames and the coding performance by using proposed AMCP compared with TMCP in panoramic video coding, are presented in section 3. Section 4 gives conclusions and future works.

2 Affine Motion Compensated Prediction

The operating principle of motion compensated video coders is to minimize the prediction error $E_n(x, y)$, that is, the difference between the nth frame being coded $I_n(x, y)$, called the current frame, and a prediction frame $P_n(x, y)$:

$$E_n(x, y) = I_n(x, y) - p_n(x, y) \ . \tag{3}$$

The prediction frame $P_n(x, y)$ is built using pixel values of a reference frame denoted by $R_n(x, y)$ and the motion vectors of pixels between the current frame and the reference frame using the equation:

$$p_n(x, y) = R_n[x + \Delta x(x, y), y + \Delta y(x, y)] \ . \tag{4}$$

where the pair of values $[\Delta x(x, y), \Delta y(x, y)]$ represent the motion vector of the pixel in location (x, y) in the current frame. By combining the equation (1) and (2), the prediction error $E_n(x, y)$ can be calculated as the equation below shows:

$$E_n(x,y) = I_n(x,y) - R_n[x+\Delta x(x,y), y+\Delta y(x,y)] \ . \tag{5}$$

In the international standards, such as ISO MPEG-1, MPEG-2, MPEG-4 and ITU-T Recommendations H.26L, the used motion compensated prediction is TMCP which is based on a translational motion model. It requires only two coefficients a_0, b_0 to describe horizontal and vertical translational motion. The motion vector [$\Delta x(x, y)$, $\Delta y(x, y)$] are given by the following equations:

$$\begin{aligned} \Delta x(x, y) &= a_0 \\ \Delta y(x, y) &= b_0 \end{aligned} \ . \tag{6}$$

Since this traditional motion model only take into account the translational motion, it cannot describe accurately the complex motions in panoramic videos which contain translational motions and non-translational motions. To deal with this problem, this paper proposes an AMCP based on an affine motion model which is given by the following equations:

$$\begin{aligned} \Delta x(x, y) &= a_0 + a_1 x + a_2 y \\ \Delta y(x, y) &= b_0 + b_1 x + b_2 y \end{aligned} \ . \tag{7}$$

where the function of a_0, b_0 model the translational effect, a_1x, b_2y model the zooming effect and a_2y, b_1x model the rotation effect. As discussed above, these are the main motion effects existing in the panoramic video.

As (7) shows, to obtain the values of motion vector [$\Delta x(x, y)$, $\Delta y(x, y)$] in affine motion model, the key point is to estimate the values of motion coefficients (a_0,a_1,a_2,b_0,b_1,b_2). The rest of this section gives the calculation of motion coefficients in detail.

Let C = (c_0,c_1,c_2,c_3,c_4,c_5) = (a_0,a,a_2,b_0,b_1,b_2), then C can be calculated by minimizing the mean square prediction error which is given below:

$$\begin{aligned} &\sum_{x\in m, y\in m} DPD^2(x,y,c) = \sum_{x\in m, y\in m} E_n^{\,2}(x,y,c) = \\ &\sum_{x\in m, y\in m} (I_n(x,y) - R_n[x+\Delta x(x,y,c), y+\Delta y(x,y,c)])^2 \end{aligned} \ . \tag{8}$$

where m denotes the coded block in current frame.

For performing the full search of all possible coefficients combinations is very computational, an alternative method which uses Gauss Newton iterative minimization algorithm to obtain a set of motion coefficients for the coded block, is summarized in the following:

$$G\delta = g \ . \tag{9}$$

where $\delta = C^{new} - C^{old}$, C^{old} denote the motion coefficients input of Gauss Newton iteration and C^{new} denote the motion coefficients output of Gauss Newton iteration. G is a 6×6 positively semi-definite symmetric matrix which estimates second derivative

of the $DPD^2(x, y, c)$ with respect to the coefficients C. G has elements as following equations show:

$$G(c)=\frac{DPD^2(x,y,c)}{\partial c_i \partial c_j}=\frac{DPD(x,y,c)}{\partial c_i}\cdot\frac{DPD(x,y,c)}{\partial c_j} \quad . \tag{10}$$

I= 0~2 and j = 0~2
i.e.

$$G(c)=\begin{bmatrix} \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_0 \partial c_5} \\ \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_1 \partial c_5} \\ \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_2 \partial c_5} \\ \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_3 \partial c_5} \\ \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_4 \partial c_5} \\ \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_0} & \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_1} & \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_2} & \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_3} & \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_4} & \frac{\partial^2 DPD(x,y,c)}{\partial c_5 \partial c_5} \end{bmatrix} . \tag{11}$$

Corresponding the elements $\Delta x, \Delta y$ in the prediction error function $DPD(x, y, c)$, there are two equations to calculate $\frac{\partial DPD(x, y, c)}{\partial c_i}$:

$$\frac{\partial DPD(x, y, c)}{\partial c_i}=\frac{\partial DPD(x, y, c)}{\partial \Delta x}\cdot\frac{\partial \Delta x}{\partial c_i} \quad . \tag{12}$$

$$\frac{\partial DPD(x, y, c)}{\partial c_i}=\frac{\partial DPD(x, y, c)}{\partial \Delta y}\cdot\frac{\partial \Delta y}{\partial c_i} \quad . \tag{13}$$

As defined above, $C = (c_0,c_1,c_2,c_3,c_4,c_5) = (a_0,a_1,a_2,b_0,b_1,b_2)$, combined with (7), the values of $\frac{\partial \Delta x}{\partial c_i}, \frac{\partial \Delta y}{\partial c_i}$ can be calculated as shown in the following equations:

$$\frac{\partial \Delta x}{\partial c_0}=\frac{\partial \Delta y}{\partial c_3}=1, \frac{\partial \Delta x}{\partial c_1}=\frac{\partial \Delta y}{\partial c_4}=x, \frac{\partial \Delta x}{\partial c_2}=\frac{\partial \Delta y}{\partial c_5}=y \quad . \tag{14}$$

Let *dx* denote the $\frac{\partial DPD(x,y,c)}{\partial \Delta x}$ and *dy* denote the $\frac{\partial DPD(x,y,c)}{\partial \Delta y}$, following 6×6 positively semi-definite symmetric matrix can be obtained by substituting (12), (13) and (14) into (11):

$$G(c)=\begin{bmatrix} dx^2 & dx^2\cdot x & dx^2\cdot y & dx\cdot dy & dx\cdot dy\cdot x & dx\cdot dy\cdot y \\ dx^2\cdot x & dx^2\cdot x^2 & dx^2\cdot x\cdot y & dx\cdot dy\cdot x & dx\cdot dy\cdot x^2 & dx\cdot dy\cdot x\cdot y \\ dx^2\cdot y & dx^2\cdot x\cdot y & dx^2\cdot y^2 & dx\cdot dy\cdot y & dx\cdot dy\cdot x\cdot y & dx\cdot dy\cdot y^2 \\ dy\cdot dx & dy\cdot dx\cdot x & dy\cdot dx\cdot y & dy^2 & dy^2\cdot x & dy^2\cdot y \\ dy\cdot dx\cdot x & dy\cdot dx\cdot x^2 & dy\cdot dx\cdot x\cdot y & dy^2\cdot x & dy^2\cdot x^2 & dy^2\cdot x\cdot y \\ dy\cdot dx\cdot y & dy\cdot dx\cdot x\cdot y & dy\cdot dx\cdot y^2 & dy^2\cdot y & dy^2\cdot x\cdot y & dy^2y^2 \end{bmatrix}. \quad (15)$$

In (9), g is a vector of size 6 with elements as the following equation show:

$$g(c)=DPD(x,y,c)\cdot\frac{\partial DPD(x,y,c)}{\partial c_i}, i=0\sim 5\ . \quad (16)$$

Similarly, Let *dx* denote $\frac{\partial DPD(x,y,c)}{\partial \Delta x}$, *dy* denote $\frac{\partial DPD(x,y,c)}{\partial \Delta y}$ and E_n denote *DPD(x, y, c)*, (16) can be transformed into the 6×1 matrix below by combining (12), (13) and (14):

$$g(c)=\begin{bmatrix} dx\cdot E_n \\ dx\cdot x\cdot E_n \\ dx\cdot y\cdot E_n \\ dy\cdot E_n \\ dy\cdot x\cdot E_n \\ dy\cdot y\cdot E_n \end{bmatrix}. \quad (17)$$

For δ in (9), it is a 6×1 matrix which can be represented by the following matrix:

$$\delta=\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}. \quad (18)$$

Finally, (9) can be solved by substitute (15) (17) and (18) into (9) and δ is outputted as a result. For $\delta = C^{new}-C^{old}$, a new set of motion coefficients C^{new} are given by the following equation:

$$C^{new} = \delta + C^{old} . \tag{19}$$

When a Gauss-Newton iteration step has been done, the method deciding whether the outputted motion coefficients are the final results, is to evaluate the prediction error. If the difference between mean square prediction error after and before the step is smaller than a threshold, stop the Gauss-Newton iteration and output C^{new} as the final result. Where, the threshold is set as 0.1. If not, look C^{new} as C^{old}, input it to the next Gauss-Newton iteration step to obtain new set of motion coefficients. When final $C^{new} = (c_0,c_1,c_2,c_3,c_4,c_5) = (a_0,a_1,a_2,b_0,b_1,b_2)$ have been obtained, substitute it to (7) to calculate the values of motion vector $[\Delta x(x, y), \Delta y(x, y)]$ for each pixel in the block m.

3 Simulated Results

In the simulated tests, we integrate the proposed technique into the H.264/AVC reference software JM90 [18]. The test sequences include panoramic video sequences [19] Village and Hall. Both sequences are 1920×352 format with 25 fps and coded with GOP structure "IPPPPP…". The Peak Signal to Noise Ratios (PSNR) presented in the results is calculated by the following equation:

$$SNR = 10\log_{10}\frac{M \cdot N \cdot 255^2}{\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}(I_n(x, y)-\tilde{I}_n(x, y))^2} . \tag{20}$$

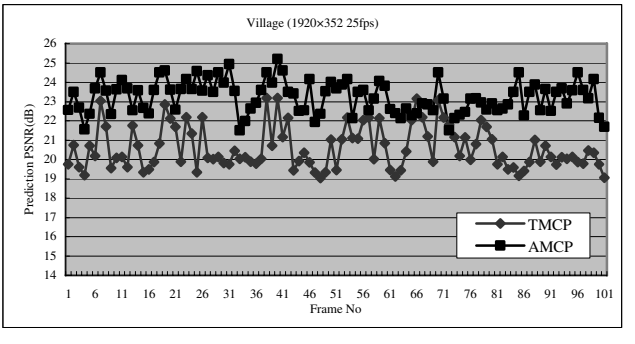
where, $M \times N$ is the image size, $I_n(x, y)$ is the original frame luminance function and $\tilde{I}_n(x, y)$ is the prediction frame luminance function when calculates the prediction PSNR which is presented in Fig 2. The prediction PSNR gives an indication of how similar the prediction is to the original frame. The higher the prediction PSNR is, the better is the prediction. In Fig 3, the reconstruction PSNR presented in rate-distortion curves, also is calculated by using (20), where $\tilde{I}_n(x, y)$ is the reconstruction frame luminance function.

It is clearly observed from Fig 2 that the prediction images PSNR are improved highly when panoramic video coding using AMCP compared with TMCP. Table 1 shows the allocation of bits between coding motion vectors and coding prediction error, as well as average prediction PSNR. Although the affine motion model using more motion coefficients than the translational motion model, which results in a slight increase of coding bits for motion vectors, the motion compensated prediction is more accurate and the prediction PSNR is better. To evaluate the coding performance of proposed technique compared with traditional technique, the rate-distortion curves are

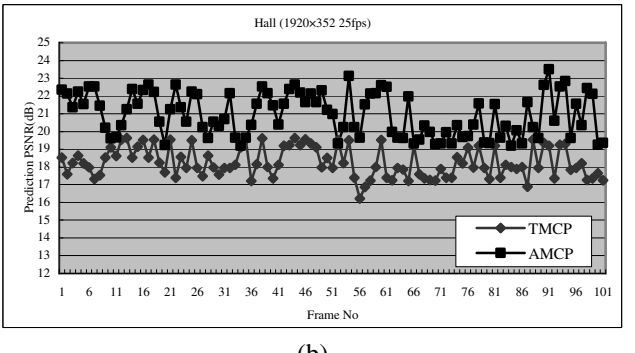
presented in Fig 3. The average gain with the proposed AMCP is up to 1.36 dB in Village sequence and 1.31 dB in Hall sequence.

4 Conclusion

The proposed affine motion model is a six-parameter motion model which more accurately approximates the changes between panoramic frame by predicting not only translational motions, but also non-translational motions, such as zooming and rotation. When integrated with AMCP, the better quality of prediction frame can be obtained and the panoramic videos are compressed more effectively than using traditional TMCP. Considering the affine motion model achieves more accurate motion compensated prediction at the cost of more motion coefficients needed to transmit to decoder, how to decrease the number of these motion coefficients and quantize them effectively according to the motion characteristics of the panoramic video, while remain the coding performance without degenerating highly, will be the issue being worth studying in future work.



(a)

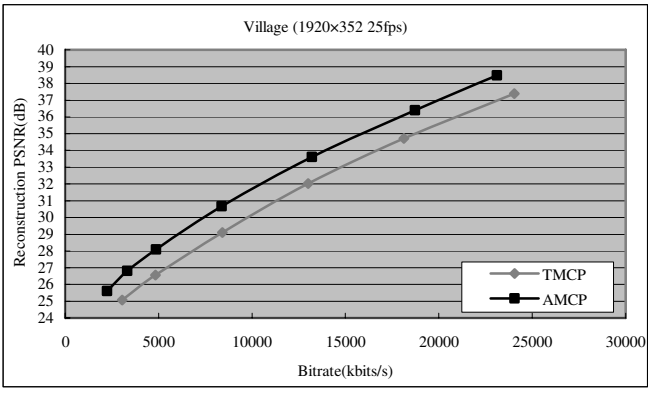


(b)

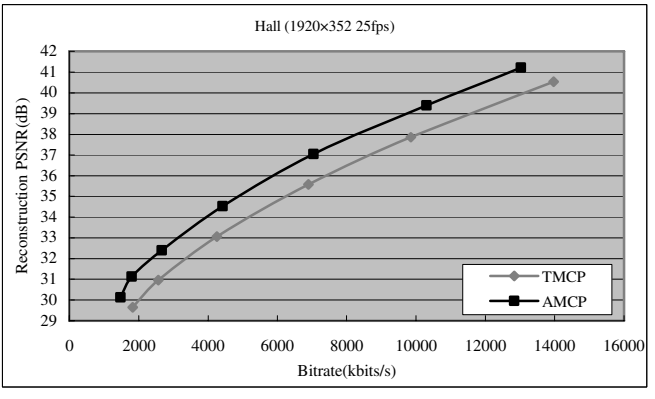
Fig. 2. Prediction PSNR for each P frame in coding panoramic video sequences (a) "Village" and (b) "Hall" by using AMCP compared with TMCP

Table 1. Average bits spent on coding motion vectors and prediction error and average prediction PSNR of each P frame by using AMCP compared with TMCP in panoramic video coding

Sequence	Motion compensated prediction method	Motion Vector (kbits)	Prediction error (kbits)	Prediction PSNR (dB)
Village	TMCP	25.62	611.25	21.5
	AMCP	58.09	582.36	23.86
Hall	TMCP	14.6	183.38	19.86
	AMCP	31.3	171.6	21.83



(a)



(b)

Fig. 3. Rate-distortion curve of coding panoramic video (a) "Village" and (b) "Hall" by using AMCP compared with TMCP

Acknowledgments

This work is partially supported by the National Nature Science Foundation of China under the contract No.60302028, the Beijing Science and Technology Planning Program of China under the contract No.D0106008040291.

References

1. Shum, H.Y., Kang, S.B., Chan, S.C.: Survey of image-based representations and compression techniques. IEEE Transaction on Circuit and Systems for Video Technology 13(11), 1020–1037 (2003)
2. Fletcher, R.: Practical Methods for Optimization, 2nd edn. chapter 3 and chapter 6, John Wiley & Sons, Chichester, (1987)
3. Wong, T.T., Fu, C.W., Heng, P.A., Leung, C.-S.: The plenoptic illumination function. IEEE Transaction on Multimedia 4(3), 361–371 (2002)
4. Szeliski, R., Shum, H.Y.: Creating full view panoramic image mosaics and texture-mapped models. Computer Graphics (SIGGRAPH'97), pp. 251–258 (August 1997)
5. Chen, S.E.: Quick Time VR – an image-based approach to virtual environment navigation. Computer Graphics (SIG-GRAPH'95) , pp. 29–38 (August 1995)
6. Ng, K.T., Chan, S.C., Shum, H.Y.: On the data compression and transmission aspects of panoramic video. IEEE Transaction on Circuit and Systems for Video Technology 15(1), 82–95 (2005)
7. Li, Y., Shum, H.Y., Tang, C.K., Szeliski, R.: Stereo reconstruction from multiperspective panorama. IEEE Transaction on Pattern Analysis and Machine Intelligence 26, 45–62 (2004)
8. Yamada, K., Ichikawa, T., Naemura, T., Aizawa, K., Saito, T.: Generation of a disparity panorama using a 3-camera capturing system. IEEE Transaction on Image Process 2, 772–775 (2000)
9. Tang, W.K., Wong, T.T., Heng, P.A.: A system for real-time panorama generation and display in tele-immersive applications. IEEE Transaction on Multimedia 7(2), 280–292 (2005)
10. Schechner, Y.Y., Nayar, S.K.: Generalized mosaicing: polarization panorama. IEEE Transaction on Pattern Analysis and Machine Intelligence 27(4), 631–636 (2005)
11. Wiegand, T., Steinbach, E., Girod, B.: Affine multipicture motion-compensated prediction. IEEE Transaction on Circuit and Systems for Video Technology 15(2), 197–209 (2005)
12. Karczewicz, M., Nieweglowski, J., Haavisto, P.: Video coding using motion compensation with polynomial motion vector fields. Signal Process: Image Commun 10, 63–91 (1997)
13. Utgikar, A., Badawy, W., Seetharaman, G., Bayoumi, M.: Affine scheme in mesh-based video motion compensation.In: Proc. IEEE Workshop SIPS 2003, pp. 159–164 (August 2003)
14. Huang, J.C., Hsieh, W.S.: Automatic feature-based global motion estimation in video sequences. IEEE Transaction on Consum Electron 50(3), 911–915 (2004)
15. Servais, M., Vlachos, T., Davies, T.: Affine motion compensation using a content-based mesh. In: Proc. IEEE Vis. Image Signal Process, vol. 152(4), pp. 415–423 (August 2005)
16. Keller, Y., Averbuch, A.: Fast motion estimation using bidirectional gradient methods. IEEE Transaction on Image Process, 13(8), 1042–1054 (2004)
17. Gahlots, A., Arya, S., Ghosh, D.: Object-based affine motion estimation. In: Proc. Conf. TENCON 2003, vol. 13, pp. 1343–1347 (October 2003)
18. http://iphome.hhi.de/suehring/tml/download
19. ftp://ftp.tnt.uni-hannover.de

Robust Bootstrapping of Speaker Models for Unsupervised Speaker Indexing

Fu ZhongHua

School of Computer Science, Northwestern Polytechnical University,
Xi'an 710072, P.R. China
mailfzh@nwpu.edu.cn

Abstract. The conventional bootstrapping approaches of speaker models in unsupervised speaker indexing tasks are very sensitive to the bootstrapping segment duration. If the duration is insufficient to build speaker model, such as in telephone conversations and meetings scenario, serious problems will arise. We therefore propose a robust bootstrapping framework, which employs Multi-EigenSpace modeling technique based on Regression Class (RC-MES) to build speaker models with sparse data, and a short-segment clustering to prevent the too short segments from influencing bootstrapping. For a real discussion archive with a total duration of 8 hours, we demonstrate the significant robustness of the proposed method, which not only improves the speaker change detection performance but also outperforms the conventional bootstrapping methods, even if the average bootstrapping segment duration is less than 5 seconds.

Keywords: Unsupervised speaker indexing, speaker model, eigenvoices.

1 Introduction

Unsupervised speaker indexing sequentially detects points where a speaker identity changes in a multi-speaker audio stream, and categorizes each speaker segments, without any prior knowledge. This paper considers the on-line unsupervised speaker-indexing problem.

Mainly the approaches to unsupervised speaker indexing are classified into metric-based and model-based methods. The metric-based methods such as the Bayesian information criterion (BIC) [1] and the generalized likelihood ratio (GLR) [2] are usually implemented for speaker change detection tasks. They use hypothesis test based on some distance measure between two adjacent speech segments. The model-based approaches, such as Gaussian mixture models (GMM) adapted from universal background models (UBM) [3] or sample speaker models (SSM) [4], are based on speaker identification techniques to cluster every segment according to the log likelihood score. Generally, to enable on-line indexing, metric-based approaches and model-based approaches are combined together, where the metric ones are always used to separate the utterances of specific speaker from the audio stream, and the model ones then use these utterances to bootstrapping his model and to index them [4].

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 122–129, 2007.
© Springer-Verlag Berlin Heidelberg 2007

To correctly index the utterances, the parameters of each speaker model need to be estimated reliably with sufficient data. However, it is almost impossible in unsupervised real-world on-line processing scenario, because only the data seen thus far can be used for modeling. In telephone conversations and meetings, for example, there are many short utterances with large duration variations. In the discussion data here, the duration can be 1 seconds or 2 minutes. The ratio of the utterances that are less than 10 seconds is about 75%. In these circumstances, data are usually not sufficient to bootstrapping speaker-dependent models. Although a model still can be roughly built, it is apt to cause decision errors due to potential uncertainty in the unsupervised learning. We therefore focus on the more effective model bootstrapping method.

Speaker model training with limited data is one of the most challenging issues in speaker recognition. There are some possible ways to solve this problem by using eigenvoices [5] or maximum likelihood linear regression (MLLR) [6] techniques. In our previous work, we have proposed a better approach named multi-eigenspace modeling based on regression class (RC-MES) [7], which separates the speaker differences and phoneme differences involved in speech data. Based on this approach, a robust framework for speaker model bootstrapping in on-line unsupervised speaker indexing is proposed. To prevent the too short segment from influencing the bootstrapping, a short-segment clustering based on GLR criterion is employed.

The rest of this paper is organized as follows: Section 2 briefly introduces the RC-MES approach. Section 3 presents the proposed framework. The description of the experiments and their results are given in Section 4. Section 5 concludes this paper.

2 Speaker Model Adaptation Using RC-MES

To train speaker dependent model with sparse data is one of the most challenge issues in speaker recognition. The common eigenvoices approach [5] aims at reducing the number of parameters to be estimated during model training. It considers that any speaker specific model is a projection inside an eigenspace, i.e. the speaker model is constrained as a linear combination of a small number of basis vectors, which are obtained offline from a set of reference speakers. Thus the model training becomes the estimation of weights of each basis vectors. However, the acoustic differences comprised in utterances from different speakers are not only caused by speaker's voice identity, but also the context differences, i.e. if different speakers utter same words, then the only reason causes the acoustic differences is their voiceprint.

The RC-MES approach adopts the idea of MLLR and implements the eigenvoices inside each regression class (RC). If adaptation data is sufficient, each RC corresponds to a phoneme. Otherwise, according to the data available, the phoneme set is divided into different number of RCs based on acoustic similarity. All the possible divisions constitute a regression class tree.

In off-line procedure, every node in the regression class tree is trained adequately to build a speaker independent (SI) model and a group of eigenvoices. In on-line adaptation procedure, according to the data available, the correct division of RCs is chosen to ensure each class has sufficient data to estimate its weights of eigenvoices.

Then the training data is separated into those regression classes based on their SI models. The maximum likelihood eigen-decomposition (MLED) is used to estimate the weights inside each RC.

The eigenvoices of RC s can be written as

$$\mathbf{e}^s(j) = \left[e_0^s(j)^T, e_1^s(j)^T, \cdots, e_m^s(j)^T, \cdots, e_M^s(j)^T\right], j = 0,1,\cdots,K \tag{1}$$

where $e_m^s(j)$ is the means vector of m-th Gaussian component in RC s, K is the feature dimension. The weights of eigenvoices are iterative re-estimated by using

$$\sum_{t=1}^{T}\sum_{m=1}^{M} r^{(m)}(t)\cdot\left[e_m^s(i)\right]'\cdot C_m^{-1}\cdot o_t^{(s)} = \sum_{t=1}^{T}\sum_{m=1}^{M} r^{(m)}(t)\cdot\left\{\left[\sum_{k=0}^{K} w_s(k)\cdot e_m^s(k)\right]'\cdot C_m^{-1}\cdot e_m^s(i)\right\} \tag{2}$$

where $o_t^{(s)}$ is observation vector at time t, which belongs to RC s, C_m^{-1} is the inverse of covariance matrix of m-th component in RC s, $r^{(m)}(t)$ is the occupation probability

$$r^{(m)}(t) = P\left(i_t = m \middle| o_t^{(s)}, \lambda_s\right) = p_m^s \cdot b_m^s\left(o_t^s\right) \Big/ \sum_{k=1}^{M} p_k^s \cdot b_k^s\left(o_t^s\right) \tag{3}$$

where p_m^s and $b_m^s()$ are the weight and pdf of m-th Gaussian component in RC s.

In eq. (2) there are k+1 equations to solve for the k+1 unknown weights ($w_s(i)$ values) of RC s. After the estimation of weights, the supervector of the new speaker is obtained, so does the means of each Gaussian component in each RC. Then the components in each RC are bound together to build the GMM for the new speaker, the corresponding covariance matrixes come directly from SI of each RC. The component weights are normalized and then re-estimated once using expectation-maximization (EM) algorithm.

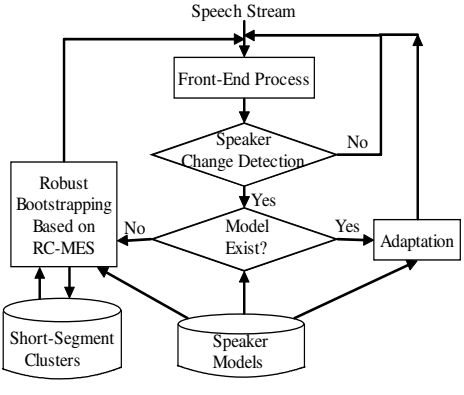


Fig. 1. Flow diagram of the proposed unsupervised speaker indexing

3 Speaker Change Detection and Model Bootstrapping

The flow diagram of the proposed on-line unsupervised speaker-indexing framework is illustrated in Fig.1. The crucial parts of the framework include: robust speaker change detection, bootstrapping of each initial speaker model, adaptation of speaker model with the new onset segmentation, and speech segmentation clustering. Moreover, in real system design, the front-end processing modules such as feature extraction and noise cancellation are also important for better performance, but they are not concerned in this paper.

3.1 Speaker Change Detection

Robust speaker change detection is the most critical prerequisite for speaker modeling and speech clustering. Since no prior knowledge is known beforehand, the change point detection between utterances of different speakers can be coarsely implemented merely based on some metric-based criterion, such as BIC or GLR. However, the parameters used for evaluation and thresholds have to be tuned empirically according to different acoustic and environmental conditions. Besides, in order to find finer resolution of change points, a small window shift is needed so that the computational complexity will severely increase. Therefore, we choose a robust GLR-based change point detection approach [8] with localized search algorithm [4].

Suppose there are two feature vector sets X and Y, which are respectively extracted from two adjoining segments comprised in the current analysis window. The feature vector sets are denoted as $X=\{x_1, x_2, \cdots, x_N\}$ and $Y=\{y_1, y_2, \cdots, y_N\}$, where N is the number of frames of each segment. Let Z be the union of X and Y. The data X and Y are individually modeled with a Gaussian model (denoted as λ_X and λ_Y) using the maximum likelihood (ML) estimation, and the data Z is modeled with a two-component Gaussian mixture model (denoted as λ_Z) estimated by using EM algorithm. Then the resulting GLR dissimilarity d_{GLR} between the two segments is calculated as

$$d_{GLR} = L(Z|\lambda_Z) - \left(L(X|\lambda_X) + L(Y|\lambda_Y)\right) \tag{4}$$

where

$$L(Z|\lambda_Z) = \sum_{i=1}^{N} \log p(x_i|\lambda_Z) + \sum_{i=1}^{N} \log p(y_i|\lambda_Z) \tag{5}$$

$$L(X|\lambda_X) = \sum_{i=1}^{N} \log p(x_i|\lambda_X),\ L(Y|\lambda_Y) = \sum_{i=1}^{N} \log p(y_i|\lambda_Y) \tag{6}$$

All local maximum of d_{GLR} greater than 0 are considered as speaker change point. Since the metric-based approaches cannot avoid the false detection points and the skip

of real change points. A lower threshold, denoted as θ_{GLR} is employed in our experiments to further improve the performance.

To simplify the computational complexity while keeping finer resolution, LSA, a fixed analysis segment and a variable window shift are used. The analysis window firstly shifts with a large step until the d_{GLR} between the two adjoining segments inside the window falls below the threshold, where speaker may have changed in that window. Then a small step is used for finer localization of the real change point. In the experiments following, the length of segment is 2s and the steps are 1s and 0.2s.

3.2 Bootstrapping and Clustering

The metric-based speaker change detection will offer speech segments, which are supposed, from different speakers, and then the corresponding speaker models are initialized or adapted with these new data. In the unsupervised scenario, the first initial model of each speaker have to be trained or adapted from generic models, such as UBM and SSM, with only the data seen thus far. The common procedure is like this: The first speaker model is always created from UBM (or SSM pool) using the first speaker segment. With the second segment, the likelihood score is calculated using the model just constructed. If the score falls blow the threshold, then a new speaker model is created also from UBM (or SSM pool) in the same way. Each time the new encountering segment is evaluated against all available speaker models to judge whether a new speaker model is needed. This procedure implies that all information used to create a new speaker model is coming from his first segment, which in most realistic situation is far beyond enough. For example, in our experimental data, the utterance duration differs from 1s to 120s, and most short utterances are concentrated at the beginning and ending of the conversation. If the model is still roughly created, then the following discrimination of each segment will be confused.

To recover the shortage of normal bootstrapping strategy, we adopted RC-MES approach, which can created speaker model with sparse data. And to keep high confidence of the model training, the shorter segments are separated and united by using metric-based approach until the length of union segments is enough for RC-MES training, i.e. the speaker change detection procedure will save all short segments and cluster them correspondingly until the data of one cluster is enough to using RC-MES. Let $S = \{s_1, s_2, \cdots, s_K\}$ be the set of speaker specific clusters, where K is the possible number of speaker. The GLR between cluster i and cluster j is denoted as $d_{GLR}(s_i, s_j)$, where $d_{GLR}(s_i, s_j) > \theta_{GLR}$. The data length of each cluster is shorter than the bootstrapping segment duration, denoted as θ_{RC-MES}. The new detected segment s_{new} is then compared with each cluster inside S. If $d_{GLR}(s_{new}, s_i) < \theta_{GLR}$, then s_{new} is combined with cluster i. If no such s_i is found, then a new cluster comprising s_{new} is created. After this clustering procedure, the length of each cluster is checked. The cluster, which length is longer than θ_{RC-MES}, is selected to train the corresponding speaker model using RC-MES.

The bootstrapping and clustering algorithm runs as follows:

(1) Find a new segment s_{new} using GLR
(2) Evaluated the likelihood score against every speaker model
(3) If the biggest score, corresponding to speaker model λ_i, is bigger than the identification threshold θ_{id}

Adapt λ_i with s_{new} by using maximum a posteriori (MAP) and then go to (1)
Else cluster s_{new} to the short segment set S

(4) Check the length of each cluster in S
(5) If the longest length corresponding to cluster s_j is longer than θ_{RC-MES}

Create a new speaker model using RC-MES with data in s_j, and delete s_j from S

(6) Go to (1)

Note that when there is no speaker model at the beginning, the likelihood score is set to 0 and the original S is empty.

4 Experiments and Results

The materials for the following speaker indexing experiments are collected from several meetings, symposiums, conversations, etc of our laboratory, which consist of about 8-hours usable speech data. The numbers of speakers inside these scenarios are from 2 to 5. The durations of utterances are from 1s to 120s and about 75% utterances are less than 10 seconds. All audio data are sampled at 16 kHz (16bit and mono) and pre-emphasized. Energy and zero-crossing parameters are used for speech end-point detection. The acoustic features consist of 26 components of 12 MFCCs, energy and their deltas. 30-ms Hamming window shifted by 10 ms is used for short time analysis.

The off-line of RC-MES includes training each phoneme with 6 components GMM, building the regression class tree by using divergence measure, random selecting 100 speakers from TIMIT database, and training eigenvoices in each regression class with corresponding data from the selected speakers. In on-line step, because the bootstrapping data duration of RC-MES, θ_{RC-MES}, is predefined, the actual division number of regression class is also known beforehand.

The first experiments are designed for appropriate choice of θ_{GLR}, which influences the recall (RCL) and the precision (PRC) well. The theoretical θ_{GLR} is 0, but to avoid skipping the real change point, we decrease it slightly so that the following indexing procedures can refine the detection results. The RCL and PRC are respectively defined as

$$PRC = \frac{number\ of\ correctly\ found\ \ changes}{total\ \ number\ of\ \ changes\ found} \tag{7}$$

$$RCL = \frac{number\ of\ correctly\ found\ \ changes}{total\ number\ of\ \ correct\ changes} \tag{8}$$

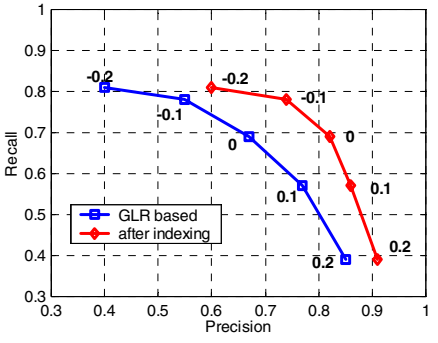


Fig. 2. Recall-Precision tradeoff of Speaker change detection with different θ_{GLR}

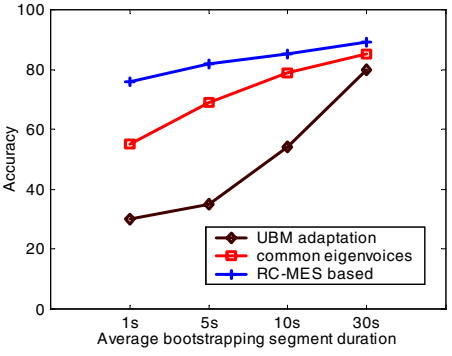


Fig. 3. Speaker indexing accuracy for various bootstrapping approaches with different bootstrapping segment duration

The detection results are show in Fig. 3. It is clear that the indexing procedure can refine the detection performance by improving the PRC while keeping the RCL. When the θ_{GLR} slightly decreased, the RCL becomes higher and the loss of PRC in GLR-based approach can be further compensated after speaker indexing procedure, for example, with GLR-based approach, when the θ_{GLR} drops from 0 to –0.1, the RCL raises from 0.69 to 0.78 while the PRC drops from 0.67 to 0.55. Nevertheless, after the clustering effects of indexing procedure, the PRC is recovered to 0.74. We choose θ_{GLR} as –0.1 in the following experiments, where the RCL is 0.78 and the PRC is 0.74.

The second experiments compared the conventional bootstrapping methods such as UBM adaptation and common eigenspace approach with the proposed RC-MES based approach. In the UBM case, the UBM consists of fully trained 512 Gaussian mixtures. Each speaker model consists of 16 mixtures and is adapted from UBM by MAP. In common eigenspace case, 40 global-eigenvoices are used. In our proposed approach, those eigenvoices are trained inside each regression class and θ_{RC-MES} is set to 10s. We compared the speaker indexing accuracy with different average durations of bootstrapping segments corresponding to every speaker, which influence directly speaker models adaptation in conventional bootstrapping strategies. The experimental results are show if Fig. 4. It is clear that the proposed RC-MES approach shows the most robust indexing accuracy than the others. The short-segment clustering maintains the bootstrapping segment length for model training. In contrast, the UBM adaptation and common eigenspace are very sensitive to the bootstrapping segment durations. They all need more bootstrapping data to gain enough speaker dependent information. On the other hand, when the duration extends to 30s, the proposed approach also shows better performance than the others, which implies that the RC-MES based speaker model adaptation method can refine the speaker identification ability of the system.

5 Conclusions

A robust bootstrapping framework based on RC-MES for unsupervised speaker indexing is proposed. The conventional bootstrapping strategies of speaker models are very sensitive to insufficient data in real-world applications. The new framework chooses the RC-MES based speaker model adaptation technique to create new speaker model with sparse bootstrapping data. To prevent the too short segments from influencing the bootstrapping process, a short-segment clustering approach is employed. It roughly clusters the short segment with GLR-based criterion until the data in one cluster is enough to use RC-MES adaptation.

We evaluated the performance of the proposed framework on a dataset collected from several conversations and meetings. It consists of about 8 hours utterances with large variation, and utterances whose durations are less than 10s are about 75%. In speaker change detection experiments, slightly decreasing the θ_{GLR} can improve both the recall ratio (0.78) and the precision ratio (0.74) because of the clustering effect in our indexing procedure. The results of speaker indexing experiments demonstrated that the proposed method is not only better but also significantly robust than the other conventional bootstrapping approach. Even if the average bootstrapping duration of each speaker is less than 5 seconds, the indexing accuracy is outstanding.

As a future work, we will evaluate the proposed method on the NIST databases to demonstrate its generality.

Reference

1. Delacourt, P., Kryze, D., Wellekens, C.J.: Detection of Speaker Changes in an Audio Document. In: Proc. Eur. Conf. Speech Commum. Tech (EUROSPEECH), vol. 3, 1195–1198 (1999)
2. Moh, Y., Nguyen, P., Junqua, J.-C.: Towards Domain Independent Speaker Clustering. In: Proc. IEEE Int. Conf. Acoust. Speech. Signal Process (ICASSP), vol. 2, pp. 85–88 (2003)
3. Wu, T., Lu, L., Chen, K., Zhang, H.: UBM-Based Real-Time Speaker Segmentation for Boradcasting News. In: Proc. IEEE Int. Conf. Acoust. Speech. Signal Process (ICASSP), vol. 2, pp. 193–196 (2003)
4. Kwon, S., Narayanan, S.: Unsupervised Speaker Indexing Using Generic Models. IEEE Trans. On Speech and Audio Processing 13(5), 1004–1013 (2005)
5. Thyes, O., Kuhn, R., Nguyen, P., Junqua J.-C.: Speaker Identification and Verification Using Eigenvoices. In: Proc. IEEE Int. Conf. Acoust. Speech. Signal Process. (ICASSP), vol. 2, pp. 242–246 (2000)
6. Aubert, X.L.: Eigen-MLLRs Applied to Unsupervised Speaker Enrollment for Large Vocabulary Continous Speech Recognition. In: Proc. IEEE Int. Conf. Acoust. Speech. Signal Process (ICASSP), vol. 1, pp. 17–21 (2004)
7. Fu, ZH., Zhao, RC.: Speaker Modeling Technique Based on Regression Class for Speaker Identification with Sparse Trainging. In: Li, S.Z., Lai, J.-H., Tan, T., Feng, G.-C., Wang, Y. (eds.) SINOBIOMETRICS 2004. LNCS, vol. 3338, Springer, Heidelberg, GuangZhou, China (2004)
8. Ajmera, J., McCowan, I., Bourland, H.: Robust Speaker Change Detection. IEEE Signal Processing Letters 11(8), 649–651 (2004)

Reversible Data Embedding Technique for Palette Images Using De-clustering

Chin-Chen Chang[1,2], Yi-Hui Chen[2], and Yung-Chen Chou[2]

[1] Department of Information Engineerning and Computer Science,
Feng Chia University, Taichung 40724, Taiwan
ccc@cs.ccu.edu.tw
[2] Department of Compute Science and Information Engineerning,
National Chung Cheng University, Chiayi 62102, Taiwan
{chenyh,jackjow}@cs.ccu.edu.tw

Abstract. In this paper, a reversible data embedding method for color images is proposed. The technique of data embedding is used to hide the secret data into multimedia such as text content, audio, video, and images. In some applications, the reversibility of data embedding is an important requirement, as in distance medical treatment and military industrial applications. Many researchers have proposed reversible techniques that work on grayscale images, but these methods cannot be applied directly to color images, which in recent years have increased in popularity and have more redundant space available for embedding the secret data. Moreover, the size of color image can be reduced by being sorted storing in the palette-based format. Some researchers have proposed data embedding techniques for palette-based images. While their methods successfully achieve the purpose of secret data delivery, they do so without reversibility. In this paper, we propose a data embedding method in palette-based images with reversibility. The simulation result shows that the proposed method embeds 255Kb into a palette image sized 512×512 pixels and that the restored image and the original image are the same.

Keywords: Color palette image, data hiding, declustering reversible, steganography.

1 Introduction

Steganography, also known as data hiding, is the practice of hiding data in cover media in order to securely deliver the secret message to the recipient without attracting attention from an unexpected user. Data hiding techniques can be applied in military, industrial and distance medical treatment applications. Intelligence agents may use this technique to deliver information to an agency. In long distance medical treatment applications, a patient's x-ray may be sent to a remote medical center with private information such as the patient's name, treatment data, and so on.

Data hiding can be classified into two categories. The first category is irreversible data hiding, in which the secret message is extracted without restoring the cover image. The second category is reversible data hiding, in which the secret message is extracted from the embedded image, and the original image is also completely

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 130 – 139, 2007.
© Springer-Verlag Berlin Heidelberg 2007

restored. The reversibility of data hiding techniques is an important property in medical treatment and military industrial applications.

In past years, many researchers [1, 3, 6, 7, 8, 10, 11, 12] developed reversible data embedding methods by using different techniques. However, all of these methods only applied to grayscale images. If these methods are applied to hiding data in color images, then the stego-images have large distortions. With advances in digital devices, color image have become common. In order to reduce the size of a color image, a true color image can be represented as a color palette image. The concept of a palette-based image is to train the significant colors and to use the color's index to represent the image. In this way, just the indices table and palette information need to be kept, and the size of the image will be significantly reduced.

Many researchers have proposed data embedding methods for color images [4, 5, 15, 16]. In 1999 [9], Fridrich proposed a data hiding method to embed secret data into a palette-based image with low distortion. Fridrich's method modifies the color pixels by using a suitable color from a sorted palette. Tzeng *et al*. proposed a data hiding method for palette-based images in 2004 [13]. In their method, a color sequence is constructed by sorting an embedding pixel reference to a color sequence constructed by sorting the neighboring color pixels of the current embedding color pixel. According to the secret data and color sequence, the current embedding pixel is modified in order to imply the corresponding secret data. Furthermore, Chan and Chang [2] proposed a data embedding method for a palette-based image in 2006. Their method utilizes the SMVQ method to embed the secret data. However, the above mentioned methods cannot achieve reversibility.

In this paper, we propose a reversible data embedding method for palette-based images to achieve reversibility. According to the characteristic of natural image, the proposed method references the neighboring pixels of the current embedding pixel with the de-clustering concept to modify the pixel to imply the secret data. To achieve reversibility, in some cases, the proposed method needs extra information to indicate the original color information. In our simulation, a color palette image sized 512 × 512 successfully carries 255 Kb secret data and successful restored the cover image.

The rest of the sections are organized as follows. Section 2 briefly introduces Fridrich's method, Tzeng *et al*.'s method, and Chan and Chang's method. The details of the proposed method are presented in Section 3. The simulation and comparison results are presented in Section 4. Finally, the conclusions are included in Section 5.

2 Related Works

2.1 Fridrich's Method

Fridrich's method embeds secret data into a palette-based image with low distortion. The main concept of Fridrich's method is that the selected pixels are replaced by their closest pixel satisfying the embedding rule, respectively. The key steps of Fridrich's method are described as follows. First, a cover image I is sized $H \times W$ pixels, and the secret message S is sized N_S bits. Then, N_S pixels are selected by using a Pseudo Random Number Generator (PRNG) with a user defined seed. Second, the color parity is computed by $(r+g+b) \bmod 2$, where r, g and b are the Red, Green and Blue values of a color, respectively. If the parity value of a color equals the secret bit, then it needs not

to be changed, otherwise a suitable color is used to replace the embedding pixel. Here, the suitable pixel is the closest color from the current embedding color pixel, and its color parity equals that of the secret bit. The distance between two colors c_i and c_j is computed by Euclidean distance $D(c_i, c_j) = ((r_i-r_j)^2+(g_i-g_j)^2+(b_i-b_j)^2)^{1/2}$, where $c_i=(r_i, g_i, b_i)$ and $c_j=(r_j, g_j, b_j)$. The embedding process is finished when all of the secret bits have been embedded into the cover image.

After the embedding process, the sender sends the embedded image, known as the stego-image, to the receiver over Internet. After that, the secret data can be extracted by following manner. First, the embedding positions are generated by using the same PRNG with a pre-shared seed. Second, the color parity of the embedding position is computed by $(r+g+b)$ mod 2. Finally, the meaningful secret message is gained by reorganizing the extracted data. With Fridrich's method, using the closest color to replace the embedding pixel, the stego-image results in lower image distortion. However, Fridrich's method cannot achieve reversibility.

2.2 Tzeng *et al.*'s Method

Tzeng *et al*. proposed a palette-based data embedding method in 2004 [13]. The main idea of Tzeng *et al*.'s method is to use the current embedding pixel's neighboring pixels to guide the modification in order to imply the corresponding secret bit. Fig. 1 is a diagram of the current embedding pixel and its neighboring pixels. For instance, $p(i, j)$ is the current embedding pixel and $p(i\text{-}1, j\text{-}1)$, $p(i\text{-}1, j)$, $p(i\text{-}1, j\text{+}1)$, and $p(i, j\text{-}1)$, a colored gray area, are $p(i, j)$'s neighboring pixels.

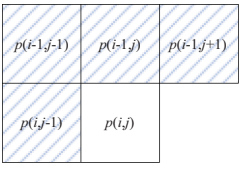


Fig. 1. The current embedding pixel $p(i, j)$ and its neighboring pixels

Data embedding can be done by the following data embedding operations. First, the neighboring pixels of the current embedding pixel are sorted by their luminance values. The luminance value is calculated by the following equation:

$$v = 0.3 \times r + 0.59 \times g + 0.11 \times b, \tag{1}$$

where v, r, g, and b denote to the luminance, red, green, and blue values of a color, respectively [14]. The sorting rule is defined as follows:

$$\begin{cases} c_1 > c_2, & \text{if } (v_1 > v_2) \text{ or} \\ & \quad (v_1 = v_2 \text{ and } r_1 > r_2) \text{ or} \\ & \quad (v_1 = v_2 \text{ and } r_1 = r_2 \text{ and } g_1 > g_2) \\ c_1 = c_2, & \text{if } (r_1 = r_2 \text{ and } g_1 = g_2 \text{ and } g_1 = g_2) \\ c_1 < c_2, & \text{otherwise} \end{cases}, \tag{2}$$

where c_1 and c_2 represent two color pixels. Second, it is decided whether the pixel can be embeddable or not. An embeddable pixel satisfies the following three conditions: (1) number of different colors of neighboring colors is larger than the predefined threshold T_c, (2) the distance between current color pixel and the largest color pixel in neighboring pixels is smaller than a predefined threshold T_d, (3) if the pixel satisfies Conditions 1 and 2, then the distance between the embedded color pixel and the largest color pixel in the neighboring pixels must be smaller than the predefined threshold T_d. Third, the current embedding pixel is modified by the following equation to imply the secret data.

$$f(c_0, c_1, c_2, c_3, c_4) = \begin{cases} 0, \text{ if } c_0 \geq c_1 \\ 1, \text{ if } c_1 > c_0 \geq c_2 \\ 0, \text{ if } c_2 > c_0 \geq c_3 \\ 1, \text{ if } c_3 > c_0 \geq c_4 \\ 0, \text{ otherwise} \end{cases}, \tag{3}$$

where c_0 is the embedded color pixel and c_1, c_2, c_3, and c_4 are the sorted neighboring color pixels.

The data extraction operation follows the embedding conditions to extract the secret data by using Eq. (3). Tzeng *et al.*'s method successfully achieves the delivery of data, but the stego-image cannot be restored to the original image.

2.3 Chan and Chang's Method

In 2006 [2], Chan and Chang proposed a data embedding method for palette-based images by using the concept of Side-match Vector Quantization (SMVQ). According to the characteristics of the palette image, the color palette can be seen as a codebook which contains N_C colors, also called codewords. Chan and Chang's method embeds the secret data in the following manner. First, the palette is sorted by principal component analysis or by luminance-sorted algorithm. Second, the sub-palettes of each pixel are constructed by following the concept of SMVQ. Finally, according to the secret data, the index of the color pixel is modified by using their embedding rules.

Chan and Chang's method not only embeds the secret data into a cover image but also maintains the visual quality of the stego-image by using a predefined threshold T. T is the maximum bound of distortion for a color pixel. An extra bitmap is used to denote each pixel's type because that pixel may be modified for data embedding, or not used in order to maintain visual quality.

The data extraction operation is used to extract the secret data when a receiver received the stego-image. The secret data is extracted by the following steps. First, the indices table and the extra bitmap of the stego-image are reconstructed when the receiver receives the stego-image. Second, according to the indicator, the secret data is extracted by the extraction rules. Thus, Chan and Chang's method successfully delivers the data. However, their proposed method cannot achieve reversibility.

3 The Proposed Method

According to the characteristics of a natural image, a smooth area has a similar color pixel distribution. Based on this property, we utilize the concept of de-clustering to modify the color pixels in order to imply the corresponding secret data. Each pixel in a color palette is composed three primary colors, *Red*, *Green* and *Blue*, denoted as *r*, *g* and *b*, respectively. Before data embedding and extracting, pixels in the color palette must be sorted by the luminance-sorted algorithm with their luminance values. The luminance value *v* of a pixel can be calculated by Eq. (1).

An important property of a sorted color palette is that the similar color has the closer indices. Thus, the differences between two colors should be large while the indices are far apart. According to this property, an efficient de-clustering strategy is adopted in our proposed scheme. That is, as shown in Fig. 2, a color palette with size *n* can first be divided into two parts. The first part covers indices from 0-th to (n/2-1)-th. Another part consists of the indices from (n/2)-th to (n-1)-th. Color *i* of first part can pick up its dissimilar codeword (i+n/2) from another part. The de-clustering construction can be done by sender and receiver themselves and used to embed or to extract the secret data.

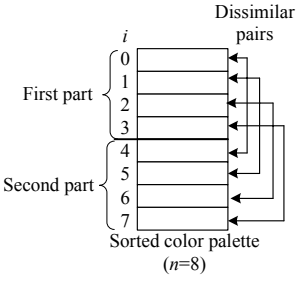


Fig. 2. Example of the color palette's de-clustering

3.1 Embedding Procedure

The sender embeds the secret data into a palette-based image by using the embedding procedure. Fig. 3 illustrates the key steps of the embedding procedure. The designed hiding procedure is divided into two hiding strategies: *Phase-1* and *Phase-2*. The notation *b* in Fig. 3 is to represent a secret bit. Before demonstrating the hiding strategies, an estimation policy is for the inputted pixel to determine that secret data can be embedded by using either the *Phase-1* or *Phase-2* hiding strategy. First, the encoder scans image *I* by raster-scanning order and looks for its past causal neighbors for each scanned pixel, as shown in Fig. 4.

The number of causal neighbors is adaptive according the value of *N*, where *N* means the pixels of N_{th} layer near the current pixel. For example, the causal neighbors of pixel P_0 are $\{x_1, x_2, x_3, x_4\}$ and $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$ once N = 1 and N = 2, respectively. Subsequently, we can use the causal neighbors of the inputted pixel to generate a predicted pixel as P_k. The predicted pixel value is composed of components *R*, *G* and *B*, which are calculated by averaging all the values of its causal

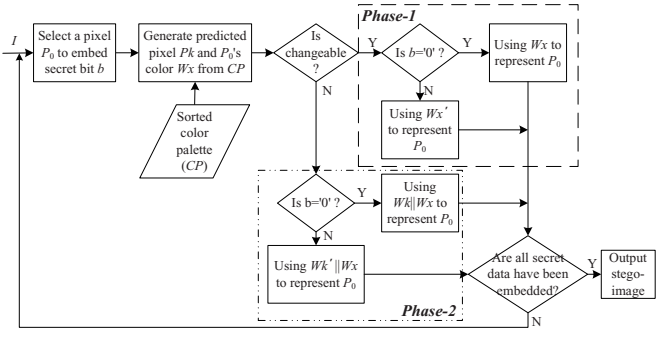


Fig. 3. The design of embedding procedure

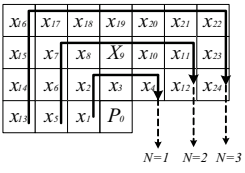


Fig. 4. The nearest causal neighbors of current pixel P_0

neighbors' components *R*, *G* and *B*, respectively. Next, from the sorted color palette, the encoder finds the best codewords *Wr* and *Wx* for the predicted pixel *Pk* and the inputted pixel P_0, respectively. In addition, from the de-clustering results, the dissimilar codewords of *Wr* and *Wx* can be looked for, and they are depicted as *Wr* and *Wx*. Finally, a temporary codeword *Wk* can be generated by calculating $\lfloor Wr+Wr' \rfloor$ and also its corresponding dissimilar codeword *Wk* can be found from the de-clustering results.

Furthermore, the encoder decides to adopt either *Phase-1* or *Phase-2* to hide the secret data according to the determination of the estimation policy. For clarity, the distortion between *A* and *B* is denoted as $D(A, B)=|A - B|$. In the estimation policy, the encoder first computes the distortions between pixel *Pk* and its best codeword *Wx* and dissimilar codeword *Wx* to denote as $D(Pk, Wx)$ and $D(Pk, Wx)$, respectively. The estimation policy is that the pixel can embed secret data used in *Phase-1* while both $D(Pk, Wx)$ is less than $D(Pk, Wx)$ and *Wx* is not equal to *Wk*. If the pixel does not satisfy any one of the above conditions, it hides secret data by *Phase-2* data embedding instead of *Phase-1*.

Phase-1 data embedding is a simple method in which the pixel is replaced with codeword *Wx* while the secret bit is 0. Otherwise, it is replaced with codeword *Wx* In *Phase-2*, the pixel is encoded with codeword *Wx* and indices of codewords *Wk* and *Wk* are used to denote the hidden data as '0' and '1', respectively. That is, if the secret data is '0', the encoder concatenates the index of codeword *Wk* in front of *Wx*. Conversely, the pixel is encoded by indices of codewords *Wk* and *Wx* once the secret data is '1'. Later, the encoded indices are sent to the receiver side, and no extra data should be reserved.

3.2 Extracting Procedure

In the extracting procedure, pixels are directly decoded by the color palette when they are located in the first row and the first column because they are not used to hide any secret data. For the residual pixels, receivers can apply the extracting strategies to extract the secret data and reconstruct the stego-image. Notably, the quality of stego-image is the same as that of the image encoded by the color palette without any secret data embedded. The flowchart of extracting procedures is shown in Fig. 5. The extracting strategy is divided into *Phase-1* and *Phase-2* to extract secret data which are embedded by *Phase-1* and *Phase-2* data embedding policies, respectively. First, the decoder scans image *I* in a raster-scanning order and computes a predicted pixel *Pk*, a temporary codeword *Wk*, and its corresponding dissimilar codeword *Wk*′, which are described in Subsection 3.1. The compression code is transformed into index *Wd*. While *Wd* is equal to either *Wk* or *Wk*′, the pixel can be decoded with *Phase-2*. Otherwise, it must be decoded with *Phase-1*. After *Phase-1* and *Phase-2* extracting, the decoder can extract the secret data and reconstruct the stego-image simultaneously.

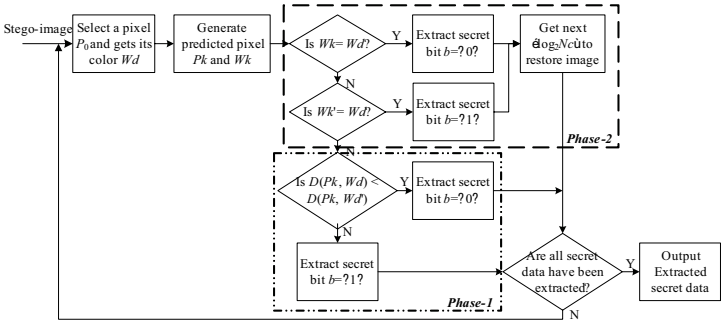


Fig. 5. Flowchart of data extracting procedure

In *Phase-1* extracting, decoder computes the distortions $D(Pk, Wd)$ and $D(Pk, Wd')$ first. Then, the decoder can decode the pixel with codeword *Wd* and extract the secret bit '0' when $D(Pk, Wd)$ is less than $D(Pk, Wd')$. Otherwise, the pixel is decoded by codeword Wd' and the retrieved secret bit is '1'.

In the *Phase-2* extracting policy, the decoder retrieves the next compression code to transform it into an index *Wb*. Then the decoder can reconstruct the pixel with the index *Wb*. The hidden data is judged as '0' and '1', and for the related index, *Wd* is *Wk* and *Wk*′, respectively. That is, if *Wd* is the same as *Wk*, the secret bit is '0'. Otherwise, the retrieved bit is '1'.

4 Experimental Results

In our experiments, three color images of 512×512 pixels were served as test images: "Lena", "Pepper" and "Baboon," as shown in Fig. 6. Two measures, hiding capacity and *PSNR* (peak signal-to-noise ratio), are used to evaluate the performance of our

proposed scheme. The hiding capacity represents the number of secret data embedded into a color image. The value of *PSNR* as defined in Eq. (4) simply indicates the image quality of a stego-image generated by our proposed scheme.

$$PSNR = 10\log_{10}\frac{255^2}{MSE}dB. \tag{4}$$

Here, 255 represents the maximum value of each pixel, and the *MSE* (mean square error) for an image is defined in Eq. (5).

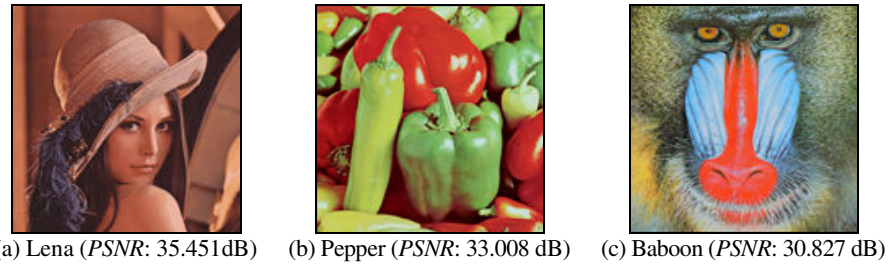
$$MSE = (\frac{1}{ht \times wd})\sum_{i}^{ht}\sum_{j}^{wd}(p_{ij} - p'_{ij})^2. \tag{5}$$

Here, the notations *ht* and *wd* represent the height and width of an image, respectively. p_{ij} is the pixel value of the coordinate *(i, j)* in an original image, and p'_{ij} is the pixel value after embedding processing. Because all test images are color images and each color image consists of three planes, we obtain the *PSNR* for a color image from averaging the *PSNRs* of these three planes in the following experiments.

(a) Lena (b) Pepper (c) Baboon

Fig. 6. The three 512×512 color images used as test images

In general, a higher image quality results in a higher PSNR value. On the contrary, if the PSNR value is small, that indicates that the image quality is worse. Actually, it is difficult for the human eye to detect the difference between a stego-image and its original image when the PSNR value of the stego-image is larger than or equal to 30 dB in the human visual system. The visual qualities of the stego-images "Lena",



(a) Lena (*PSNR*: 35.451dB) (b) Pepper (*PSNR*: 33.008 dB) (c) Baboon (*PSNR*: 30.827 dB)

Fig. 7. Three reconstructed images

"Pepper" and "Baboon" reconstructed in the extracting procedure are shown in Figs. 7(a)—7(c), respectively. Notably, PSNRs presented in Fig. 7 are the same as those of the original images.

The performance of our proposed scheme in the image quality of reconstructed image and hiding capacity are listed in Table 1. In essence, each pixel except for the first row and the first column pixels can embed one secret bit, so each color image can embed about 255 Kb secret data into it.

Table 1. The *PSNRs* and hiding capacity (*H*) of our proposed scheme

Lena		Pepper		Baboon	
PSNR	*H*	*PSNR*	*H*	*PSNR*	*H*
35.446 dB	255 Kb	33.007 dB	255Kb	30.827 dB	255Kb

The comparisons between our and other schemes are listed in Table 2. Notably, it shows the comparative results of three color images in terms of image quality *PSNR* and hiding capacity *H*. Here, the used color palette size is $n = 256$. Among them, our proposed scheme can outperform others under the same hiding capacity while offering a higher image quality of the reconstructed image.

Table 2. The comparisons of image qualities *PSNRs* between others' scheme and our proposed scheme under the same hiding capacity *H*

Images / Methods	Lena		Pepper		Baboon	
	PSNR (dB)	*H* (Kb)	*PSNR* (dB)	*H* (Kb)	*PSNR* (dB)	*H* (Kb)
Fridrich [7]	31.89	255	28.78	255	26.72	255
Tzeng *et al*. [11]	33.28	255	29.43	255	26.52	255
Chan and Chang [2]	33.52	255	30.56	255	25.22	255
The proposed scheme	35.44	255	33.00	255	30.82	255

5 Conclusions

Digital images are widely used as cover media for secret data delivery. With advances in digital equipment, color images have increased in popularity. The proposed method effectively utilizes the characteristics of a color image with the concept of declustering to embed the secret data into a palette-based image. In the experimental results, the proposed method not only achieves the purpose of secret data delivery but also restores the stego-image to the original. In addition, the proposed method hides 255 Kb secret data into a palette-based image sized 512×512 pixels. In a word, we have a proposed a method to achieve the purpose of reversibility of data embedding for medical, military, and other applications.

References

1. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. IEEE Transactions on Image Processing 13(8), 1147–1156 (2004)
2. Chan, C.S., Chang, C.C.: A color image hiding scheme based on SMVQ and modulo operator. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4352, pp. 461–470. Springer, Heidelberg (2006)
3. Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recognition 37(3), 469–474 (2004)
4. Chang, C.C., Tsai, P.Y., Lin, M.H.: An adaptive steganographic scheme for color image. Fundamenta Informaticae 62(3–4), 137–146 (2004)
5. Chang, C.C., Wang, Y.C., Hu, Y.C.: An adaptive data hiding scheme for palette images. In: Pan, J.S., Fang, W.C., Huang, H.C., Jain, L.C. (eds.) Intelligent Multimedia Data Hiding, Chapter 5, pp. 103–122. Springer-Verlag, Berlin-Heidelberg, Germany (2006)
6. Chang, C.C., Lu, T.C.: A difference expansion oriented data hiding scheme for restoring the original host images. Journal of Systems and Software 79, 1754–1766 (2006)
7. Chang, C.C., Lin, C.Y.: Reversible Steganography for VQ-compressed Images Using Side Matching and Relocation. IEEE Transactions on Information Forensics and Security 1(4), 493–501 (2006)
8. Chang, C.C., Tai, W.L., Lin, C.C.: A Reversible Data Hiding Scheme Based on Side Match Vector Quantization. IEEE Transactions on Circuits and Systems for Video Technolo 16(10), 1301–1308 (2006)
9. Fridrich, J.: A new steganographic method for palette-based images. In: Proceedings of IS&T PICS Conference, Savannah, Georgia, April 1999, pp. 285–289 (1999)
10. Fridrich, J., Goljan, M.: Practical steganalysis of digital images—state of the art. In: Delp, E.J., Wong, P.W. (eds.) SPIE. Proceedings of SPIE Security Watermarking Multimedia Contents IV, vol. 4675, pp. 1–13 (2002)
11. Ker, A.: Steganalysis of LSB matching in grayscale images. IEEE Signal Processing Letters 12(6), 441–444 (2005)
12. Tian, J.: Reversible data embedding using a difference expansion. IEEE Transactions on Circuits and Systems for Video Technology 13(8), 831–841 (2003)
13. Tzeng, C.H., Yang, Z.F., Tsai, W.H.: Adaptive data hiding in palette images by color ordering and mapping with security protection. IEEE Transactions on Communications 52(5), 791–800 (2004)
14. Tzeng, C.H., Yang, Z.F., Tsai, W.H.: Adaptive data hiding in palette images by color-ordering and mapping with security protection. IEEE Transactions on Communications 52(4), 791–800 (2004)
15. Wang, X., Yao, T., Li, C.T.: A palette-based image steganographic method using colour quantization, In: Proceedings of IEEE International Conference on Image Processing, vol. 2, pp. 1090–1093 (September 2005)
16. Wu, M.Y., Ho, Y.K., Lee, J.H.: An iterative method of palette-based image steganography. Pattern Recognition Letters 25, 301–309 (2004)

The PT-2 Caching Algorithm in the Transcoding Proxy Cluster to Facilitate Adaptive Content Delivery

YongJu Lee[1], YuHyeon Bak[1], OkGee Min[1], HagYoung Kim[1], and CheolHoon Lee[2]

[1] Dept. of Internet Server Group, Digital Home Research Division, Electronics and Telecommunications Research Institute, 161 Gajeong-Dong, Yuseong-Gu, Daejeon, Korea
[2] System Software Lab., Dept. of Computer Engineering, Chungnam National University, Daejeon 305-764, Korea
{yongju,bakyh,ogmin,h0kim}@etri.re.kr
chlee@ce.cnu.ac.kr

Abstract. With the advent of wireless network devices, clients such as PDAs and mobiles have become heterogeneous in terms of their capabilities. For the purpose of maintaining different categories of clients, transcoding has emerged as a major field in adaptive multimedia delivery. However, little work has been performed in the area of caching transcoded streaming media in a cluster environment. We therefore propose a new partial caching algorithm for improving cache utilization and adapting heterogeneous content delivery to reduce startup latency.

1 Introduction

Adaptive content delivery over the Internet must address both the multimedia nature of the content and the capabilities of diverse client categories. Specifically, transcoding is the major issue in terms of scalable multimedia delivery to various mobile devices. Traditional caching research has focused on optimizing the caching algorithm to cope with space limitation and network capability. Furthermore, an area of transcoding localizes Web content such as image or text, and studies have paid relatively little attention to transcoded streaming data caching in a cluster environment. In this paper, we propose the PT-2 caching algorithm for efficiently transcoded a streaming set of heterogeneous multimedia content to facilitate seamless playback via adaptive delivery in a cluster environment. Experimental results show that the proposed caching algorithm is efficient with less than computation overhead, effective with cache hit and transcoding rates. The remainder of this paper is organized as follows. In

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 140–150, 2007.
© Springer-Verlag Berlin Heidelberg 2007

section 2, we describe recent research in various fields. In sections 3 and 4, we provide a detailed description of the PT-2 caching algorithm and a cluster model which comprised of several transcoding-enabled servers and a streaming dedicated server for the adaptive delivery, and discuss the adaptation scheme in our original media streaming server equipped with streaming accelerated hardware. In section 5, we evaluate the simulation results, and in section 6, we summarize our findings.

2 Related Work

Much work has focused on adapting multimedia content delivery by various means. With regard to Web content and mobile client capabilities, the InfoPyramid has two key components that define a representation scheme to provide multimodal, multi-resolution,and the selection method is achieved via client capabilities. In addition, the InfoPyramid contains individual components of the multimedia content, such as text, images, video, and audio,that has to be adapted to different client devices, and customizes its characteristics and processes via off-line transcoding, not by caching or streaming[1]. Traditional partial caching schemes store initial parts of popular videos on the proxy to reduce playback latency and achieve better performance via consideration of the IP network, the prefix size, and so on[2,3]. In content layering or versions[4], the video streaming of multiple versions is compared to that of multiple layers in a caching environment,and mixed distribution/caching strategies have been shown to provide the best overall performance. However, layer-encoded formats (e.g., scalable coding) may not always be available in the real world. TranSquid [5] maintains the server-directed transcoding information as part of meta-data and uses this information to convert its fidelity, modes, and user heuristics, while considering three distinct events (miss, partial hit, and hit). When the cache already contains a higher fidelity variant but no object, a partial hit event occurs. Outstanding transcoding-enabled algorithm for adaptive content delivery are the full version only(FVO) and transcoded version only(TVO) schemes[6], and TEC[7,8]. FVO and TVO utilize caching algorithms with transcoding capabilities, in which the video object is cached in the proxy. Under the TEC scheme, Transcoding-enabled Caching(TeC), which is defined by TEC-11, TEC-12 and TEC-2, also defines three distinct events(exact hit, transcode hit, and miss) in a TeC Proxy. Tests of their performance using a synthesized and enterprized trace-driven simulation showed that TEC-11 and TEC-12 perform better than TEC-2 under similar network capacity conditions, and that the performance of TEC-2 is superior with heterogeneous network connectivity. However, all TEC algorithms evict one complete version even if the newly transcoded version is sufficiently smaller than the size of the victim. Moreover, they are not considered in a cluster environment, which is necessary to reduce a high CPU workload.

3 PT-2 Caching Algorithm

Let us assume that the origin content server has various video objects: S is the sum of Si and each video object O has n versions at bitrates b_0,b_1,. . .,b_{n-1} (i.e. $b_0 > b_1$,.., $> b_{n-1}$) and the transcoding-enabled proxy has various content versions. For example, if Ob_0 is highest bitrate version of video object O and O_{b1} is next highest one, Otb_{0-1}(b_0>tb_{0-1}>b_1) is a transcoded bitrate version of video object Ob_0, and a video object Otb_{0-1} is divided into two equal parts. A partial version of a video object defines a prefix object of two, and a suffix object of two is evicted at the time of replacement.
We define the following events in a transcoding proxy:

** Full Fetch (FF)*: the requested or transcodable version of the video object does not exist in a transcoding proxy and the requested version does not need to be transcoded.
** Full Hit (FH)*: the exact requested version of the video object has a transcoding proxy.
** Full Transcoding (FT)*: the requested version of the video object does not exist in a transcoding proxy, but has a transcodable version.
** Full Fetch Transcoding (FFT)*: the requested or transcodable version of the video object does not exist in a proxy cache and the transcodable version has been fetched and transcoded from the origin content server.
** Partial Hit and Fetch (PHF)*: the exact requested partial version of the video object does exist in a transcoding proxy, but does not have the versions suffix.
** Partial Transcoding and Fetch Transcoding (PTFT)*: the exact requested partial version of the video object does not exist in a proxy cache, but a transcodable partial version exists. A transcoding proxy has transcoded the partial transcodable version and fetched the suffix object from the origin server and has transcoded it.
When a transcoding proxy in the cluster does not have an exact or transcodable version such as FF, FFT, PHF, PTFT in the fetch event, we also define additional events in a transcoding proxy cluster:
** Sibling Full Hit (SFH)* : the exact version of the video object exists in other proxies.
** Sibling Partial Hit (SPH)* : the exact partial version of the video object exists in other proxies.
** Sibling Full Transcoding Hit (SFTH)* : the exact version does not exist in other proxies, but a transcodable exact version exists in other proxies.
** Sibling Partial Transcoding Hit (SPTH)* : a transcodable partial version exists in other proxies.

The pseudocode for our algorithm is below(for simplicity, evicting the actual threshold value has not been included.)

Table 1. Partial Transcoded Multiple Version(PT-2) Caching Algorithm

Version bi of a video object in the transcoding proxy cluster is requested.
if *bi's exact full version is already in the proxy cluster*
// Sibling Full Hit(SFH)
Redirect the current request to other proxy **if** *necessary.*
// Full Hit(FH)
Stream Obi to the user in the proxy cache.
else if *bi's exact partial version is already in the proxy cluster*
// Sibling Partial Hit(SPH)
Redirect the current request to other proxy **if** *necessary.*
// Partial Hit and Fetch(PHF)
If the proxy cache size is not large enough **for** *suffix-Obi's fetching,*

Call Evict-algorithm
Stream prefix-Obi and suffix-Obi after fetching
else if *bi's transcodable full version is already in the proxy cluster*
// Sibling Full Transcoding Hit(SFTH)
Redirect the current request to the other proxy **if** *necessary.*
// Full Transcoding(FT)
If the proxy cache size is not large enough **for** *the transcoded version,*
Call Evict-algorithm
Stream Otbi to the user after transcoding Obi
else if *bi's transcodable partial version is already in the proxy cluster*
// Sibling Partial Transcoding Hit(SPTH)
Redirect the current request to the other proxy **if** *necessary.*
// Partial Transcoding and Fetch Transcoding(PTFT)
If the proxy cache size is not large enough **for** *sum of prefix-Otbi and suffix-Otbi and suffix-Obi, Call Evict-algorithm*
Stream Otbi to the user after transcoding Obi
else if *bi's transcodable full or partial version does not exist and transcoding is needed*
// Full Fetch Transcoding(FFT)
Choose one of the cluster nodes in order of cache space, CPU, and redirect the proxy.
If the proxy cache size is not large enough **for** *sum of Obi and Otbi's sizes, Call Evict-algorithm*
Stream Otbi to the user after transcoding Obi
else
// Full Fetch (FF)
Choose one of cluster nodes in order of cache space, CPU and redirect the proxy.
If the proxy cache size is not large enough **for** *Obi, Call Evict-algorithm*
Fetch Obi from the origin server and stream it to the user.
Evict-algorithm: Call Evict-folding or -trash until there is enough space.
Evict-trash: Trash victims using the Least Recently Used(LRU) algorithm.
Evict-folding: Folding victims using the LRU algorithm, divide by two(e.g., in equal), trash suffix-object of two.

4 Transcoding the Proxy Cluster with Streaming Accelerated H/W

4.1 Transcoding the Proxy Cluster Model

Our transcoding proxy cluster can be implemented in a network in the form of several transcoding proxy streaming severs(TPSSs) and a content streaming server(CSS) to perform streaming and transcoding on-the-fly. There are same nature that TPSS and CSS can deliver the streaming data to the client, but especially, TPSS has the additional functionality that it transcodes streaming data to adaptive stream rate via hardware Digital Signal Processing units(DSPs) or software transcoders. Adaptive clients such as PC, portable PDAs, and laptops request stream data via the proxy cluster. Then the proxy cluster calls on a dispatcher node, which determines a suitable node depending on whether only streaming or on-the-fly transcoded streaming. The dispatcher node redirects its request to the user player using the RTSP REDIRECT method which informs the client that it must connect to another server location[9,10]. A selected node delivers the specific streaming service to the user. At this time, additional connection between client and server are established so that the information including the network usage, the CPU overhead, and etc, can be exchanged.

4.2 CSS/TPSS Architecture

As shown in Fig. 1, a media streaming server consists of a main process (hereafter "MP") and an ns process (hereafter "NP") whose number is the same as the number of the network storage card[11]. The media streaming server also has one or two processors, system memory, a memory controller hub(or bridge), an I/O

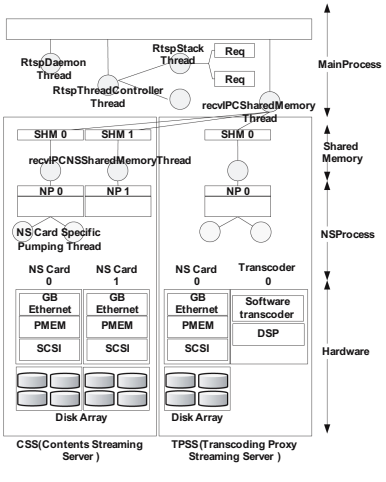


Fig. 1. MainProcess(MP)/NSProcess(NP) Architecture

controller hub, and several PCI/PCI-X controller hubs. In particular, for efficient storage, retrieval, and transmission of large amounts of video data, the server node employs a new H/W device called a network/storage card(NSCard)[12,13]. A NSCard comprises a stream disk array driver, a PMEM driver, and a TOE driver. A special purpose disk array(SDA) is optimized for large sequential disk access and a PMEM equipped with DRAM memory is dedicated to a temporary buffer on disk-to-network data path. A TOE driver is enabled to transmit video data from the PMEM to the network interface directly.

An RtspDaemon thread of the media streaming server creates an RtspThreadController thread, which controls the creation and assignment of an RtspStack thread and has a FIFO, which maintains several RTSP connections. The RtspThreadController thread also passes an RTSP connection to the RtspStack Thread through PIPE. The RtspStack thread manages the maximum number of RTSP connections, manipulates RTSP messages, and stores them to the specific shared memory area. If a recvIPCNSSharedMemory thread in a specific NP receives an RTSP request, the session manager adds it to the session manager's map table. In the case of CSS, we can equip two NScard PCI devices, and focus on the streaming capacity. In the case of TPSS, we can extend the server configuration for the heterogeneous client environments.

4.3 Multiple Buffering Scheme for Adaptation

Figure 2 illustrates the overall client demultiplexing and decoding procedures with newly encapsulated components, which are depicted as a brighter-box in this figure. When streaming data is received by the client, it is received through the DataReceiver/Controller. The data then passes through the CurrentReceiverBuffer to general decoding components and it is ready to be decoded by the Codec as drawn in the shaded-box. At this time, each decoding unit must be reported to the AdaptiveAnalyzer which allows the DecisionEngine to determine which version is appropriate to guarantee seamless playback. In detail, a video decoder on a fast machine decodes all I frames immediately, but a slower machine can decode an I picture if it has enough time to decode it before displaying it. Hence, the IPictureDecodingRate of the video object Obi is calculated by

$$IPictureDecodingRate(Obi) = t0 - t > tauI + VDELTA \text{ in given period,} \quad (1)$$

where t0 is the presentation timestamp of picture 0, t represents the current

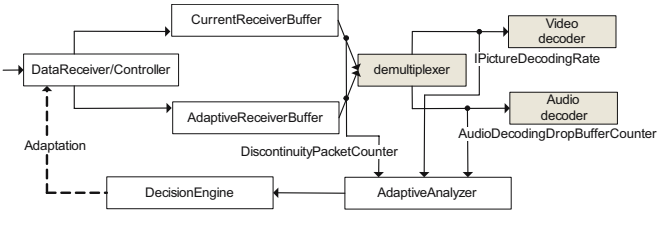


Fig. 2. Demultiplexing and Decoding Enhanced Layer

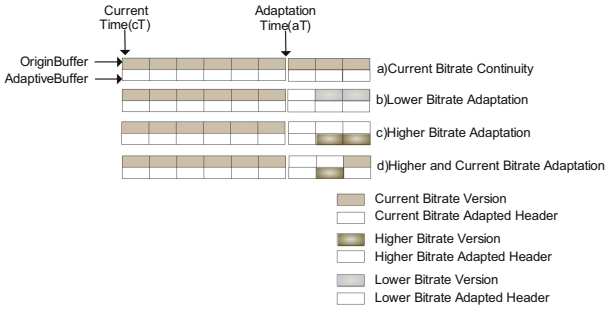


Fig. 3. States of Each Adaptation Phase

timestamp at the moment of decoding, tauI is the mean time to decode an I picture, and VDELTA is a given error margin. The AudioDecodingDropBuffer-Counter is also calculated by

AudioDecodingDropBufferCounter(Obi) = t0 - t > ADELTA in given period. (2)

The DiscontinuityPacketCounter is obtained through the discontinuity indicator of a stream in the case of MPEG-2 TS.

For adaptive playout buffering, we use multiple buffers, one for current version's stream, the other for higher version's adaptation. We have experimentally determined that lower version's individual buffer is not adequate because of shorter playback delay. Figure 3a) and b) depicts the current bitrate version's continuity in general, the lower bitrate adaptation at the OriginBuffer to cope with network congestion,respectively. In case of 3c), the DecisionEngine attempts to change the current version's stream to a higher version's after timing delay. In this case, the decoder receives the higher version's adaptation header information for seamless streaming. In case of 3d), if attempting the higher version leads to poor performance because of network utilization, a rollback sequence is performed with the previous bitrate version in the OriginBuffer and then the AdaptiveBuffer's data are trashed permanently.

5 Simulation Results

A simulation using the proposed algorithm was measured via experiments with a proxy workload. The synthesized trace's parameters were 500 original video objects and each object had four bitrate original versions with their running time varying from 300 to 600 seconds. Its popularity was assumed by a Zipf distribution with of 0.47[14] and an access arrival interval time of 14 seconds through a Poisson process [15]. Figure 4 illustrates the variance of total cumulative gigabytes as a cache size increases up to 8Gb. Figure 4(a) shows the cumulative transfer bytes. TEC-11 exhibited the smallest transfer size; its cumulative transcoded bytes were also the highest and its cumulative hit bytes were the lowest because TEC-11 always utilizes the higher bitrate version. Therefore, TEC-11 has great advantage with high CPU computing power. The TEC-12

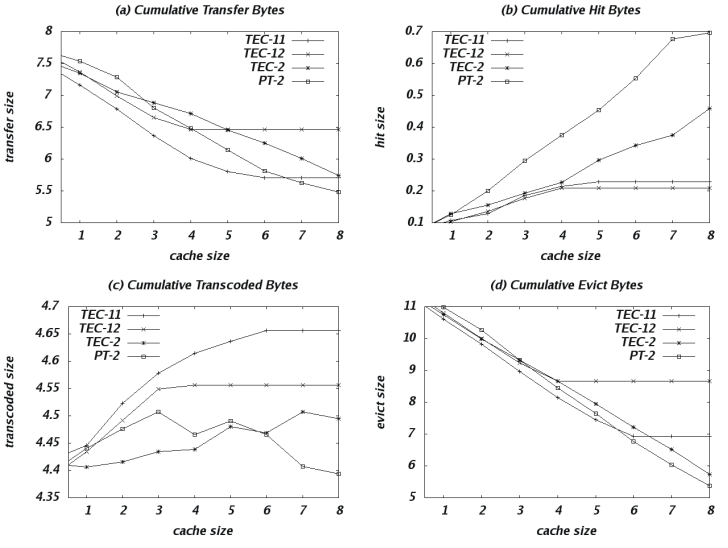


Fig. 4. Variance of Cumulative Transfer/Hit/Transcoded/Evict Bytes

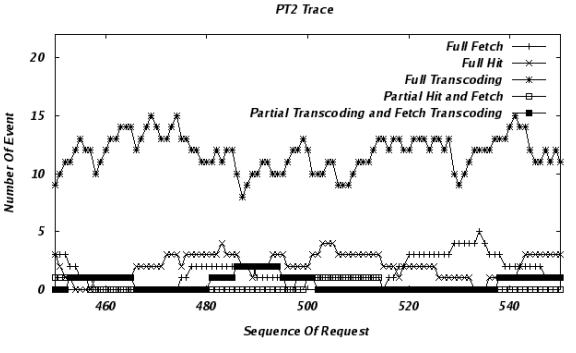


Fig. 5. Event-driven Concurrent Sessions in our PT-2 Algorithm

result was lowest transcoded bytes, but its hit bytes were not performed better than PT-2's. TEC-2 with multiple versions' caching showed the lowest transcoder rates. However, its transfer data and hit bytes were not better than our PT-2 algorithm as its cache space was larger. Our PT-2 algorithm provided the best results in terms of highest hit bytes and lower transfer and transcoded bytes.

Figure 5 presents the concurrent number of sessions as request sequences increased up to 4GB cache size. The results from the 450th to the 550th requests are shown. The average number of full transcoding events was from 10 to 15, and these events' number was diminished by the chances of occurring partial events at the rate of 11 - 16 percents.

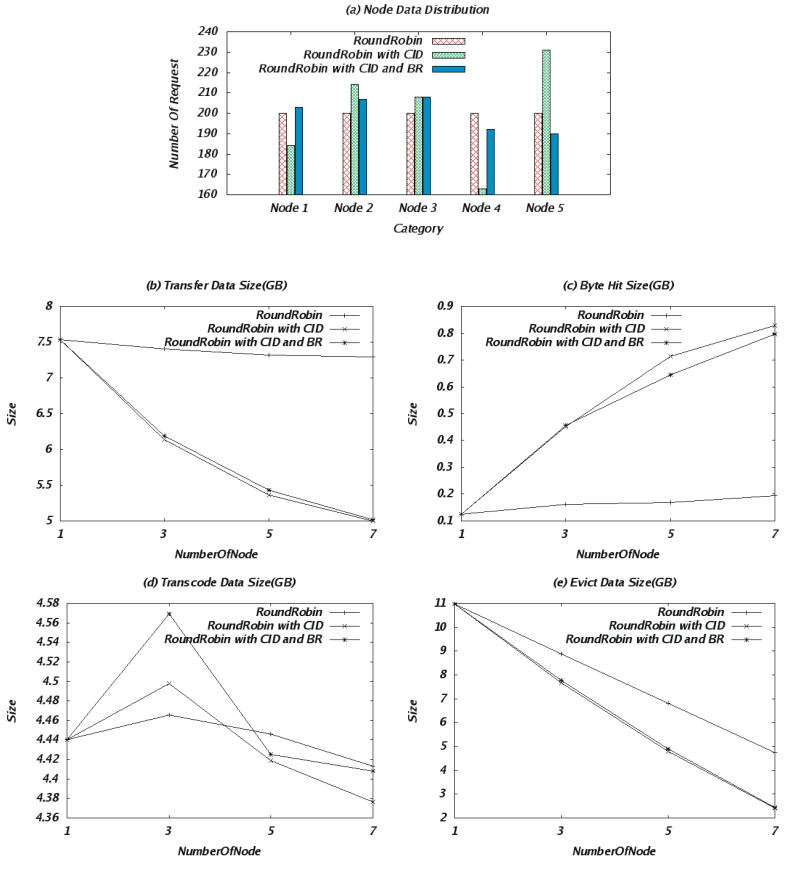


Fig. 6. Cluster Nodes' Workload Distribution and Cumulative Bytes

To extend our results to cluster nodes, we repeated our experiments with the following simple round-robin approach under various node configuration. Figure 6 shows the cluster nodes' workload distribution and cumulative bytes in our PT-2 caching algorithm. In terms of node data distribution, each node is processing equal number of requests at any instant in a round-robin fashion. Figure 6(a) shows diverse distribution patterns. Pure *RoundRobin* is exactly equal number of requests in all nodes, and *RoundRobin with CID* is extremely scattered because of considering only similar contents. *RoundRobin with CID and BR* is moderately distributed[16].

From Fig. 6(b) and (c), we can see that pure round-robin requests had the highest transfer data ratio and the lowest hit data size, implying that caching similar contents in the same node such as *RoundRobin with CID (and BR)* is more effective in saving network bandwidth. As shown in Fig. 6(d), the

round-robin requests with CID and bitrate(BR) had the highest transcoded data rates with high data concentration and heavy CPU costs. Figure 6(e) also shows *RoundRobin with CID (and (BR))* has similar evict sizes.

6 Conclusion

In this paper, we propose a PT-2 caching algorithm for adaptive content delivery in the transcoding proxy cluster to improve the accessibility of various client devices. Preliminary simulation results show that our algorithm performs well with few negative effects in the case of smaller cache size, and also demonstrate its effectiveness as a content distribution model in terms of cluster nodes' workload and byte ratio. In our ongoing work, we are attempting to equip a hardware-level DSP to improve both the transcoding capacity and heuristic measurement for selection among cluster nodes.

References

1. Mohan, R., Smith, J.R., Li, C.-S.: Adapting multimedia Internet content for universal access. IEEE Trans. Multimedia 1(1), 104–114 (1999)
2. Sen, S., Rexford, J., Towsley, D.: Proxy prefix caching for multimedia streams In: Proc. IEEE INFOCOMM'99, New York, NY, USA, pp. 1310–1319 (March 1999)
3. Chen, S., Shen, B., Wee, S., Zhang, X.: Adaptive and lazy segmentation based proxy caching for streaming media delivery. In: Proc. ACM NOSSDAV 2003, Monterey, CA, USA, pp. 22–32 (June 2003)
4. Hartanto, F., Kangasharju, J., Reisslein, M., Ross, K.W.: Caching video objects: Layers vs Versions? In: Proc. IEEE ICME 2002 August 2002, Lausanne, Switzerland (2002)
5. Maheshwari, A., Sharma, A., Ramamrithan, K., Shenoy, P.: TranSquid: Transcoding and caching proxy for heterogeneous e-commerce environments In: Proc. IEEE RIDE 2002, February 2002, San Jose, CA,USA (2002)
6. Tang, X., Zhang, F., Chanson, S.: Streaming media caching algorithms for transcoding proxies Proc. ICPP 2002, August 2002, Vancouver, Canada (2002)
7. Shen, B., Lee, S.-J., Basu, S.: Caching strategies in transcoding-enabled proxy systems for streaming media distribution networks HP Laboratories, Technical Report HPL-2003-261 (December 2003)
8. Shen, B., Lee, S.-J., Basu, S.: Streaming media caching with transcoding-enabled prox-ies HP Laboratories, Technical Report HPL-2002-210R1 (August 2002)
9. Schulzrinne, H., Lanphier, R., Rao, A.: Real time streaming protocol (RTSP) RFC 2326, IETF (April 1998)
10. Schulzrinne, H., Casner, S., Frederick, R., Jacobson, V.: RTP: a transport protocol for real-time applications RFC 1889, IETF (January 1996)
11. Lee, Y.-J., Min, O.-G., Mun, S.-J., Kim, H.-Y.: Enabling high performance media streaming server on network storage card In: Proc. IASTED IMSA, August 2004, USA (2004)
12. Min, O.-G., Kim, H.-Y., Kwon, T.-G.: A mechanism for improving streaming throughput on the NGIS system In: Proc. IASTED IMSA USA (August 2004)

13. Ahn, B.-S., Sohn, S.-H., Kim, C.-Y., Cha, G.-I., Baek, Y.-C, Jung, S.-I., Kim, M.-J.: Implementation and evaluation of EXT3NS multimedia file system In: Proc ACM international conference on Multimedia, June 2004, New York, NY, USA (2004)
14. Zipf, Law: http://en.wikipedia.org/wiki/Zipf's-law
15. Poisson Distribution: http://mathworld.wolfram.com/PoissonDistribution.html
16. Round-Robin scheduling: http://en.wikipedia.org/wiki/Round-robin_scheduling

Constructing and Application of Multimedia TV News Archives*

H.T. Pao[1], Y.Y. Xu[2], S.C. Chung[2], and H.C. Fu[2]

[1] Department of Management Science,
[2] Department of Computer Science
National Chiao Tung University, Hsinchu, Taiwan, ROC
htpao@cc.nctu.edu.tw,
{yyxu,scchung,hcfu}@csie.nctu.edu.tw

Abstract. This paper addresses an integrated information mining techniques for broadcasting TV-news. The utilizes technique from the fields of acoustic, image, and video analysis, for information on news title, reporters and news background. The goal is to construct a compact yet meaningful abstraction of broadcast TV news, allowing users to browse through large amounts of data in a non-linear fashion with flexibility and efficiency. By using acoustic analysis, a news program can be partitioned into news and commercial clips, with 90% accuracy on a data set of 400 hours TV-news recorded off the air from July 2005 to August of 2006. By applying additional speaker identification and/or image detection techniques, each news stories can be segmented with a better accuracy of 95.92%. On screen captions and screen characters are recognized by video OCR techniques to produce the title of each news stories. Then keywords can be extracted from title to link related news contents on the WWW. In cooperation with facial and scene analysis and recognition techniques, OCR results can provide users with multimodal query on specific news stories.

Keywords: TV-News archives, Multimedia information mining, Multimodal query, Video OCR, Speaker identification.

1 Introduction

Among the major sources of news program, TV has clearly been the dominant influence at least since the 1960s. Yet it is easy to find the old newspaper in microfilm in any public library, but it is impossible to find the old footage of television news in the same library. Thus, Mr. Paul C. Simpson founded the Vanderbilt University Television News archive in 1968. In [1], a team in the University of Missouri-Columbia decided to do a content analysis of the three US network coverage of the 1989 Tiananmen Massacre, they located these news items in the Vanderbilt Archive Index [2]. The Vanderbilt archive promptly

* This research was supported in part by the National Science Council under Grant NSC 94-2213-E009-139.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 151–160, 2007.
© Springer-Verlag Berlin Heidelberg 2007

provided the 11-hour video clips all related to the Tiananmen Massacre. At the same time, the Missourian team also planned to do a comparable study of Taiwanese reportage on Tiananmen Massacre. But the equivalent material of the Vanderbilt archive did not exist in Taiwan then. Therefore, that study only contained the US perspective of the Tiananmen Massacre. In fact, an economical and effective way for the construction as well as utilization of a multimedia TV news archive will be highly desired for countries and places where traditional TV news archives are not available. As described in [3], a multimedia TV News system was implemented to achieve the following goals:

1. Academic and applied aspects: This archive will greatly improve the quality of TV news broadcasting. In an interview with Texas Monthly, Dan Rather, the retired CBS anchorman, once mentioned that he lives with two burdens -*the ratings and the Vanderbilt Television News Archive* [4]. Therefore, once the archive is there, the researchers and the public may do content analysis on the TV news. Then, journalists will be more careful in what they report.
2. Timing factor: Vanderbilt archive started its project with Betacam videotapes in 1968. There will be a problem of preservation because these tapes deteriorate along the years. Today, we can save lots of TV news contents in hard discs, VCDs and/or DVDs.

Infomedia[5] is an integrated project launched in Carnegie Mellon university. Its overall goal is to use modern AI techniques to archive video and film media. VACE-II, a sub-project of Informedia, automatically detects, extracts, and edits highly interested people, patterns, and story evolves and trends in visual content from news video.

This paper proposes an integrated information mining technique that can automatically generate semantic labels from news video[6], as to discover hidden news information. We expect that the following significance will come true.

- Although web-news provides an efficient way to access news, however most of web-news contains only text-based news instead of multimedia contents.
- Since there are many TV-news channels, most people needs an efficient way of searching a favor channel for desired news program.
- Although almost every channel claimed their objectivity on all news items, but truly objectivity is hard to maintain without any outside supervision. Some kinds of monitoring and evaluation are needed to check if a channel is really objectivity.

The rest of this paper is organized as follows. Section 2 introduces overall concepts of the multimedia TV news archive. In Section 3, methods of generating necessary semantic labels from the recording TV news video are presented. Section 4 focuses on describing the information mining on the proposed TV news archive. Finally, summary and concluding remarks are given in Section 5.

2 TV-News Archive

An automated Web-based multimedia TV-News archive [3] needs to contain the following modules: (1) TV news video acquisition, (2) news content analysis, and

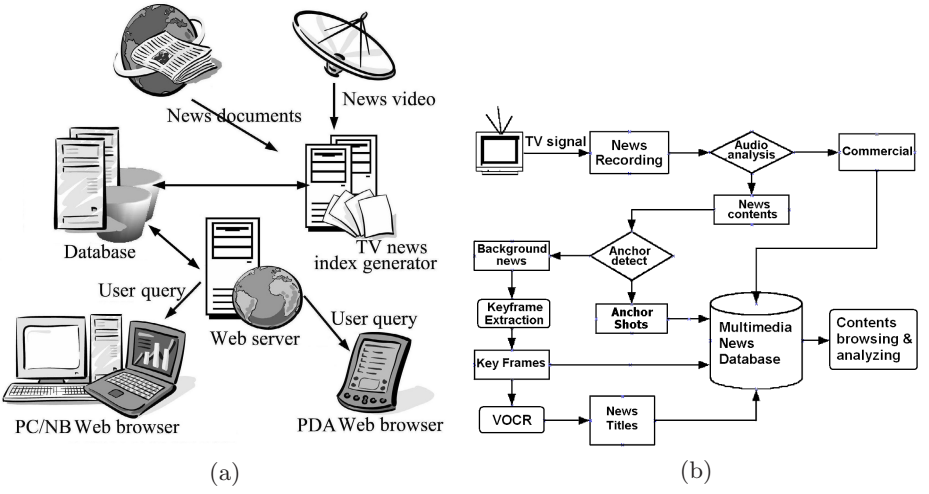


Fig. 1. (a) The overall architecture and information processing flow of the proposed fully automated web-based TV-news system; (b) The flow diagram of TV news acquisition and content segmentation

(3) user interface for news query and retrieval. Figure 1(a) depicts the architecture and interaction of these three modules. Detailed information processing of these three modules is depicted in Figure 1(b). At the beginning, a completed period of TV news program is captured and encoded in stream format[7], and then is stored in a database. In the meantime, a shot detector is used to segment the streaming video into scene based shots for news story segmentation and keyframe extraction[8]. Within a scene shot, speaker identification techniques are then applied to identify anchor frames[9]. By using video OCR techniques[10], close captions of each anchor frames are recognized as candidates for the news title and keywords of each news story. The extracted keywords can then be used to match with (1) Internet news stories to construct links between TV news stories and Internet news, and (2) the users' query words for retrieving interested and/or related news stories. In addition, the extracted text characters can also be used as semantic labels of each news units. Detail description of the semantic labels is presented in Section 3.3.

3 News Information Tree Generation

The basic elements in a news story writing contain 5 W's, that are who, what, when, where and why, which are crucial to catch a reader's attention and to present essential facts of a story. Based on this discipline, the news archive system introduced in Section 2, may be further enhanced to extract more information contents from a recorded news program video. A *news information tree*[11] is suggested to further structure the recorded TV news contents.

3.1 News Information Tree

Figure 2(a) illustrates the hierarchical structure of a news information tree. The hierarchical tree contains five types of news information records: (1) Date, (2) Channel, (3) Title, (4) Content, and (5) Commercial. The title record contains the starting time, length, and brief description of the corresponding news story unit. Thus, the content record may also be further divided into three sub-records: (a) on-site locations, (b)interview contents, and (c) data tables and/or quoted words.

3.2 Content Analysing Units

In general, a complete TV-news program contains following items: news stories, commercials and weather reports. The flow chart of segmenting a TV news program is shown in Figure 2(b). A complete description of TV news shot detection and scene segmentation can be found in [12]. Among various video scene shots, anchor shots are detected first[9]. In general, anchor shots are the most frequently appeared video shots, thus we propose to use a model-selection[9] based method to cluster anchor shots from other video shots. When all the anchor shots are detected, a SVM based video shots classifier[13] is used to detect weather report shots. The following video characteristics and techniques are integrated to achieve a high performance commercial detector: (1) Rates of zero crossing; (2) Short time energy changes and rates; (3) Shot change rates; (4) Video clip lengths. Finally, commercials may be detected and separated from news stories. At this stage, anchor's briefing, weather reports, commercials, and background stories are all separated and identified from each other. As shown in Figure 3, on-site scene shots (background stories) are further segmented and

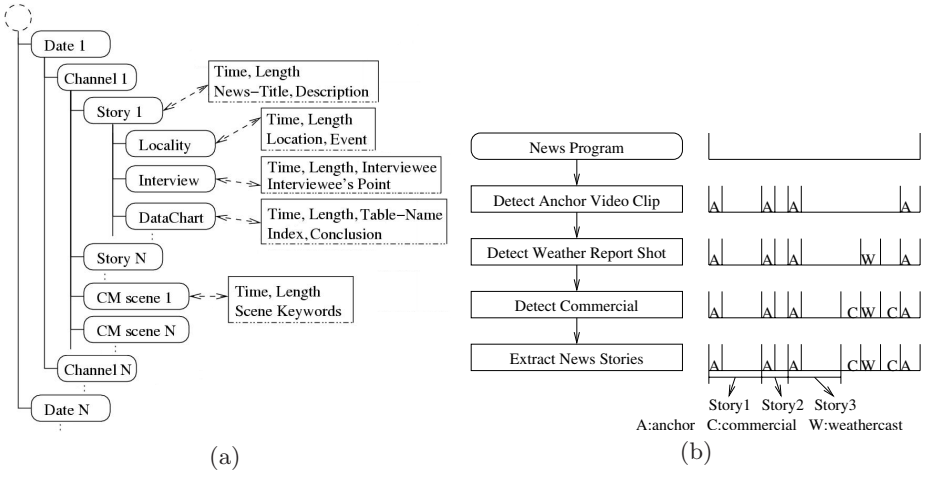


Fig. 2. (a)The data structure of a news information tree; (b) The flow diagram of the proposed news story analysis and information extraction processes

classified into three categories: scene locations, interview contents, and date tables and/or quoted texts. In general, the location scene comes with reporters' narration. Most of the time, the reporters' on-site narration scene is not as active as the interview scene, the interview scene can be distinguished from location scene. Based on location or street names, location scenes can be distinguished from data tables and/r quoted words scenes.

3.3 Semantic Labels of News Story Units

This section describes how to assign each segmented unit with semantic labels. First, the text words for semantic labels are extracted from close-caption or on-screen characters. Usually, a TV-news story scene often contains a stream of words, such as locations and peoples' names, and keywords of news events, ... etc., as an overview of each news story at on-screen caption area. In general, these keywords contain sufficient enough information for labeling each segmented units. Figure 4 shows a block diagram of how a news information tree can be

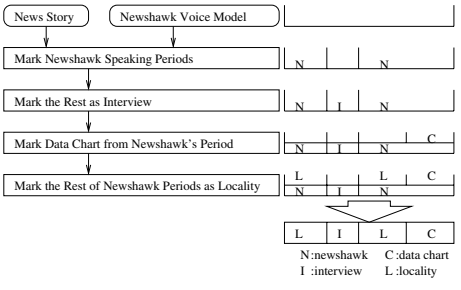


Fig. 3. The flow diagram of on-site scene contents segmentation. First, on-site reporters' narration and interview contents can be detected by using speaker identification techniques to distinguish the interview contents from the rest of scenes. By detecting the screen characters, scenes associated data tables and quoted words can be detected and distinguished. Then, rest of scenes can be identified as background scenes.

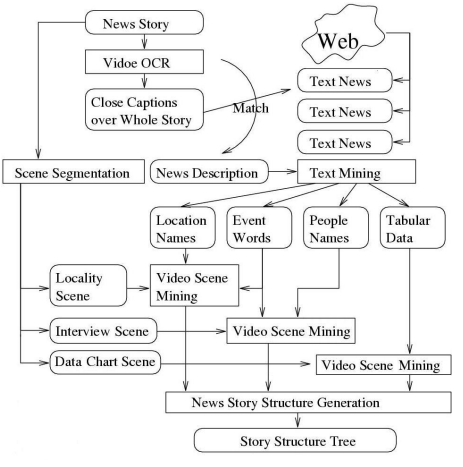


Fig. 4. The news information processing for the generation of a news information tree

established. The establishing of news information tree contains two phases: story and scene processing phases.

In story processing phase, video OCR (optical character recognition) techniques are applied to extract text characters in the close-caption area. The extracted characters will then be used query keywords to retrieve news documents over the Internet news sites. The title and contents of best matched web news

Fig. 5. (a) An example of locality scene frame. The locality scene is used to show where and what the news occurred. Thus, the location information and event description can be retrieved from the close-captions of a locality scene; (b) An example of interview scene frame. The peoples and reporters' names and the keywords of their opinions can be extracted from screen characters or closed captions.

documents can be used as the labels and text contents of the corresponding news story in the news information tree. In addition, more key words, such as associated people's, event locations', and reporters' names etc., can be extracted to form a label candidate pool for scene processing phase. In scene processing phase, the major work is to select proper semantic labels from label candidate pool for each scene and/or video units. As shown in Figure 5(a), the on-screen captions of locality scene contain location name and some event descriptions. Therefore, the location names and event associated keywords are selected from a label candidate pool for locality scenes.

Figure 5(b) is an example for interview scene. In the interview scene, people's names and their view points are given in on-screen captions. Therefore, in an interview scene, we can search for people names, and event words from the label candidate pool. An example of data chart scene is shown in Figure 6(a). As we can see that text items in data chart usually can be semantic labels or sub-labels for data items in a column or a row of a table in a data chart scene.

4 Data Mining on the News Information Tree

This section presents how and what to mine from a news information tree (NIT). In Section 4.1, we propose to mine the favored or preferred news contents of a TV-station. The news information tree can also be used to track the evolution of a news story (see Section 4.2). In addition, the mining results from the NIT and the realtime ratings can be combined to provide a very useful guidance to TV-news commercial buyers.

4.1 Mine the News Preference of a TV Station

Generally speaking, a TV-station arranges the broadcasting sequence of each story in a news program according to their impact and attractiveness to audience. Thus, a preferred news story may be assigned more air time. By analyzing the sequence order and the length of news stories, the preferred or the favored news stories and/or news types of a TV-station can be roughly figured out. Knowing favored or preferred types of news story of a TV station will help audience to select favor news channels or TV stations.

The proposed news mining method is described as follows: Given N sets of keywords, $K_1, K_2, ..., K_i, ..., K_N$, which correspond to N news topics (or subjects), let the following delta function $\delta(k, K_i)$ define the relations between a keyword k and a keyword set K_i:

$$\delta(k, K_i) = \begin{cases} 1, \text{ if keyword } k \in K_i \\ 0, \text{ otherwise.} \end{cases}$$

1. Extract keywords $\{k^l_{s_j};\ l = 1, \cdots, L_j\}$ from a scene unit s_j in a news program.
2. For each scene units s_j, compute its association frequency $F(K_i|s_j)$ with respect to a subject K_i,

$$F(K_i|s_j) = \frac{\sum_{l=1}^{L_j} \delta(k^l_{s_j}, K_i)}{\sum_{i=1}^{N} \sum_{l=1}^{L_j} \delta(k^l_{s_j}, K_i)}$$

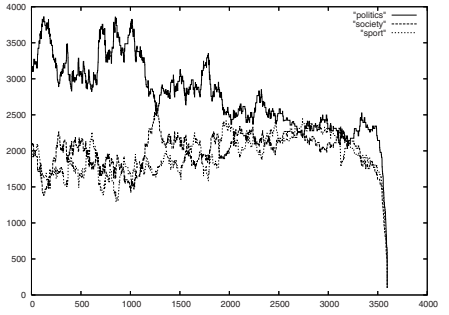
3. Compute the the association frequency of news program $F_d(t|K_i)$ at hour t of a day d:

$$F_d(t, K_i) = \begin{cases} F(K_i, s_1), & \text{for } s_1.start \le t \le s_1.end \\ F(K_i, s_2), & \text{for } s_2.start \le t \le s_2.end \\ \quad \vdots & \quad \vdots \\ F(K_i, s_M), & \text{for } s_M.start \le t \le s_M.end, \end{cases}$$

where $s_j.start$ and $s_j.end$ are the start and the end time of a scene unit s_j in a news program at hour t of a day d.

The associated frequency distribution from a segment of a news program is not enough to represent the overall preference or trend of a news channel, thus long term statistics is desired. By accumulating a longer period (say one month) of associated frequency of news subjects, the preference of a channel can be discovered. As shown in Figure 6(b), appearing frequency keywords that are related to social, political, and entertainment news are applied to associate with and to accumulate frequency of news topics. As we can see in this example, the monitored TV news channels or stations' favor social and political news more than entertainment news.

(a)



(b)

Fig. 6. (a) An example of data chart scene frame. The data chart scene or frame can provide numerical and statistical information in a organized manner; (b) Three sets of (representative) keywords are used to associate the appearing frequency of *social*, *political* and *sport* news in a news program. The scale on **X**-axis is in second, and the appearing frequence is shown on **Y**-axis.

4.2 The Evolution of a News Event

The evolution of a news story can also be mined from the news information tree. By associating the keywords of a specific event with recorded news scenes over a period of days, then the accumulated association frequency of matched scene units profiles a life cycle of the specific news event. Figure 7 shows the life-cycle of a particular news event. In addition, the spreading of the specific events to other places, e.g., cities, counties, countries, etc., can also be retrieved from the associated names of locations in the matched scene units. For example, one can query a news story by using a particular people's name, then the person's daily schedule and/or whereabouts can be retrieved from the the recorded NIT.

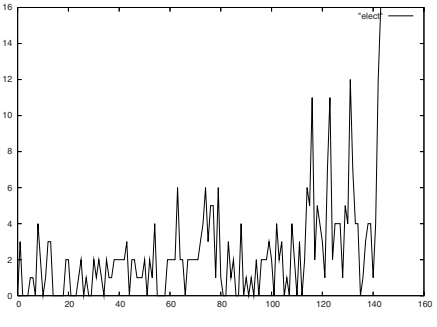


Fig. 7. The life-cycle of a specific news events along with a period of time in days along **X**-axis

4.3 The Mining on TV Commercials

In addition to background stories, commercials are also valuable information. Huang et al., [12] proposed commercial detecting and identifying methods in TV video clips. When a commercial frame contains screen characters and/or, video OCR techniques can be used to extract keywords to label the corresponding commercial clips. Otherwise, keyblock-based image retrieval methods [14] may be utilized to represent and to identify each commercial clips. However, manual annotation is needed to label the keyblock. By gathering statistical information of these labels and keywords in news programs, cross relationship between TV commercials, realtime ratings, and news stories can be discovered and analyzed to achieve a useful marketing database. Two example areas, *customer modeling* and *cross-selling*, in database marketing are discussed in the followings.

Customer modeling. The basic idea behind customer (i.e., the commercial buyers and news audience) modeling is to improve audience response rates by targeting prospects that are predicted as most likely to respond to a particular advertisement or promotion. This is achieved by building a model to predict the likelihood that groups of news audience will respond based on news type, viewing time and news channels as well as previous viewing behavior. In addition, by targeting more effectively to prospects and existing commercial buyers, TV station operators can improve and strengthen customer relationships. The customer can perceives more value in TV news and commercials (i.e., both commercial buyers and news audience receive only products and/or services of interest to them).

Cross selling. The basic idea behind cross selling is to leverage the existing customer base by selling them additional products (commercial time slots) and/or news services. By analyzing the groups of products or services that are commonly purchased together and predicting each customer's affinity towards different products using historical data, a TV-station can maximize its selling potential to the existing customers. Cross selling is one of the important areas in database marketing where predictive data mining techniques can be successfully applied. Using historical purchase data of different commercials from the customer database along with news type, viewing time and news channels, commercial buyers can identify their products that are most likely to be of interest to targeted news audience. Similarly, for each type of product (i.e., commercial or groups of commercials), a ranked list of different types of news or groups of audience, that are most likely to be attracted to that product. Then, arrangement of commercials with matched types of news to achieve a high likelihood of audience response rate.

5 Conclusion

This paper addresses techniques and possible applications of automated information mining on a multimedia TV-news archive. The proposed automated information mining techniques contain the followings: (1) segmenting a TV-news

video recording into story based units, (2) using video OCR techniques to extract and to recognize close-caption and/or screen characters into keywords for each scenes, (3) using extracted keywords to generating semantic labels for each scenes, and (4) segmenting commercial video clips from news units. Information associated with various labels and scenes (e.g., the starting and ending time of a scene) are stored in a *news information tree*. Performing statistical analysis on the data items in the news information tree, one can reveal hidden information, like popular channels and profile of some hot news stories. These information can help general multitude in finding their favored or desired news-channel, searching focal point person, tracking hot news stories, matching more favorable commercials, ... , and so on.

References

1. Huffman, S., Yang, T.E., Yan, L., Sanders, K.: Genie out of the bottle: Three u.s. networks report tiananmen square. In: Proceedings of the annual meeting of Association for Education in Journalism and Mass Communication, Minneapolis, Minnesota, USA (1990)
2. Vanderbilt television news archive, `http://www.vanderbilt.edu/vtna`
3. Lai, P., Lai, L., Tseng, T., Chen, Y., Fu, H.C.: A fully automated web-based tv-news system. In: Proceedings of PCM2004, Tokyo, Japan (2004)
4. Dan rather interview with texas monthly, `http://tvnews.vanderbilt.edu/about.pl`
5. Informedia. `http://www.informedia.cs.cmu.edu/`
6. Daniel, G., Daniel, J.: Automatic labeling of semantic roles. Comput. Linguist. 28, 245–288 (2002)
7. Wang, Y., Ostermann, J., Zhang, Y.Q.: Video processing and communications. Prentice Hall Press, Englewood Cliffs (2002)
8. Patel, N.V., Sethi, I.K.: Video shot detection and characterization for video databases. Pattern Recognition 30, 583–592 (1997)
9. Cheng, S.S., Wang, H.M., Fu, H.C.: A model-selection-based self-splitting gaussian mixture learning with application to speaker identification. EURASIP Journal on Applied Signal Processing 17, 2626–2639 (2004)
10. Lin, C.J., Liu, C.C., Chen, H.H.: A simple method for chinese video ocr and its application to question answering. International Journal of Computational Linguistics and Chinese Language Processing 6, 11–30 (2001)
11. Feinstein, C., Morris, P.: Information tree: a model of information flow in complex organizations. Systems, Man and Cybernetics, IEEE Transactions 18, 390–401 (1988)
12. Huang, T.Y., Lai, P.S., Fu, H.C.: A shot-based video clip search method. In: Proceedings. of CVGIP2004, Taipei, Hualien, ROC (2004)
13. Sun, S.Y., Tseng, C.L., Chen, Y.H., Chuang, S.C., Fu, H.C.: Cluster-based support vector machine in text-independent speaker identification. In: Proceedings of International Joint Conference on Neural Networks IJCNN 2004, Budapest, Hungary (2004)
14. Zhu, L., Rao, A., Zhang, A.: Theory of keyblock-based image rerieval. ACM Trans. on Information Systems 20, 224–257 (2002)

A Three-Level Scheme for Real-Time Ball Tracking

Xiaofeng Tong[1], Tao Wang[1], Wenlong Li[1], Yimin Zhang[1]
Bo Yang[2], Fei Wang[2], Lifeng Sun[2], and Shiqiang Yang[2]

[1] Intel China Research Center, Beijing, P.R. China, 100080
{xiaofeng.tong,tao.wang,wenlong.li,yimin.zhang}@intel.com
[2] Department of Computer Science and Technology,
Tsinghua University, Beijing 100084
bo.yang02@gmail.com, fei-wang04@mails.tsinghua.edu.cn,
{sunlf,yangshq}@tsinghua.edu.cn

Abstract. A three-level method is proposed to achieve robust and real-time ball tracking in soccer videos. It includes object-, intra-trajectory-, and inter-trajectory-level processing. Due to much noise and frequent occlusion, it's difficult to get the solely ball in one frame. Thus, in object level, multiple objects instead of a single one are detected and taken as ball candidates with shape and color features. Then at intra-trajectory level, each ball candidate is tracked by a Kalman filter in successive frames, which results in lots of initial trajectories in a video segment. These trajectories are thereafter scored and filtered according to their length and relationship in a time-line model. With these trajectories, we construct a distance graph, in which a node represents a trajectory, and an edge means distance between two trajectories. We use the Dijkstra algorithm to get the optimal path in the graph at the inter-trajectory level. To smooth the trajectory, we finally apply cubic spline interpolation to bridge the gap between adjacent trajectories. The algorithm is tested on broadcast soccer games in FIFA2006 and got the *F*-score 80.26%. The whole speed far exceeds real-time, 35.6 fps on mpeg2 data.

Keywords: Ball tracking, Kalman filtering, trajectory inference, data association, sports video.

1 Introduction

Semantic object (ball, player, goalmouth, face, jersey number, etc) detection and tracking have attracted much attention and been used for advanced semantics analysis in addition to highlights extraction in sports video [8]. In field-ball sports, the ball is the focus of competition. Ball detection and tracking, trajectory analysis is of great benefit to semantics mining, such as tactics, team activity analysis, etc [4].

In previous literatures, X. Yu [8] proposed a detection-based tracking method for ball trajectory selection, inference and extension, and applied it in semantic event detection. D. Liang [2] used Kalman filtering to track the ball based on detection verification in initial multiple successive frames. X. Tong [5] presented a Condensation based ball tracking algorithm using color and shape features. F. Yan [7] used trajectory level verification scheme to infer and determine the ball trajectory in tennis

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 161 – 171, 2007.
© Springer-Verlag Berlin Heidelberg 2007

video based on data association. At early times, chromatic and morphological features were utilized to detect ball in [3].

There are some challenges for soccer ball tracking. Firstly, there is camera motion in broadcast video; thus, the estimated ball motion is mixed by intrinsic ball and camera motion. Secondly, it is not easy to construct an effective model to detect ball because there are constantly size and shape change, occlusion and noise (See Fig 1). Thirdly, there is much occlusion, frequent motion speed and direction change, flying outside of playfield, which challenges the robustness of ball tracking in a long time.

Fig. 1. Difficulties of ball detection (from left to right: shape blur, occlusion and noise)

The above challenges come into being much difference between ball and other traditional object tracking algorithms. Firstly, it needs more than one object for tracking initialization. The unreliable detection and much noise make tracking can not simply rely on detection. Secondly, long time tracking without instant rectification is not stable due to frequent occlusion and shot transition. We deem the tracking as data association based on detection is feasible and practical.

In this paper, we propose a three-level including object-, intra-trajectory-, and inter-trajectory-level, scheme to realize effective ball tracking solution. The idea is to discriminate and optimize the unique and true ball trajectory from multiple trajectories generated by many ball-like candidates. At the object level, we use the video dominant color (field color in soccer) to extract non-field objects including ball, players, line-marks and others within play-field. Then a filter with shape and color cues is applied to remove noises, and several qualified candidates are reserved in each frame. At the intra-trajectory level, many possible trajectories are generated through motion fitting with ball candidates in consecutive frames. Some false alarms are removed through trajectory filtering in term of the length and their relationship in time-line. At this phase, there are still many trajectory candidates in a video segment. To separate the truth trajectories from others, at the inter-trajectory level, a graph, in which a node represents a trajectory, and an edge is the distance between a pair of trajectories, is constructed to find the optimal path by the Disjkstra algorithm. In order to smooth and linkup the final ball trajectory, a cubic spline is used to interpolate the gap between two adjacent trajectories.

Compared with previous methods, the contributions of this paper are:

1) Propose a practical and hierarchical scheme for real-time tracking.
2) Apply time-line model to define trajectory distance.
3) Use Dijkstra distance matrix to optimize the optimal ball path.
4) Utilize cubic spline interpolation to predict the ball position in missing frames.

We tested the method on FIFA2006 soccer games and got the *F*-score of 80.26%. The speed is 35.6 fps on mpeg2 data (720x576 pixels, 25fps, Intel Core 2 Dual).

2 Framework

The whole flowchart is illuminated in Fig 2, which consists of: (1) pre-processing for view-type classification. In this step, only global views are selected for ball tracking process. (2) Object level processing, where multiple ball candidates are identified through image segmentation and connect-component analysis. To eliminate noises, Hough line detection, shape and color cues verification are performed. With the filtering, the left candidates have high probability to be the ball. (3) Intra-trajectory level,

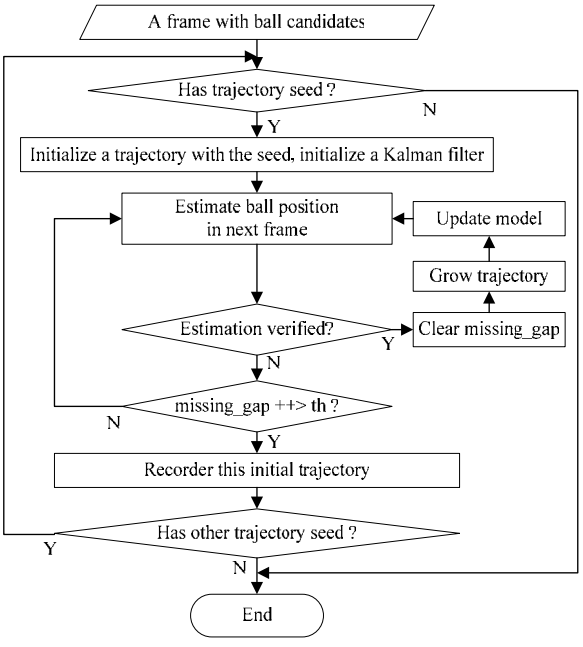


Fig. 2. Framework of the method

connect ball candidates and generate initial ball trajectories, delete the false ones. (4) Inter-trajectory processing, define distance between trajectories, construct a distance graph representing the trajectories and their distance, and find the optimal path consisting of non-overlap trajectories. Finally, we use cubic spline interpolation to smooth and linkup the gap between two adjacent trajectories.

3 Methodology

3.1 Pre-processing

The pre-processing module performs view-type classification, selects global views for ball detection and tracking. We firstly learn the dominant color (corresponds to grass

color) by accumulating HSV color histograms with lots of frames. Then we segment the play-field with the dominant color. According to the area of play-field and the size of non-field object, we classify each view using a decision tree into one of four types of views: global view, medium view, close-up, and out of view (Fig 3) [1].

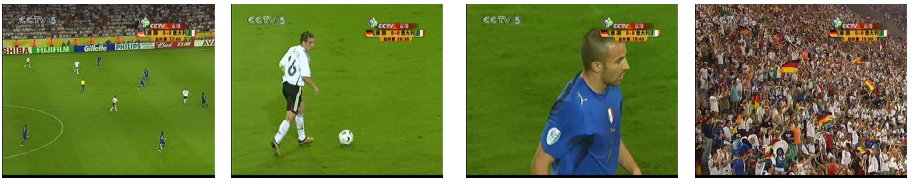
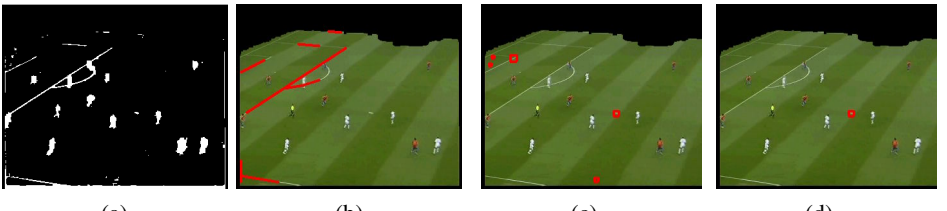


Fig. 3. View-types: global view, middle view, close-up, and out-of view

3.2 Object Level Processing

In a global view, we extract the non-field regions using dominant color segmentation (Fig.4(a)). The ball is included in these regions (white regions). But there is much noise: line-marks, player regions, false segmented regions, etc. To discriminate the ball and remove noise, we apply a three steps filtering:

1) We use Hough transform to detect direct lines within play-field and remove them.
2) We filter out the unqualified regions with shape features: (a) size, the size of the longer axis of a region; (b) area (A), area of the region; (c) eccentricity, ratio of longer axis to shorter axis; and (d) circularity factor, represented as $4*\pi*A/P^2$, P is the perimeter.
3) We utilize an SVM classifier with color cue to further verify the ball candidates. A four-dimension color feature vector is used: $<r, b, s, v>$, in which $r = R/G$, $b = B/R$ in RGB color space, s and v are saturation and value components in HSV space. The middle results are presented in Fig.4.



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

Fig. 4. Results in object level processing. (a) Segmented binary image by dominant color; (b) remove line-marks; (c) filtering with shape feature; and (d) filtering with color.

After get ball candidates in each frame, we divide the video into shots (called ball-shot), then perform tracking process in each ball-shot. If the frame number interval between two adjacent ball frames is larger than a threshold ($interval > th_delta$), a

ball-shot boundary at this time-point is declared. In this work, the *th_delta* is set to be 2**fps* (frame per second of the video).

3.3 Intra-trajectory Level Processing

Once get the ball candidates in each frame, we generate initial trajectories through linking the adjacent candidates in spatial-temporal domain. The predication is performed by Kalman filtering with the formation:

$$X_k = AX_{k-1} + w_k$$
$$Z_k = HX_k + v_k$$

Where X_k and Z_k are the state and measurement vector at time k, w_k and v_k are system and measurement noise. A and H are state transition and measurement matrix, respectively. In this work, we set:

$$X = \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}, Z = \begin{bmatrix} x \\ y \end{bmatrix}, A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

where (x, y) is the ball's center, (v_x, v_y) is the velocity in x- and y-direction.

A trajectory is initialized by a seed candidate that is not included by any trajectory, and grows if the position predicted by Kalman filter is verified by a candidate in the next adjacent frame. The procedure is summarized in Fig 5, which is similar to [8].

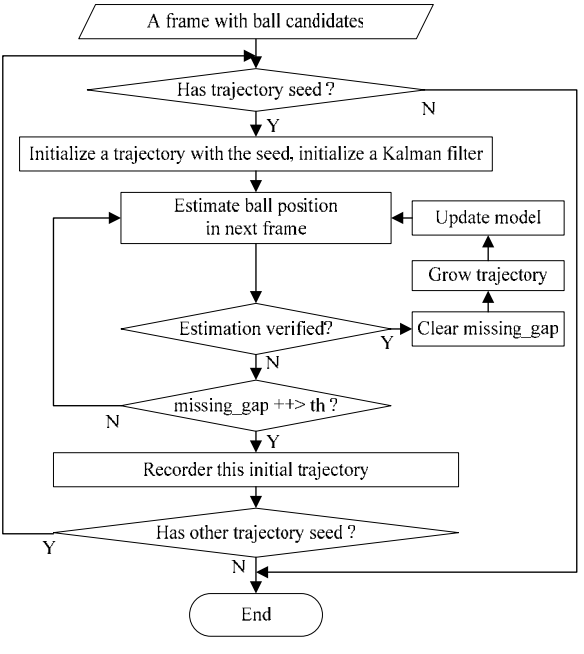


Fig. 5. Flowchart of initial trajectory generation

The initial trajectories link all possible candidates along a reasonable path. We set confidence for each trajectory according to its length (during):

$$confidence = \begin{cases} false, & if \quad length < T_1 \\ true, & if \quad length > T_2 \\ pendent, & \text{otherwise} \end{cases}$$

We then remove the unconfident trajectories. If a pendent trajectory is covered (A covered by B means: A.start > B.start, and A.end < B.end) by another one (see Table 1), it is also deleted. The others reserved will be identified at inter-trajectory level. The procedure is shown in Fig 5. The ball candidates are shown in Fig 6(a). The initial trajectories are displayed in Fig 6(b). Fig 6(c) shows the procedure of trajectories filtering. The circled trajectories are false. Fig 6(d) is the filtering result.

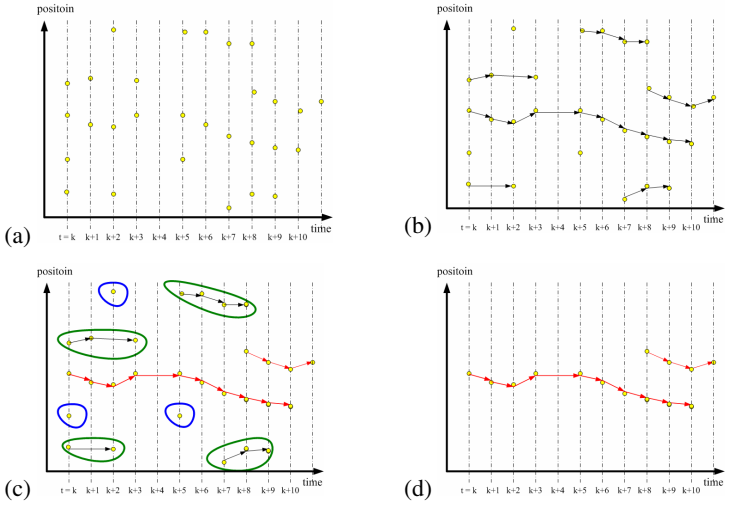


Fig. 6. Intra-trajectory filtering

3.4 Inter-trajectory Level Processing

We further discriminate the truth trajectories via path optimization at the inter-trajectory level. The ball's trajectory within a shot should be smoothing and continuous. We define the distance between two trajectories and generate a distance graph. Then, we find the optimal path with Dijkstra algorithm.

We apply the time-line model to define the distance between two trajectories. Let A be a trajectory, A.start and A.end be its start and end time. The symbol "<" represents "earlier", ">" denotes "later", "<=" is "no later", ">=" is "no earlier". "a" and "b" are ball candidates in trajectory A and B. We define the distance of two trajectories below.

Table 1. Trajectory distance definition between trajectories A and B

No	Description	Example	Distance
1	A.end < B.start	A B	dist(A.end, B.start)
2	A.end > B.start and A.end <= B.end	A B	min(dist(a,b)), a,b ,
		A B	
3	A.end > B.end	A B	∞
		A B	
		A B	

Assume $a = <x_a, y_a, t_a>$, $b = <x_b, y_b, t_b>$ be two ball points, then

$$dist(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} + \alpha |t_a - t_b|$$

Where α is a weighted factor for time interval. In this work, we set α = 7.

In the trajectory graph, a trajectory is taken as a node, and the linkage between trajectories are edges. The distance between two trajectories is set to be edge weight. In the graph, we put two additional start and end nodes. The start node is a trajectory including one ball-frame which has a ball candidate with the start time of the ball-shot. The distance from the start node to its neighbors is represented by the temporal distance (the spatial distance is zero). The end node is defined vice verse. With the graph, we apply Dijkstra algorithm to find the optimal path from start to end. The resultant path consists of optimal time-ordered trajectories. Fig 7 portrays an example. Trajectories before/after optimization are shown in (a) and (b) (blue solid line).

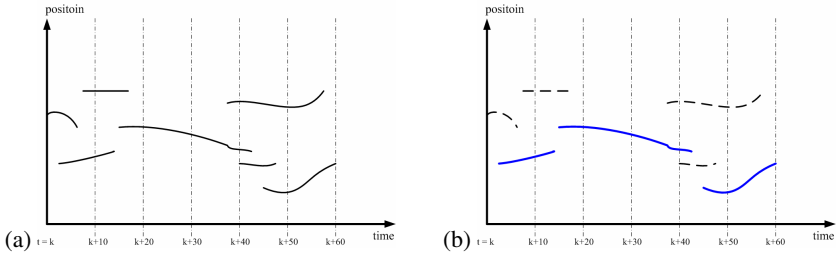


Fig. 7. Inter-trajectory optimization. (left) multiple trajectory candidates, (right) filtered trajectories(blue and solid lines).

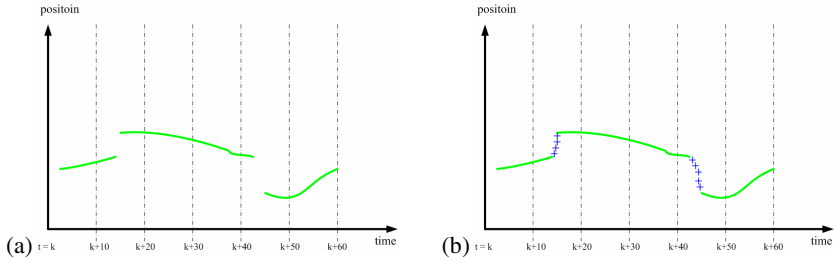


Fig. 8. Trajectory interpolation. (a) trajectories before interpolation, (b) interpolation result (points with blue "+" are interpolated results).

The interval between two adjacent trajectories is generated due to candidate missing. The missing often occurs when there is a change of the motion speed and direction (motion model) due to, for example, take a pass, dribble and then pass the ball to others. The interval (gap) can be interpolated by the two sandwiched trajectories. We apply cubic spline interpolation to make up the interval (Fig 8).

4 Experiments and Analysis

4.1 Data-Set and Evaluation Criterion

We tested this algorithm on FIFA2006 broadcasting soccer videos with mpeg2 format. We manually labeled one minutes video (1497 frame). We manually located the ball's position and its radius in every frame if: 1) the ball is visible or 2) it's occluded but its position can be estimated by neighbor frames. The evaluation criterion is based on overlap between ground truth and tracked regions.

4.2 Middle Results

In the experiments, we extract the middle results of the first ball-shot and show in Fig 9. The detected/tracked results are displayed with blue "+"; and true results are showed by red "o". The horizontal axis is time, and the vertical axis is the width of image frame. The ball candidates are shown in Fig 9(a). Trajectories linked by these candidates and filtered results are displayed in Fig 9(b). We can see that the most false trajectories are removed out. Fig 9(c) shows the optimal path by the Dijkstra algorithm. Fig 9(d) is the linking and smoothing result. It can be seen that most correct candidates are reserved, and most missing cases are smoothingly interpolated.

The smoothing and interpolation results from frame 50 to 89 are shown in Fig 10. The ground truth is displayed by red dash line, the tracking and interpolated result is displayed by blue filled circles. Actually, some ground truth is got by estimation and impossible to be detected due to occlusion or blur.

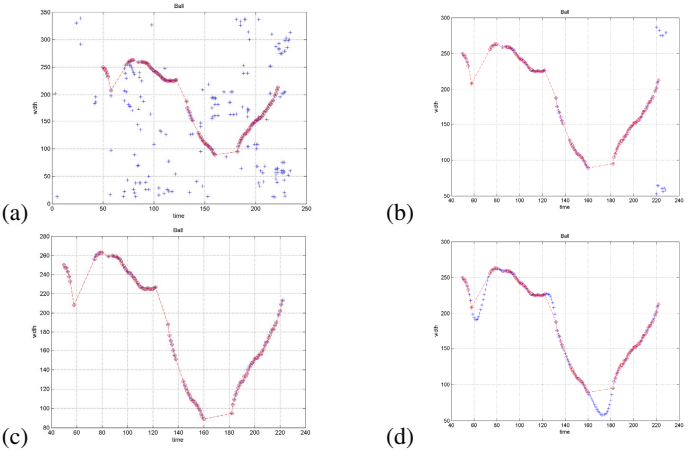


Fig. 9. A ball tracking result. (a) Ball candidates; (b) filtered trajectories; (c) optimal path; (d) final trajectories. Detection/tracking results are displayed by blue "+",truth is shown by red "o".

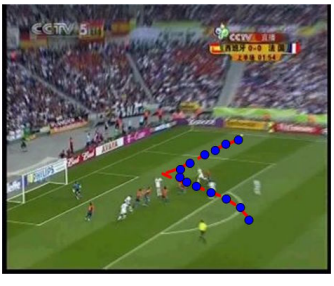


Fig. 10. An interpolation result from frame 50-89

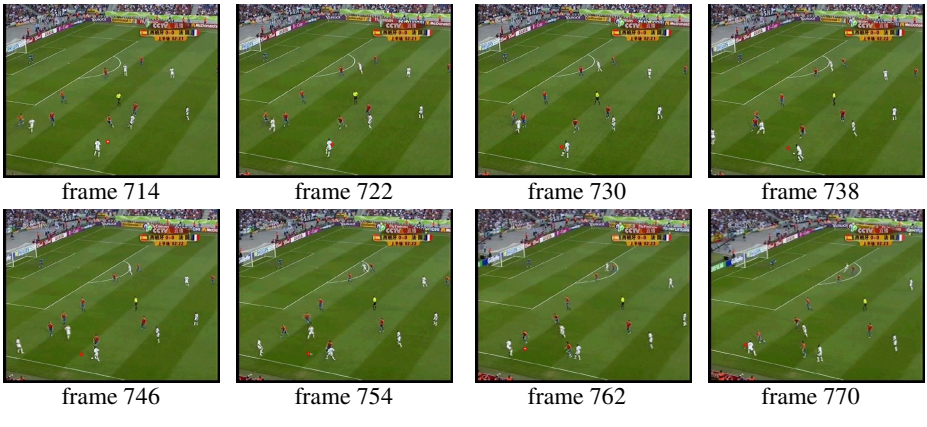


Fig. 11. An example of interpolation in a sequence

Another example of sequence is shown in Fig 11. We extract 8 representative frames from successive 80 frames. In these results, most frames in two occluded segments frame 714-738, and 762-770 are corrected interpolated.

4.3 Tracking Results

We compared the tracking results with the ground truth frame by frame on the one-minute video. The information is given in Table 2. In the total 1497 frames, there are 3 ball-shots, 928 global frames, in which the ball tracking is conducted. The shot boundaries are correctly obtained. In detection phase, the ball is correctly located in 356 frames. Through tracking, we finally get 572 ball-frames.

Table 2. Information of the one-minute video clip

total (frame)	global-view	ball-shot	detected ball-frame	tracked ball-frame
1497	928	3	356	572

In the results, there are correct, false and missing instances. The incremental correctly tracked frames are got by interpolation. It is actually impossible to detect the ball in these frames due to occlusion or clutter. The result is shown in Table 3.

Table 3. Tracking performance

Correct	False	Missing	Precision	Recall	F-score
494	78	165	86.36%	74.96%	80.26%

In the final result, we found that some fragmental trajectories (interval and short length trajectories) are usually missing. It is because that the short trajectories are removed during the intra-trajectory processing.

We also tested the method on a complete game between Spain and France in FIFA2006. The result is also satisfied. Most ball-frames are correctly tracked. The processing including all modules (view-type classification, line-marks detection with Hough transform, three-level ball tracking) is beyond real-time with two threads on Intel Core 2 Dual machine with 2.4G CPU, 2.0G RAM. The testing data is mpeg2 format (720x576 pixels, 25 fps). The operation speed is given in Table 4.

Table 4. Processing speed

Module	Time (second)	Speed (fps)
only view-type	24.48	61.15 (1497/24.48)
only ball tracking	17.50	53.03 (928/17.50)
view-type + ball tracking	41.98	35.66 (1497/41.98)

5 Conclusions

We proposed a three-level scheme for ball tracking in broadcasting soccer video in this paper. To solve the difficulty of solely ball determination, we detect multiple candidates in a frame in the object level. Then, we link these candidates to generate initial trajectories and apply intra-trajectory processing to filter out uncertain trajectories. To get the final optimal path in a shot, we utilize the inter-trajectory procedure to optimize the ball path with the Dijkstra algorithm. We tested the algorithm on real broadcasting soccer video and got satisfying performance. The whole processing far exceeds real-time. In the future, the verification of path composed by small fragment trajectories should be improved.

References

1. Li, J., Wang, T., Hu, W., Zhang, Y.: Soccer highlight detection using two-dependent Bayesian network, In: ICME 2006
2. Liang, D., Liu, Y., Huang, Q., Gao, W.: A scheme for ball detection and tracking in broadcast soccer video. In: Ho, Y.-S., Kim, H.J. (eds.) PCM 2005. LNCS, vol. 3768, Springer, Heidelberg (2005)
3. Gong, Y., Sin, L., Chuan, C., Zhang, H., sakauchi, M.: Automatic parsing of TV soccer programs. In: Proc. of 2nd International Conference on Multimedia Computing and Systems, pp. 167–174 (1995)
4. Sullivan, J., Carlsson, S.: Tracking and labeling of interesting multiple targets. In: ECCV 2006 (2006)
5. Tong, X., Lu, H., Liu, Q.: An effective and fast soccer ball detection and tracking method", ICPR 2004, pp. 795–798 (2004)
6. Welch, G., Bishop, G.: An introduction to the Kalman filter, Technical Report TR95-041, University of North Carolina at Chapel Hill (1995)
7. Yan, F., Kostin, A., Christmas, W., Kittler, J.: A novel data association algorithm for object tracking in clutter with application to tennis video analysis, In: CVPR 2006 (2006)
8. Yu, X., Xu, C., Tian, Q., Leong, H.: A ball tracking framework for broadcast soccer video, In: ICME 2003 (2003)

Video Object Mining with Local Region Tracking

Arasanathan Anjulan and Nishan Canagarajah

Department of Electrical and Electronic Engineering
University of Bristol, Bristol, UK.
{A.Anjulan,Nishan.Ganagarajah}@bristal.ac.uk

Abstract. This paper describes a novel object mining system for videos. An algorithm published in a previous paper by the authors is used to segment the video into shots and extract stable tracks from them. A grouping technique is introduced to combine these stable tracks into meaningful object clusters. These clusters are used in mining similar objects. Compared to other object mining systems, our approach mines more instances of similar objects in different shots. The proposed framework is applied to a full length feature film and improved results are shown.

1 Introduction

Data mining is the process of analyzing data from different perspectives and finding useful patterns or information which can be used in a variety of applications. Text data mining is popular among researchers and have a large application domain varying from business databases to web applications [2, 3]. However despite the vast amount of visual information present in digital media, the research done in visual data mining is limited. There are several reasons for this. The semantic gap between the low level visual features (such as colour, texture and motion) and a higher level user domain (such as object and event identification) is one of the main hurdles for the development of a visual data mining system. Moreover, there is no meaningful clustering or segmentation method that can be universally applied to all kind of visual media. In text, words can be easily clustered into groups. However this is difficult for images and videos. In visual media an object may appear in different imaging conditions (different camera angles, zoom positions, or lighting conditions) and can also be occluded. All these variations make visual data mining more challenging compared to text data mining.

Existing work in video data mining can be divided into two categories: mining similar motion patterns and mining similar objects. The first type of systems use motion information to mine similar event patterns or identify peculiar events [4, 5, 6]. The second category systems aim to group frequently appearing objects in

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 172–183, 2007.
© Springer-Verlag Berlin Heidelberg 2007

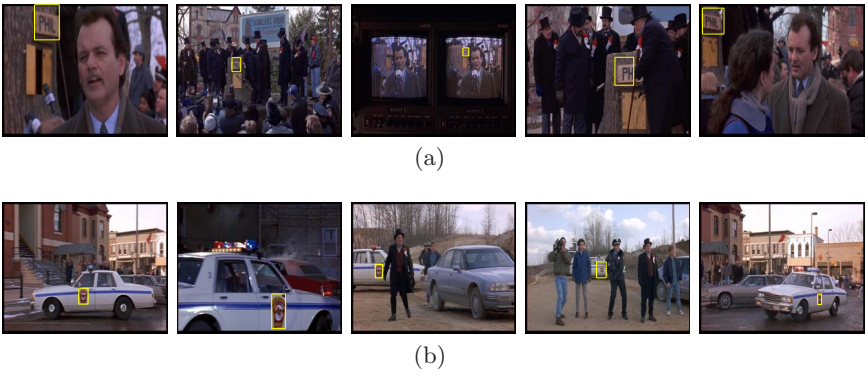


Fig. 1. An example scenario of object appearance in videos. Each row shows an object taken from different shots of the movie *Groundhog Day*. We define these different appearances of the same object among the shots as different instances of the same object. The purpose of the object mining system is to group these different instances of the same object which appear in different parts of the video.

videos. An example scenario of object appearance in videos is given in Fig 1. Each row shows an object taken from different shots of the movie *Groundhog Day*. We define these different appearances of the same object among the shots as different instances of the same object. The purpose of the object mining system is to group these different instances of the same object which appear in different parts of the video. The problem is difficult because the object can appear in different ways in different parts of the video due to different imaging conditions, lighting conditions, back grounds and occlusions.

In [7, 8], authors use vector quantize and spatial neighborhood techniques to mine frequently appearing objects in video. Their system is based on key frames and any interesting objects appearing outside the key frame will be missed and any failure in key frame extraction will affect the mining system. Moreover different spatial neighborhood sizes needed to mine different objects.

This paper use an earlier work published [1] by the authors to segment the video into shots[1] and extract features from stable local invariant tracks. These features are grouped into clusters to represent relevant objects. Each cluster approximately corresponds to an object in a shot. These object clusters may contain a number of similar instances of the same object, and these instances are grouped together in the mining stage. The effectiveness of the system is demonstrated by performing object mining in a full length feature film. The rest of this paper is organized as follows. The proposed mining frame work is described in section 2 and results from a real movie are presented in section 3. We conclude in section 4 with suggestions for future work.

[1] A shot is the basic element in a video and can be defined as an uninterrupted sequence of frames recorded from a single camera operation.

2 Proposed Framework

2.1 Shot Segmentation and Representation

The first stage of the system divide the video into a number of meaningful shots, and obtain representation for objects that are present in the shots. Both of these tasks are performed simultaneously using the local invariant features obtained at each frame. An algorithm previously published by the authors [1] are used for this purpose. In this paper, MSER [11] is used to extract local invariant regions (LIRs) and an ellipse fitting algorithm is used to approximately fit an ellipse to each of these regions. SIFT [12] is used to assign a 128 dimensional descriptor for each of the extracted ellipse regions. Although we propose to use MSER and SIFT, the system can work with any combination of region extractor and descriptor.

We briefly explain the algorithm proposed in [1]. The shot boundaries are detected based on the number of LIRs from the current frame that are matched with the LIRs from the next frame. A continuous matching of these LIRs across a number of frames is considered as a *track*. A track can start, either from a local region detected at the first frame of the shot or from local invariant regions at intermediate frames which are not matched to any of the LIRs in the previous frame. Once a shot boundary is detected we obtain a meaningful representation of the shot by obtaining a collective descriptor for each of the track by averaging the descriptors of the local invariant regions which are associated with that track. The discontinuous nature (due to occlusion or failure of the MSER algorithm due to extreme conditions) of the tracks are handled by further comparing each track to all other mutually exclusive (in terms of frames) tracks within the shot (using the L2 distance between the average descriptors of the tracks). If that distance is less than a threshold, then the two tracks are joined together and a new average descriptor is calculated. A shot is now represented by the average shift descriptors of the detected LIR tracks within that shot. Once the track descriptors are calculated for all the shots of the video, each track is compared with the rest of the tracks, and any track that matches more than a threshold number of tracks, is removed. This will help to remove very common tracks which frequently appear in all the shots. This step will enable to reduce the mismatches among the shots caused by these common tracks. This is similar to the stop list exploited by [7].

2.2 Grouping of Tracks

In the previous section we described how local invariant regions are connected from frame to frame to form tracks within a shot. While track descriptors are sufficient for some applications such as scene retrieval [1], further refinement is needed for object mining and other applications. Hence the next important step is the grouping of tracks from a shot into clusters that approximately correspond to a number of objects in that shot. We define two tracks belong to the same object in a shot if the spatial distance between the tracks are small and approximately constant in the frames where they both appear (common frames to both

(a)

(b)

Fig. 2. Some of the clusters obtained from the movie *Groundhog Day*. Clusters in each row approximately cover the same object regardless of the different imaging conditions.

tracks). The task of grouping the tracks into meaningful clusters is difficult due to the discontinuous nature of the tracks. Moreover any clustering algorithm should cope with size variation of an object in different shots. It is not straight forward to use standard clustering algorithms to achieve these goals. This section describes a heuristic procedure which works in practice.

We start by defining a distance measure between an object cluster and a track in a shot (equation 1, Algorithm box 1). This distance is essentially an average of distances between the cluster center and track center in the shared frames, normalized by the covariance matrices of the cluster at each frame. Given a number of object clusters, a track is assigned to the cluster with a minimum distance, if that minimum distance is below a threshold value. Otherwise a new cluster is started with the track as it's first member. Cluster centers and covariance are recalculated after we have completed the assignment of all the tracks. The procedure is repeated until convergence.

It is important to normalize the cluster-to-track distance by the covariance matrix of the cluster to handle the scale variation of objects in different shots. For example, without this normalization a face appearing in a long shot will be clustered into a single object, whereas the same face in a close shot may be grouped into a number of different object clusters with eyes and nose appearing in separate clusters. Fig 2 shows the starting tracks of some of the clusters obtained from the film *Groundhog Day*. The clusters are localized by a bounding rectangle, using the spatial positions of the LIRs in the tracks. Clusters in each row of this figure covers the same object which appear in different parts of the film. As seen in the figure, they approximately cover the same objects regardless

Fig. 3. LIR tracks of the clusters *Both clusters obtained from a shot from the movie* Groundhog Day*. They (start from the first frame and last until the final frame of the shot. The LIR trakcs of the clusters are kept together even if the objects covered by the clusters disappear or are occluded in some of the frames of the shot (between 17551 and 17565 and between 17858 and 17902).*

of the different imaging conditions (scaling in Fig 2(a) and camera angle in Fig 2(b)). Fig 3 shows the LIR tracks for two of the clusters obtained from a shot from the movie *Groundhog Day*. Both of these clusters start from the first frame and last until the final frame of the shot. These clusters approximately cover two interesting objects: face of a person and a name board with letters *PHIL*. The LIR trakcs of the clusters are kept together even if the objects covered by the clusters disappear or are occluded in some of the frames of the shot. The face of the actor in Fig 3 disappears in some continuous frames (between frames 17551 and 17565 and between 17858 and 17902) due to the movement of the face away from the camera. However the LIR tracks are recovered and kept together within the same cluster once the face reappear in the shot. As explained in the previous section, the discontinuous tracks help to keep the LIRs within a cluster through occlusions.

2.3 Object Mining

The previous sub-section described a scheme for grouping tracks into object clusters within a shot. This section explains a object mining module for grouping

different instances of the same object from different clusters into a single object group. Input to the module is the object clusters from all the shots in the video. Each object cluster contains a number of tracks. A track itself is described by the average descriptor of LIRs in it. A similarity value between two object clusters is obtained as the percentage of matched tracks. Consider two object clusters $\mathcal{X}$

Algorithm 1. Obtaining object clusters from tracks in a shot

Notation

$\mathbf{m}_c^f$ center of object cluster c at frame f
Ψ_c^f covariance of object cluster c at frame f
$\mathbf{m}_t^f$ center of track t at frame f

1. Initialize

The algorithm is initialized with a single cluster with the longest track in the shot.

2. Iterate

a **Assignment of Tracks**
For each track t in the shot, we calculate the distances to each of the current object cluster $\mathcal{O}_c$

$$D_{ct} = \sum_f D_{ct}^f = \sum_f \left(\mathbf{m}_c^f - \mathbf{m}_t^f\right)^T \left(\Psi_c^f\right)^{-1} \left(\mathbf{m}_c^f - \mathbf{m}_t^f\right) \tag{1}$$

and the closest object cluster c^* is given by

$$c^* = \underset{c=1}{\overset{K}{\operatorname{argmin}}}\, D_c t \tag{2}$$

if $D_{c^*t} < \tau_r$ then the track t is assigned to the object cluster c^*, otherwise a new object cluster is started with track t as it's first member.

b **Estimating the parameters of object clusters**
for each cluster object cluster $\mathcal{O}_c$ and for each frame associated with that cluster, calculate the following parameters

$$\mathbf{m}_c^f = \frac{1}{N_\mathcal{O}} \sum_{t \in \mathcal{O}_c} \mathbf{m}_t^f \tag{3}$$

$$\Psi_c^f = \frac{1}{N_\mathcal{O}} \sum_{t \in \mathcal{O}_c} \left(\mathbf{m}_c^f - \mathbf{m}_t^f\right)^T \left(\mathbf{m}_c^f - \mathbf{m}_t^f\right) \tag{4}$$

3. Convergence

The algorithm is terminated when there is no change in assignment of tracks.

and $\mathcal{Y}$ with N_x and N_y tracks. First an optimal one-to-one matching between the tracks are obtained by minimizing the following cost function using the Hungarian algorithm [13].

$$C(\mathcal{X}, \mathcal{Y}) = \sum_t C(\mathbf{X}_t, \mathbf{Y}_{\phi(t)}) \tag{5}$$

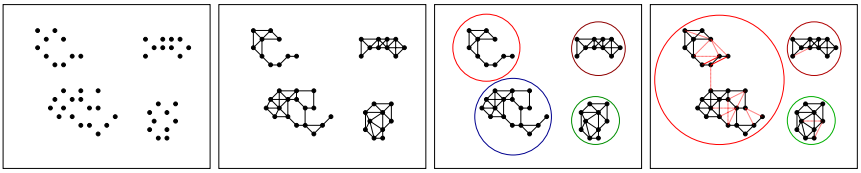


Fig. 4. Grouping of object clusters. (a) Each object cluster is considered as a vertex in a graph. **(b)** Two vertices are connected if the matching ratio of the corresponding object clusters is above a threshold τ_{ob} (see text). **(c)** Mutually exclusive groups of connected vertices are identified and classified as object groups using standard graph connected components algorithm. **(d)** Additional connections, introduced due to a smaller τ_{ob}, are displayed in red color. This leads to a merging of two, previously separate, object groups into a single object group. A smaller τ_{ob} will give fewer and larger object groups.

where, the individual matching cost between the two tracks, $C(\mathbf{X}_t, \mathbf{Y}_{\phi(t)})$, is the L2 distance between the average SIFT descriptors of the tracks $\mathbf{X}_t$ and $\mathbf{Y}_t$. The matching between the two set of tracks is further pruned by only allowing matches with the individual matching cost less than the threshold value τ_s. Let N_{xy} be the number of matched tracks. If the matching ratio, $R(x, y)$, defined as

$$R(x, y) = \frac{N_{xy}}{\min(N_x, N_y)}, \tag{6}$$

is higher than a threshold value,τ_{ob} then the object clusters $\mathcal{X}$ and $\mathcal{Y}$ are defined as belonging to the same object group. Since the number of tracks in each object cluster vary, match ratio is a more appropriate measure of similarity than the absolute number of matched tracks.

However, simple grouping of object clusters may not be sufficient to identify same objects due to large appearance variation and partial occlusion of objects in shots. Therefore we devised an algorithm to group the different instances of an object into same cluster. Formerly the algorithm can be easily explained using a graph. Let each object cluster be a vertex in a graph. The edge between the object cluster V_i and V_j is defined if the two clusters are matched.

$$E_{ij} = \begin{cases} 1 \text{ if } \quad R(i, j) > \tau_{ob} \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

The graph will contain a number of connected components which are completely disconnected with respect to each other. Each connected component will correspond to an object group. Standard graph algorithms are used to calculate the number of connected components as well as to identify the individual object clusters of each object group. The algorithm is illustrated on an example in Fig 4. If $\tau_{ob} = 1$, each of the vertex will form a separate connected component. If $\tau_{ob} = 0$, all the vertexes will join to form one single connected component.

Table 1. Mined Objects The table shows results for some of the mined objects. Correct-Number of correctly mined clusters, Miss-Number of Missed clusters, False-Number of wrongly mined clusters, Pre-Precision, Re-Recall.

Object	Correct	Miss	False	Pre	Re
	81	292	1	0.988	0.217
	70	131	1	0.986	0.348
	6	4	0	1.000	0.600
	5	2	0	1.000	0.714
	13	1	0	1.000	0.929
	5	0	0	1.000	1.000
	5	0	0	1.000	1.000
	23	3	0	1.000	0.885
	8	4	1	0.889	0.500
	9	4	0	1.000	0.692
	7	7	0	1.000	0.500
	6	2	0	1.000	0.750
	8	3	0	1.000	0.727
	5	1	0	1.000	0.833
	5	5	0	1.000	0.500
	10	3	0	1.000	0.769

If $0 < \tau_{ob} < 1$, vertexes will be grouped into a number of separate connected components. A smaller τ_{ob} value (0.1) is used in the mining experiments.

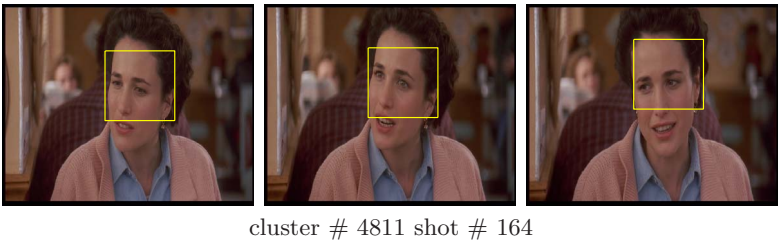
It is interesting to note that the same category objects are joined into a single group with a lower τ_{ob} value. For example, different faces are joined to form a group, different name boards are brought together into a single group and different vehicles are joined to one group. This is because they have similar features (features from eyes, nose and mouth for the face group and features from same *letters* for name board group). This is shown in Fig 4(d). Therefore with the aid of a classification algorithm, these phenomenon can be exploited in classification of object groups in videos.

3 Results

The proposed algorithm is experimented with a full length movie *Groundhog Day*. The movie is divided into shots and features are extracted from stable tracks as explained in the proposed approach. These shot features are grouped

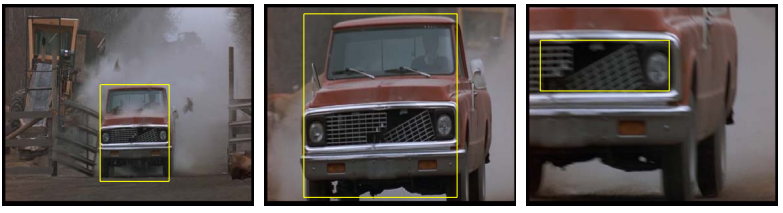
cluster # 11618 shot # 403

cluster # 5686 shot # 191



cluster # 4811 shot # 164

Fig. 5. Sample clusters of a mined face. The face of the actress is mined in 70 clusters from 70 different shots. The figure shows three of these mined clusters. Each row shows sample frame from a particular mined cluster. The cluster number and shot number of each mined instance are given under each row.



cluster # 13897 shot # 493

cluster # 13853 shot # 489

cluster # 13473 shot # 477

Fig. 6. Sample clusters of front of a vehicle. It is mined in 10 clusters from 10 different shots. The figure shows three of these mined clusters. Each row shows sample frame from a particular mined cluster. The cluster number and shot number of each mined instance are given under each row.

into clusters. The mining module mined the different instances of the objects from different clusters into relevant groups. Any object is considered as a valid mined object only if it is mined in at least five different shots. Under this condition 147 objects are mined from the whole movie. Some of the mined objects and the number of clusters they are mined from are given in Table 1. Since we mine only one cluster from a shot into any cluster group, the number of shots in which an object was mined will be equal to the number of clusters mined for that object. Table 1 also gives the number of missed and false clusters and precision, recall values for the given mined objects. The ground truth positions are manually obtained. A shot is selected as a valid representation for an object even if that object partially appear in the shot. For future comparison purposes, the total number of object appearance (in terms of number of shots) in the movie is also given in the table. Fig 5 and 6 shows some sample clusters of two mined objects. Despite the difference in imaging conditions, different instances of the same objects are mined together. It is worth noting that the system is able to

mine small objects (eg microphone) and big objects (eg front of a vehicle) with a single set of parameters.

Compared to the approach proposed in [8], our algorithm mine more instances for most of the objects. Since the ground truth values can be different, precision and recall values can not be used for this comparison. However the number of shots which are mined for a particular object is a useful measure and our approach mined more number of shots for most of the objects compared to the approach proposed in [8]. Moreover our approach able to mine most of the interesting objects with single set of parameters, while neighborhood sizes should be different to mine different size objects for the approach proposed in [8].

4 Conclusions and Future Work

A novel framework for automatic object mining based on local invariant region descriptors is proposed. An algorithm previously published by the authors is used to divide the video into meaningful shots and extract representative features from stable tracks. These stable tracks are grouped into meaningful object clusters. These object clusters may contain a number of similar instances of the same object, and these instances are grouped together in the mining stage. The performance is evaluated with a full length movie, and excellent results are shown. Future work will consider classifying video objects using the clustered features.

References

1. Anjulan, A., Canagarajah, N.: Video Scene Retrieval based on Local Region Features. In: Proc. ICIP 2006 (2006)
2. Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., Simoudis, E.: Mining business databases. Communications of the ACM 39, 42–48 (1999)
3. Etzioni, O.: The World-Wide Web: quagmire or gold mine? Communications of the ACM 39, 65–68 (1999)
4. Oh, J., Lee, J., Kote, S., Bandi, B.: Multimedia data mining framework for raw video sequences. In: Zaïane, O.R., Simoff, S.J., Djeraba, C. (eds.) MDM/KDD 2002 and KDMCD 2002. LNCS (LNAI), vol. 2797, pp. 18–35. Springer, Heidelberg (2003)
5. Zhang, Z.: Mining surveillance video for independent motion detection, In: Proc. International Conference on Data Mining, pp. 741–744 (2002)
6. Chen, S.-C., Shyu, M.-L., Chen, M., Zhang, C.: A decision tree-based multimodal data mining framework for soccer goal detection. In: Proc. International Conference on Multimedia and Expo, pp. 265–268 (2004)
7. Sivic, J., Schaffalitzky, F., Zisserman, A.: Efficient Visual Content Retrieval and Mining in Vidoes. In: Proc. Pacific-Rim Conference on Multimedia (2004)
8. Sivic, J., Zisserman, A.: Video Data Mining Using Configurations of Viewpoint Invariant Regions. In: Proc. International Conference on Computer Vision and Pattern Recognition (2004)
9. Sivic, J., Zisserman, A.: Efficient Visual Content Retrieval and Mining in Vidoes. In: Proc. Pacific-Rim Conference on Multimedia (2004)

10. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal on Computer Vision, vol. 65 (2005)
11. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: Proc. British Machine Vision Conference, pp. 384–393 (2002)
12. Lowe, D.G.: Distinctive image features from scale-invariant key points. Int. Journal of Computer Vision 60, 91–110 (2004)
13. Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38, 325–334 (1987)

Automatic Extraction of Semantic Relationships from Images Using Ontologies and SVM Classifiers*

Jin-Woo Jeong[1], Kyung-Wook Park[2], OukSeh Lee[1], and Dong-Ho Lee[1,**]

[1] Department of Computer Science and Engineering, Hanyang University
[2] Sensing & Perception Research Group, KITECH
Ansan-si, Kyeonggi-do, 426-791, Korea
{jwjeong, oukseh, dhlee72}@cse.hanyang.ac.kr,
kwpark@kitech.re.kr

Abstract. Extracting high-level semantic concepts from low-level visual features of images is a very challenging research. Although traditional machine learning approaches just extract fragmentary information of images, their performance is still not satisfying. In this paper, we propose a novel system that automatically extracts high-level concepts such as spatial relationships or natural-enemy relationships from images using combination of ontologies and SVM classifiers. Our system consists of two phases. In the first phase, visual features are mapped to intermediate-level concepts (e.g, *yellow, 45 angular stripes*). And then, a set of these concepts are classified into relevant object concepts (e.g, *tiger*) by using SVM-classifiers. In this phase, *revision module* which improves the accuracy of classification is used. In the second phase, based on extracted visual information and domain ontology, we deduce semantic relationships such as spatial/natural-enemy relationships between multiple objects in an image. Finally, we evaluate the proposed system using color images including about 20 object concepts.

Keywords: Content-based image retrieval, automatic image annotation, machine learning, ontology, support vector machine, semantic annotation.

1 Introduction

Most users often tend to use the abstract notion involved in an image when searching for an image. However, traditional content-based image retrieval approaches [1, 2] have a critical problem called semantic gap that makes it difficult to search for images using high-level concepts. Therefore, there are needs for the development of tools that support the semantic retrieval for a large image database.

Recently, to resolve this problem, researchers in this field have been studying various techniques such as machine learning techniques [3, 4] and relevance feedback [5]. Especially, automatic image annotation techniques using the support vector machine

* This work was supported by Korea Research Foundation Grant funded by the Korea Government(MOEHRD) (KRF-2006-521-D00457).
** Corresponding author.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 184 – 194, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(SVM) have been suggested [6, 7]. SVM, which is a kind of supervised learning approaches, is efficient for classification and regression of data.

However, conventional approaches using SVMs have a limitation that they cannot support various high-level semantics. For example, it is impossible for us to search by using the semantic query such as "*find an image that contains a zebra, which is located on the left-side of a tiger*". To overcome these limitations, we present a novel framework which combines SVMs with ontologies for rich semantic annotation.

Ontology provides explicit definitions of different concepts and the relationships between them in a specific domain. In particular, from the perspective of image retrieval, it can help users more easily retrieve the desired images because it is possible to explicitly describe the knowledge contained in an image.

In this paper, we propose a novel approach which automatically extracts high-level concepts such as spatial/natural-enemy relationships as well as object concepts from images. In our work, first, low-level visual features are automatically extracted and mapped to intermediate-level concepts, and a set of these concepts are used for learning object SVM-classifier. The revision module has been proposed to revise misclassifiable labels in this phase. Second, various high-level concepts are inferred from semantic processing module based on domain ontologies and visual information which are extracted in classification phase. Finally, whole image semantics are described in I*mageAnnotation* ontology. In particular, the ImageAnnotation ontology is domain-independent so that we can easily apply various fields such as medical science, astronomy. In this work, we examine the semantic-processing for extracting high-level concepts in animal images by using *ecosystem* ontology with the Image-Aannotation ontology.

The rest of this paper is organized as follows: In Section 2, we briefly review related work. Section 3 describes the architecture of the proposed system and methods that extract rich semantics from images in detail. Section 4 shows the experiment results. Finally, we conclude our work in Section 5.

2 Related Work

Many approaches based on SVM have been proposed for automatic image annotation.

In Jianping Fan et al. [6], in order to recognize n object concepts, n concept detection functions are generated. Each concept detection function is a classifier which tells whether an image is relevant or not. However, it may suffer from the problem of the parameter optimization for each detection function. Therefore, this may lead to high complexity as the number of training data and concepts increases.

Tsai et al. [8] proposed two-level learning framework with multiple SVMs. First, images are partitioned into 5 regions with the same size (i.e., the tiling image), and then visual features are extracted from each regions. In the first mapping phase, visual features are classified into corresponding terms (e.g, 'yellow'). In the second mapping phase, SVM maps the set of these terms to relevant concepts such as a tiger, a wolf.

Although this approach made an effort to annotate images more effectively and accurately, it still has some limitations. First, this approach without consideration for the weight of representative features may degrade the accuracy of classification. Generally, each concept may have their own representative visual features (e.g., tigers may have orange/black stripes as their representative visual features). Therefore, a consideration for

these representative features may help to effectively identify a certain object from other objects which have similar visual features. Second, the result of image annotation using SVMs does not contain rich semantic information such as *spatial relationships* or *natural-enemy relationships* between multiple objects in an image. Therefore, we cannot search images by using semantic query based on semantic relationships between multiple objects in an image. For example, it is impossible for us to use the query such as "*find images whose image category is a photo about carnivores*" Or "*find an image that contains an animal which has a natural-enemy relationship with a zebra*". Consequently, recent researches using only SVM-classifier just extracts fragmentary information (e.g, *this image contains a tiger and a horse*), because they did not apply any knowledge processing (e.g, deduction, induction) to the information that are extracted from images.

Therefore, in order to extract more rich semantic information, a knowledge processing must be required. In this paper, we present a novel framework which automatically extracts various high-level semantics by integrating ontologies and SVM which has a consideration for weight of representative visual features.

3 System Architecture

Figure 1 shows the framework of our system which consists of *image classification phase* and *semantic processing phase*. In the image classification phase, visual features are automatically extracted from images, and classified into the relevant object concepts. In particular, the revision module, which will be described in Section 3.1.4, corrects mis-classifiable results of SVM-classifier. This module can improve the accuracy of classification. Every kinds of information which are generated in the image classification phase are passed to *semantic inference module*. This module infers various semantic information by using *ImageAnnotation* ontology, domain ontology, and visual information generated in the image classification phase.

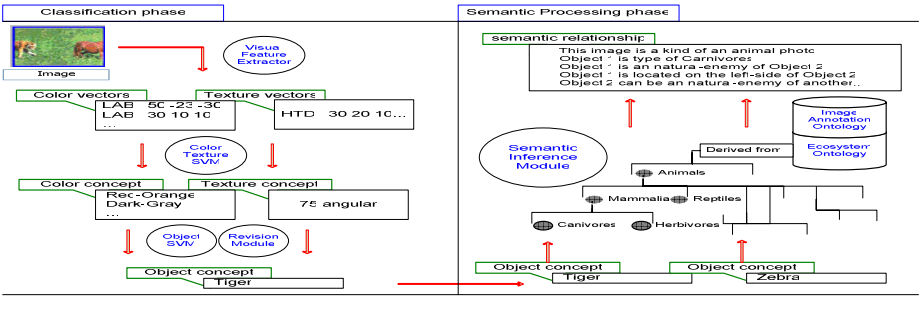


Fig. 1. Framework of the Proposed System

3.1 Image Classification

The visual pattern of images that have the same concept tends to be similar. However, since shape is so sensitive to an external environment, we use only color and texture features for image classification. Each feature is described as CIE LAB color-space and MPEG-7 Homogeneous Texture Descriptor (HTD), respectively.

3.1.1 Mapping Visual Features to Intermediate-Level Concepts

As mentioned above, each concept may have its own representative visual features. In our work, training data for the classifier consists of 4 representative colors and one representative texture. In addition, color histogram and texture histogram of object are also extracted for revision module.

Given training image set, each image is partitioned into 8x8 blocks. The partitioning of an image enables recognizing multiple objects and spatial relationships between them with low complexity for segmentation. Afterward, 4 representative colors and one texture feature are extracted from each image blocks. The extraction of representative colors is performed based on *Wu-quantization* algorithm [9].

After extracting of representative features, our system maps these low-level features to intermediate-level concepts via color/texture SVM-classifiers. Intermediate-level concepts are simple keywords such as yellow, black, 45-angular diagonal, which can describe the characteristics of each visual feature. We defined 512 color concepts and 15 texture concepts. Training data for intermediate-level SVM classifier consists of a label and feature vectors. (e.g, training data for texture SVM-classifier is represented as (45-angular diagonal, [10, 10, 10…30])).

Color SVM-classifier associates a low-level color vector (i.e, *LAB vector*) with the relevant intermediate-level color concept. For example, color vectors, which are represented in CIE-LAB [0, 0, 0] ~ [10, 10, 10], are mapped to *Black-DarkGray* concept. Similarly, texture SVM-classifier associates a low-level texture vector (HTD) with the relevant intermediate-level texture concept, which is represented as a form of 15-angular diagonal, 45-angular diagonal, and so on. In case of a tiger, the process of mapping visual features to the intermediate-level concept is as follows. First, 4 representative colors are extracted from an image by using *Wu-quantization* algorithm. And then a representative texture is extracted via MPEG-7 HTD. Second, both LAB vectors and HTD vector are classified into corresponding intermediate-level concepts. Through these procedures, intermediate-level concepts for a tiger, [Orange, RedOrange, Gray, Black, 75-angular diagonal], are generated.

3.1.2 Extraction of Object Concepts

Based on the intermediate-level concepts, training procedure for object SVM-classifier is performed. Training data vector for recognizing an object concept C_i is represented as $T_i = \{L_i, DC_{0-3}, DT\}$, where L_i is a label, DC_{0-3} are 4 representative color concepts, and DT is a representative texture concept for C_i, respectively.

Given image I_i, this image is classified into an object concept C_i by object SVM-classifier. However, as mentioned above, previous researches lack a consideration for the weight of representative features per object concept. Therefore, an object O_i may be misclassified into the irrelevant concept which has similar visual features. To solve this problem, we propose a revision module which corrects mis-classifiable results by calculating the distance of weighted features between O_i and object concepts. In general, SVM-classifier produces one prediction concept and the *confidence* values associated with each concept as the result of classification process. The confidence value represents relevance between an image and its concepts. Suppose that as a result of classification for image I_i, each confidence value associated with concepts {Tiger, Gazelle, Eagle, Horse, Zebra}, is produced as {0.7, 0.15, 0.05, 0.07, 0.03}. In this case,

the classifier determines a tiger as a prediction concept for this image. It can be mentioned that this prediction is reliable because the confidence value of a concept *Tiger* is very high. However, the following example shows a different case. Let's assume that based on the confidence set {0.35, 0.30, 0.12, 0.27, 0.03}, an image I_j is classified into a concept *Tiger* by SVM-classifier. In this case, however, there rarely exists significant difference between highly-ranked confidence values. This means that visual features of an image are not so distinctive that we correctly recognize the original concepts. Hence, this image may be classified into the irrelevant concept *Tiger* even though correct concept is *Gazelle*. Consequently, in this case, a revision module which can correct mis-classifiable labels is required.

3.1.3 Revision of Image Concept

```
Algorihm_1. ConceptDecision(p, conf[]) returns concept
// p is a predicted label, conf is a set of confidence
values that are generated by object SVM-classifier
1: Candidates <- highly ranked 5 values of conf[];
2: STD_Candidates = STDDEV(Candidates);
3: If( STD_Candidates < Th1 && confidence(p) < Th[p])
4:   Label = Revision(CH,TH);
5: Else Label = p;
6: Return Label;
```

Algorithm_1 describes how the object concept of an image is determined from predicted concept p and confidence values which are generated by object SVM-classifier. In line 1, highly ranked 5 values are selected from the confidence vector. And then, a standard deviation among these values is calculated in line 2. The value of the standard deviation is a criterion for the revision process. If this value is high enough, we can treat this problem as a first example mentioned in Section 3.1.3, or else this corresponds to the second example. Therefore, we determine whether the result by the object SVM-classifier is reasonable or not. In line 3-5, if the value of the standard deviation is less than a threshold Th_1, revision process is performed because predicted concept p may not be reliable. However, if the confidence value which is concerned with p is greater than a threshold Th_p, the result is not revised. Th_p is the minimum value which ensures the reliability of the result produced by object SVM-classifier. The optimal value for Th_1, Th_p can be determined by experiments.

```
Algorithm_2. Revision(Ch,Th)returns object concept
// Ch,Th represents color histogram and texture histo-
gram of a target image, respectively
1: ConceptVector CV,TV;
2: For each concepts
3:   For each features // color and texture
4:     For each bins
5:       Dist_C = Dist_C + square(CW*CV-CW*Ch);
6:       Dist_T = Dist_T + square(TW*TV-TW*Th);
7:     EndFor
8:   EndFor  //apply modified Euclidean equation
9:   Dist[concept] = Dist_C + Dist_T;
10:EndFor
11:Return Label(Min(Dist));
```

Algorithm_2 shows a detailed process of the revision module. In this algorithm, a variable CV, TV denotes a representative color histogram and a representative texture histogram of each object concept, respectively. And a variable CW, TW represents the weight for features of each object concept. In order to revise mis-classifiable result, first, we calculate a distance between visual features of a target image and each object concept stored in image DB by using modified *euclidean* distance function with a consideration for the weight information. And then, we choose the concept which has a minimal distance with a target image as object concept for an image. The weight for each bin x of the color/texture histogram $W_x^{c,t}$ is computed by equation (1).

$$W_x^{\,c} = 1-(bin_x^{\,c} \,/\, Area_c)^2, \;\; W_x^{\,t} = 1-(bin_x^{\,t} \,/\, Area_t)^2, \;\; c \in \{L, A, B\} \tag{1}$$

Where, $Area_{c,t}$ represents the total area of color and texture histogram, and $bin_x^{c,t}$ represents the area that each bin x holds in each histogram, respectively. We revise confused labels by weighting to frequently appeared bins in each histogram and calculating a distance between a target image and each object concept. For example, in case of a bright image (e.g, *snow image*), a luminance histogram of this image tends to be skewed to a certain side. Thus, we weight to frequently appeared bins in the histogram of snow images so that we can more correctly identify snow images.

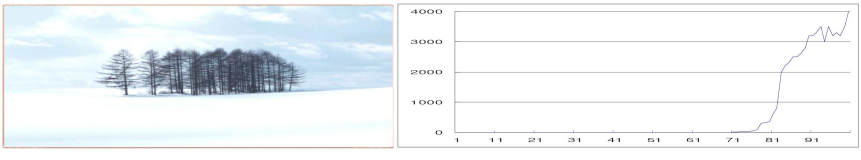


Fig. 2. Snow Image and its Luminance Histogram

3.2 Semantic Processing

As depicted in Figure 1, visual information of an image which are extracted in the classification phase are passed to *ImageAnnotation* ontology. And then we infer rich semantic information such as spatial relationship or natural-enemy relationship from the *ImageAnnotation* ontology and domain ontology. Finally, deduced semantics are represented as an OWL document.

3.2.1 ImageAnnotation Ontology

In order to effectively infer and manage various information and knowledge contained in an image, we propose the *ImageAnnotation* ontology shown in Figure 3. Detailed descriptions for important classes are as follows:

- *ImageAnnotation* class is a basic unit for describing one image. It has a relationship with *ImageCategory* class and *Object* class.
- *ImageCategory* class describes a category of an image. The category is classified into artifacts (human-arts, buildings), naturals (landscaped, animals), and so on.

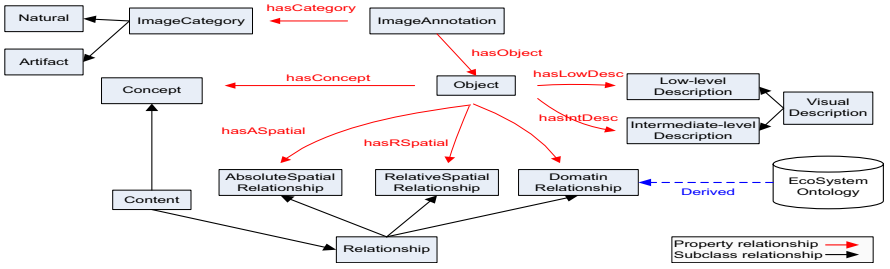


Fig. 3. A simple structure of the *ImageAnnotation* ontology

- *Object* class has various information about an object in an image, and can be generated as many as the number of objects in an image. Each object has its own object concept, visual information, semantic relationship, etc.
- *Concept* class describes the name, synonyms, definition, hierarchical level of an object concept. These data are derived from WordNet 2.1[10].
- *Relationship* class has a description about semantic relationships such as spatial / natural-enemy relationships which can be derived from domain ontologies.
- *DomainRelationship* class describes semantic relationships between objects by using terms inherited from domain ontologies. In our work, we deduce semantic relationships like *isSymbiosisOf* or *isConsumerOf* by using *ecosystem* ontology.
- *VisualDescription* class describes visual features of an object. Each description is represented as a form of a low-level vector and intermediate-level concepts.

3.2.2 Inference of Semantic Relationships

We can infer various semantic relationships by applying corresponding inference rules to images. Some rule engines for ontology reasoning have been proposed. We use *Bossam* rule engine [11] which provides a rule language called *buchingae*.

- **ImageCategory Information**

We can deduce a type of an image by applying categorization rules to images. For example, an image that contains a 'tiger on the grass' is classified into *Natural* and *Animals* photo. This inferencing is performed based on the fact that if an object contained in an image is a type of *living thing* and *animal*, this image describes *natural* and *animal* things. These terms for hierarchical level are derived from *WordNet 2.1*. The inference rule for recognizing natural-scene images is as follows:

```
Rule isNatural is:
        If hasObject(?x,?y)and hasConcept(?y,?z)and
           hasHierarchicalLevel(?z,?w) and
           contains(?w,'living thing') or
           contains(?w,'NaturalPhoenomenon')
        Then isNatural(?x)
```

Similarly, other category such as animals, landscape can be deduced easily.

- **Relative Spatial Information**

If there exists multiple objects in an image, more complicated query is available (e.g, "*find an image that contains a tiger which is located on the left-side of a bear*"). Relative spatial relationships between multiple objects can be represented as *isWestOf*, *isNorthOf, isNorthEastOf*, etc. In order to deduce spatial relationships between objects, first, area that each object holds is represented by MBR (Minimum Bounding Rectangle). Afterward, we deduce corresponding relationships by applying inference rules to objects in an image. Suppose that two object A and B are located as shown in Figure 4. One of inference rules for extracting relative spatial relationship between two objects A, B can be represented as follows:

```
Rule isNorthOf is:
            MBR(?A),MBR(?B)
            LowerRightY(?A,?y1),UpperLeftY(?B,?y2)
            IF    isLessThan(?y1,?y2)
            Then  isNorthOf(?A,?B)
```

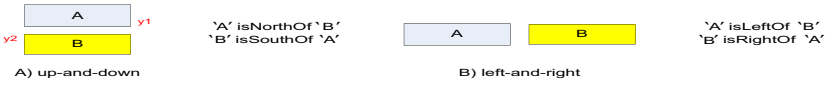


Fig. 4. Relative spatial relationships between objects

- **Natural-enemy Relationship Information**

In our work, we examine the semantic relationships about an ecosystem by using *ecosystem* ontology. These domain-dependant ontologies can be made manually, or inherited from existing ontologies. In case of *ecosystem* ontology, various semantic relationships such as *natural-enemy / Symbiosis* are described in detail. Every time an object is classified into a relevant object concept, semantic relationships related to an object concept are included in *ImageAnnotation* ontology. For example, information that tigers or lions are natural-enemy of a zebra, and zebras generally take a grass are included in images which contain a zebra.

```
1  <ImageAnnotation rdf:ID = "Image_001">
2          <hasImageCategory rdf:ID= "Animals"/>
3          <hasObject rdf:resource = "#Object_01"/>
4          <hasObject rdf:resource = "#Object_02"/>
5  <Object rdf:ID = "Object_01">
6          <hasConcept rdf:resource = "#Zebra"/>
7          <hasDomainRelship rdf:resource = "#ecosystem_zebra"/>
8                  ...
9  </Object>
10 <DomainRelship rdf:ID = "ecosystem_zebra">
11         <isConsumerOf rdf:resource = "#Grass_01"/>
12         <isProducerOf rdf:resource = "#Tiger"/>
13                 ...
14 </DomainRelship>
```

This OWL document denotes a part of *ImageAnnotation* ontology for *image_001*. All generated information like a category of image and information of objects in an image are integrated and represented in the OWL document. As shown in line 5-9, a concept of an object and various semantic relationships are described. In line 10-14, semantic relationships such as *isConsumerOf* about a concept *zebra* are described. In addition,

visual information and spatial information of each object in an image are also depicted. We can use this visual information for searching images using content-based queries (i.e, searching by color or texture keywords are also available). Based on these OWL documents, our system can support various semantic queries.

4 Experiments

In this section, we evaluate the performance of our image classification module and demonstrate a semantic query process. Evaluation is on the color image database which consists of above 6,000 color images with about 20 concepts.

4.1 Performance Evaluation of Image Classification

For objective evaluation, our system is evaluated with SVM-only and K-Nearest Neighbor which is known as an effective classification algorithm. Also, the threshold Th_1, Th_p for the revision module is set to 0.35 and 0.25, respectively. Precision ρ and recall r are used to measure the performance.

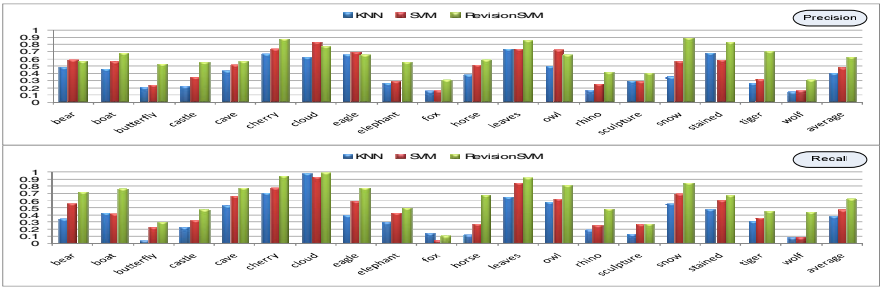


Fig. 5. Precision and recall of KNN, SVM-only, and SVM with the revision module

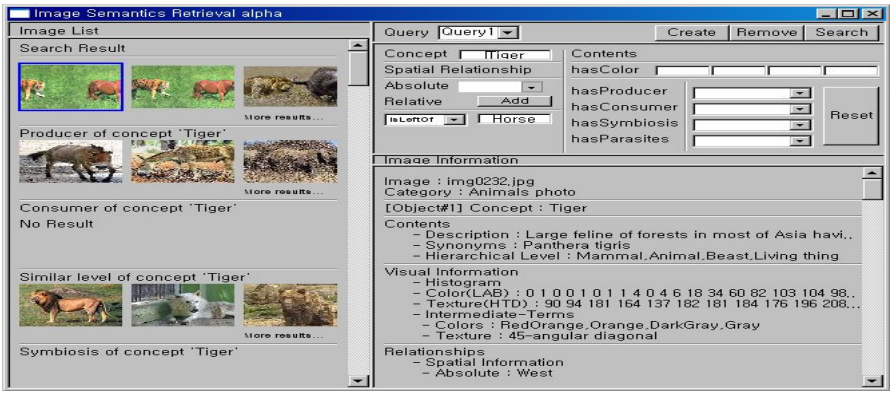


Fig. 6. Example of Semantic Query

$$\rho = \eta/(\eta+\varepsilon) \quad , \quad r = \eta/(\eta+\vartheta) \tag{2}$$

Where η is the number of images that are relevant and are detected correctly, ε is the number of images that are irrelevant but are detected, ϑ is the number of images that are related but are mis-detected. A comparison of the performance for each algorithm is given in Figure 5. As shown in Figure 5, the performance of proposed system is outstanding in terms of average precision and recall. Our system outperforms SVM-only and KNN about 50%, 30%, respectively. Also, we can see that our system increases the accuracy of annotation for objects which have no distinct visual features (e.g, rhino, elephant, castle) as well as objects which have distinct visual features(e.g, cherry, snow, etc .).

4.2 Result of Semantics Query

Figure 6 shows the prototype of semantic query process in our system. If a user queries "*find an image that contains a tiger which is located on the left-side of a horse*", our system searches for images that contain a tiger object whose value of *isLeftOf* property is the identifier of the horse object contained in the same image. Searching is based on the inference modules mentioned in Section 3.2. The result of searching is shown in the search result panel. In addition, we can see that images which are concerned with a result image are also returned. For example, images about siblings of a tiger in a food chain (e.q, *lion, bear*) are presented. When a user selects one image from the result, visual/semantic information of selected image are displayed. Moreover, various semantic queries like "*what is the symbiosis of a crocodile?*" can be processed by using semantic processing modules.

5 Conclusion

In this paper, we proposed a novel framework using a combination of SVMs and ontologies for extracting high-level concepts from images. For this, we performed classification by training multiple SVMs with the revision module, and deduced high-level semantics by applying inference rules to domain ontologies. Our main contributions are as follows: 1) our system supports various semantic processing on the image database. 2) our revision module improved the accuracy of image classification using SVMs. 3) since the *ImageAnnotation* ontology is domain-independent, our system can be applied to other domains easily. 4) it is worth noting that our system is fully automatic, with the exception of collecting training samples and including some inference rules. As a further work, we are now studying the techniques for automatic annotation on video database.

Reference

1. Smith, J.R., Chang, S.-F.: VisualSEEK: a fully automated content-based image query system. ACM Multimedia 96
2. Carson, C., Thomas, M., Belongie, S., Joseph, M., Hellerstein.: Blobworld: A System for Region-Based Image Indexing and Retrieval. Visual Information Systems (1999)

3. Minka, T.P., Picard, R.W.: Interactive Learning Using a Society of Models. Pattern Recognition' 97 (1997)
4. Zhang, Q., Goldman, S.A., Yu, W., Fritts, J.E.: Content-Based Image Retrieval Using Multiple-Instance Learning. In: Proc. Machine Learning'02 (2002)
5. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: A power tool for interactive content-based image retrieval. IEEE Trans. Circuits Syst. Video Technol (1998)
6. Fan, J., Gao, Y., Luo, H., Xu, G.: Automatic Image Annotation by Using Concept-Sensitive Salient Objects for Image Content Representation. ACM SIGIR'04 (2004)
7. Li, B., Goh, K.-S., Chang, E.Y.: Confidence-based Dynamic Ensemble for Image Annotation and Semantics Discovery, In: ACM MM'03 (2003)
8. Tsai, C.-F., McGarry, K., Tait, J.: Automatic Metadata Annotation of Images via a Two-Level Learning Framework, In: ACM SIGIR'04 workshop on semantic web, pp. 32–42 (2004)
9. Wu, X.: Color quantization by Dynamic Programming and Principal Analysis, In: TOG' 92 (1992)
10. Wordnet2.1 http://wordnet.princeton.edu/
11. Jang, M., Sohn, J.-C.: Bossam: An Extended Rule Engine for OWL Inferencing

A Robust Caption Detecting Algorithm on MPEG Compressed Video

Yaowei Wang[1], Limin Su[2,*], and Qixiang Ye[3]

[1] Department of Electronic Engineering, Beijing Institute of Technology,
Beijing, P.R. China, 100080
yaoweiwang@gmail.com
[2] Information School, Beijing Union University,
Beijing, P.R. China, 100101
limin.su@gmail.com
[3] Graduate School of Chinese Academy of Science
Beijing, P.R. China, 100080
qxye@jdl.ac.cn

Abstract. Captions (or overlay texts) play an important role in video content understanding. In this paper, an algorithm is proposed to detect captions in MPEG compressed video. First, energy features, which are used to find candidate caption blocks, are extracted from DCT coefficients. Second, temporal information is employed to verify these candidate blocks. Then, a new region growing method named "density-based region growing" is proposed to connect these blocks into candidate text regions. Finally, the regions are identified as caption or non-caption by structural information of caption regions. Experiments are conducted on news videos and it is shown that the algorithm is feasible and effective in finding captions.

1 Introduction

Content-based video retrieval is a long-term research topic to manage the increasing digital video resources. Text, which is considered containing plentiful semantic information, is a valuable clue to solve the retrieval problems. For example, captions in a news video usually annotate information on where, when, and who of the reported events [1]. The appearance of a caption in sports video probably means a "goal" event.

In recent years, many researchers study on the method to detect the video captions [1-8]. Most of the studies are in spatial domain. They use edge (gradient) analysis [2][6], color connected component analysis [4] or wavelet features [4] to detect text lines or captions. To detect the appearance and disappearance of a text line [7], temporal features are also employed.

Unlike those methods in spatial domain, methods in compressed domain yield great advantages in the following aspects:

* Author is supported by Project (Beijing Union University Scientific Research Plan for imported person(12213HC0516)) and Project (Beijing Union University Funding Plan for Middle-aged and Youthful Faculty(12213HZ8802)).

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 195 – 202, 2007.
© Springer-Verlag Berlin Heidelberg 2007

1) Smaller data volume;
2) Lower computing complexity and high speed.

Zhong [1] presented a method to automatically localize captions in the I-frames of MPEG compressed videos. Candidate captions or text regions are localized directly in DCT compressed domain by using horizontal and vertical intensity variations, which are called texture features in his paper. Although the method has good performance on most video captions; it does not work if a text region has similar luminance and different chromas to its background. The reason is that only luminance information is employed. In [8], Wang employed a supervised color model to detect news video captions in MPEG compressed domain. The DCs are used as color features of an 8 x 8 block. A DCT block is regarded as a caption block only if it meets the learned color model. Yeo and Liu [9] extracted embedded captions partially in uncompressed MPEG video. First, lower resolution video frames are reconstructed from original MPEG sequences by using DC components alone or DC components and two AC components. Then, text regions are marked wherever a large inter-frame variance is observed. Unfortunately, the algorithm fails when captions are gradually appearing.

To overcome these shortcomings, we develop a robust caption detection algorithm in MPEG compressed domain.

In the proposed method, "robust" is treated as the most important thing:

1) Robust to text font-size, color;
2) Robust to complex background;
3) Few parameters need to be modified.

And video captions' properties are summarized as follows:

1) Violent irregular luminant or chromatic variances;
2) Captions lasting a few seconds;
3) Caption characters bounding in size and its spatial structure.

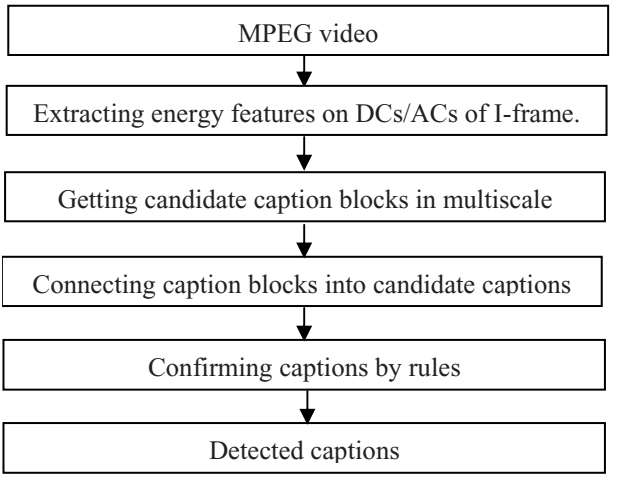


Fig. 1. Flow chart of the caption detecting algorithm on MPEG compressed domain

Violent luminant or chromatic variances in a text region can be formularized as energy features of candidate caption blocks. And candidate blocks can be verified by using successive I-frames (according to property 2)). For property 3), caption blocks can be connected into caption regions. At the same time, unreasonable captions are eliminated.

Fig.1 is the flow chart of our algorithm. There are four steps: feature extraction, candidate blocks detection and verification, connecting candidate blocks into candidate captions and captions confirmation.

The rest of the paper is organized as follows: in section 2 the algorithm is presented. Experimental results are presented in Section 3. Finally, we conclude the paper in Section 4.

2 Caption Detecting Algorithm on MPEG Compressed Domain

2.1 Multiscale Feature Extraction in I Frame

For each I-frame in MPEG video, DCT information (for each 8×8 image block) can be obtained directly without the IDCT (Inverse DCT) computing. A DC component represents the mean value of the block. AC components (ACs) demonstrate the spatial frequency and directionality. In a DCT block (the 8×8 block), ACs can be used to present the luminant or chromatic variations and their directionality. DCs of adjacent blocks are used to capture the variations among adjacent image blocks. By using ACs in just one block and ACs/DCs among adjacent image blocks, the multi-scale luminant/chromatic variations of the text region can be computed. Fig.2 shows the employed coefficients in one 8x8 DCT block and 4 adjacent DCT blocks.

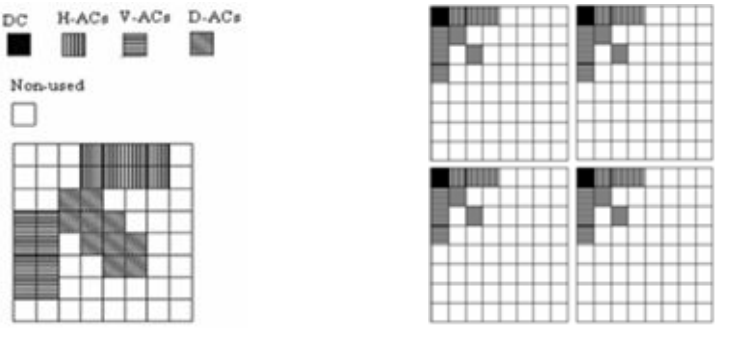


Fig. 2. DCT coefficients used in the proposed algorithm. (a) ACs for original scale and (b) DCs and ACs in four adjacent DCT blocks for second scale.

The luminant/chromatic variations are expressed as energy features. For possible violent luminant or chromatic variations in a text region, a caption region's energy features are usually higher than its background.

To detect texts in different font-size, the energy features are extracted in the first two scales, the original scale and the second scale.

In the original scale, features are obtained from the coefficients of H-/V-/D-ACs (as shown in Fig.2a). Energy features are calculated by summing up ACs on the horizontal, vertical and dialog directions.

$$E_{i,j} = E_{i,j}^{H} + E_{i,j}^{V} + E_{i,j}^{D} \tag{1}$$

Where $E_{i,j}^{H/V/D} = \sum_{i,j} \left| (C_{i,j}(u,v) \right|$, i and j represent the location in a DCT block, $C_{i,j}(u,v)$ is the DCT coefficient, H/V/D is the coefficient of H-/V-/D-ACs (Fig.2).

In the second scale, features are extracted from the coefficients represented by DC and H-/V-/D-ACs (as shown in Fig.2b). Energy features are calculated by summing up the value of ACs and the square variance of DCs in four adjacent blocks (Fig.2b).

$$E_{i,j} = E_{i,j}^{DC} + \sum_{n=0}^{3} E_{i,j}^{H} + \sum_{n=0}^{3} E_{i,j}^{V} + \sum_{n=0}^{3} E_{i,j}^{D} \tag{2}$$

Where $E_{i,j}^{H/V/D} = \sum_{A=0}^{3} \sum_{i,j} \left| (C_{i,j}(u,v) \right|$, i, j are the locations in a DCT block. n is the index of four adjacent DCT blocks in Fig.2b, and $E_{i,j}^{DC}$ is the DC variance in the four adjacent blocks.

To detect in larger scales, we can extent the adjacent blocks numbers from four to eight.

2.2 Candidate Caption Blocks Detection and Verification

Blocks, whose energy is above a threshold, will be regarded as candidate text blocks. The method proposed by Zhong and Zhang [1] is employed to identify the threshold. An adaptive threshold is that t times the average energy of the selected corresponding DCT channel.

$$B_{i,j} = \begin{Bmatrix} 1, if\ E_{i,j} > t \cdot E_{i,j}^{avg} \\ 0, otherwise \end{Bmatrix} \tag{3}$$

Where "1" is candidate block and "0" is background. Experiment results show that the algorithm performs best when t is set as 1.80.

Then, temporal information is computed to verify the caption blocks. For each caption will last for several seconds, it is reasonable to simply assume that a caption block should be found in at least three continuous I frames (about 1 second). So, those blocks that cannot last three I frames will be discarded. Fig.3 shows an example.

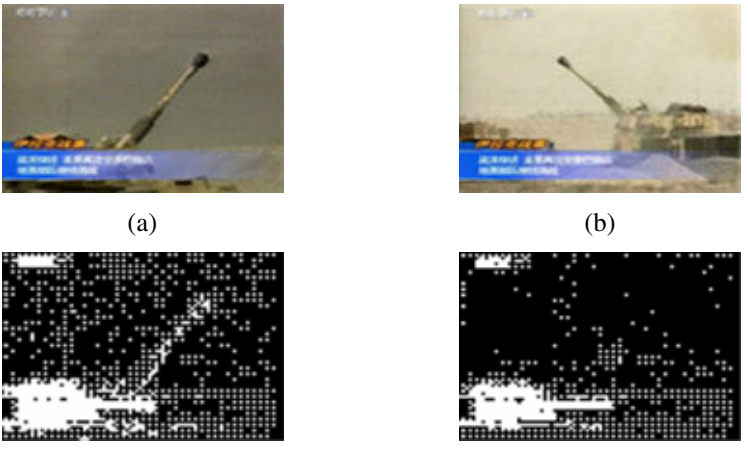


Fig. 3. Detected caption blocks. (a) is the first I-frame where the caption appears. (b) is the third I-frame containing the caption. (c) is the detected caption blocks in the first I-frame. (d) is the final detected caption blocks after verification in three I frames.

2.3 Connecting Candidate Blocks Up Into Caption Region

In general, a caption is composed of a cluster of text blocks (these blocks should be dense). Thus, isolated blocks could be treated as noise.

Most researchers [1] [9] [10] compute the connectivity of blocks by the morphological operations, such as dilation. But morphological operation cannot find the "denseblocks". All adjacent blocks will be connected to form a region whether they are members of a "cluster" or not.

In our algorithm (we name it "density-based" region growing), region growing is used to find the connectivity of blocks. A block B should be treated as a seed block of a region if there are more than 50% of blocks, which are candidate caption blocks in the neighborhood (as shown in Fig.4a) of B. In an image, we define that block B is density-connected with block B' only if B is in the neighborhood of $B^{'}$, and $B^{'}$ is a core block. Thus, we define the "density-based" region grow method.

"Density-Based" Region Growing Alrorithm:

1) Search the unlabeled blocks in an image for current seed block and current core region.
2) If a seed block B is found, a new cluster is created. Then, we iteratively collect unlabeled blocks that are density-connected with B, and label these blocks with same cluster label.
3) If there are core blocks in the image goto 2).
4) Label each cluster found in the image as a caption.
5) Merge blocks, which are not included in any caption region, into background.

Fig.4 demonstrates that the "density-based" region growing method performs better than morphological operations.

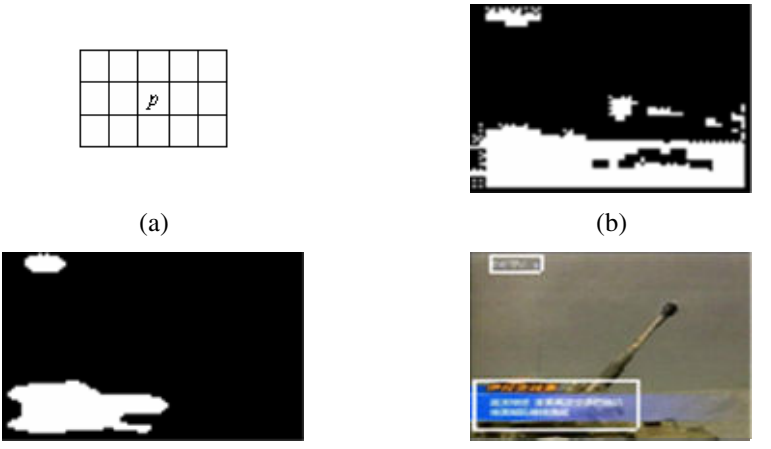


Fig. 4. (a) is the neighborhood of block B . (b) is the caption find by dilation operation and (c) is the text region found by "density-based" region growing. (d) is the final detected text.

2.4 Confirming Captions

Though we have some candidate captions now, there may be some false alarms, such as some texture regions. Following rules are employed to identify whether a candidate caption block is a real caption block.

The Rules:

1) *Caption_region_width* / *caption_block_height* >1.0 (< 1.0 for vertical caption region);
2) *Caption_region_height* is between T_1 and T_2 ;
3) The ratio between *horizontal /vertical /dialog* energy of text region is between T_3 and T_4 .

Where *text_block_height* and *text_block_width* are the height and width of a caption. Rule 1) is used to restrict that a caption area will be a rectangle, whose width is larger than height (for horizontal caption). Rule 2) is used to limit the size of the caption in terms of the screen size. Rule 3) is to restrict that captions must have a similar energy in its horizontal, vertical and dialog directions. This rule can eliminate some horizontal, vertical or dialog textures that often have a much larger energy in one direction than other directions. In this paper, T_3 and T_4 are set as 0.3 and 3.0.

3 Experimental Result

In our experiments, two videos in MPEG-7 test set [11] and four video of CCTV (Chinese Central Television) are employed to evaluate the method. There are 146 captions in these videos (denoted by people). Precision and recall [12] are employed to evaluate the performance of the algorithm.

$$\text{Re}\,call = \frac{a}{c},\ \text{Pr}\,ecision = \frac{a}{a+b} \tag{4}$$

Where a is the number of detected captions and b is the number of non-caption detected as caption, and c is the total number of truly caption in the test set.

CCTV video frame size is 72 576 and the MPEG-7 video frame size is 355 288. The detected captions' font-sizes are varying from 8 pixels to 64 pixels. Results (compare with Zhong's method) are showed in table 1.

Table 1. Performance and comparison

	Recall	Precision
Proposed algorithm	94.4%	92.0%
Algorithm of Zhong	92.7%	90.9 %

Method of Zhong is used for comparison. Zhong's method misses some captions, in which text has a large font-size or has a similar luminance with its background. The recall rate of Zhong is lower than proposed method. Zhong's method will induce more false alarms, for only spatial information is used in verification. It shows that our algorithm is more reasonable in terms of video caption detection. It also proves that the algorithm is benefited from the multi-scale spatial features, temporal features, luminant and chromatic information, and the "density-based" region growing method instead of morph operations. Fig.5 shows some examples.

It takes about 0.01 second to process one I-frame and about 0.05 second to detect a caption in the successive I-frames on Pentium IV CPU (1.7 G Hz).

Fig. 5. Examples of some detected captions

4 Conclusions

In this paper, a video caption detection algorithm in MPEG compressed is proposed. The algorithm is robust in different font-sizes, different font-colors and different kinds of videos. Experiment results show that the proposed algorithm is feasible and

effective. Our algorithm has a higher precision and a higher recall than Zhong's method. It should be noted that texts in the captions can be recognized for semantic indexing of videos by OCR software.

References

[1] Zhong, Y., Zhang, H.J., Jain, A.K.: Automatic Caption Localization in Compressed Video. IEEE Transactions on PAMI 22(4) 385–392 (2000)
[2] Chen, X., Yang, J., Zhang, J., Waibel, A.: Automatically Text Detection and Recognition in Natural Scene Images. IEEE Trans. Image Processing (2004)
[3] Smith, M.A., Kanade, T.: Video skiming for quick browsing based on audio and image characterization. Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-95-186 (1995)
[4] Jain, A.K., Yu, B.: Automatic text location in images and video frames. Pattern Recognition 31(12), 2055–2076 (1998)
[5] Li, H., Doermann, D., Kia, O.: Automatic text detection and tracking in digital Video, In: IEEE Trans. on Image Processing, 9(1) (January 2000)
[6] Lienhart, R., Wernicke, A.: Localizing and segmenting text in images and videos. IEEE Trans. on CSVT, 12(4) (April 2002)
[7] Tang, X., Gao, X.B., Liu, J., Zhang, H.: Spatial-Temporal Approach for video caption detection and recognition. IEEE Trans. on Neural Networks. 13(4) (July 2002)
[8] Wang, W.Q., Gao, W., Li, J.T., Lin, S.X.: News content highlight via fast caption text detection on compressed Video. In: Int. Conf. on Intelligent Data Engineering and Automated Learning (2000)
[9] Chen, D.T., Bourlard, H., Thiran, J.-P.: Text Identification in Complex Background Using SVM. In: Int. Conf. on CVPR (2001)
[10] Xi, J., Hua, X.-S., Chen, X.-R., Liu, W., Zhang, H.J.: A video text detection and recognition system. In: Int. Conf. on Multimedia and Exp (2001)
[11] Zhi, D.G., Lin, Y., Chua, T.-S., Kankanhalli, M.: Ground truth data and analysis of shot boundaries in the mpeg7 video content set. In: Technical Report, National University of Singapore (1999)
[12] Raghavan, V., Bollmann, P., Jung, G.: A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. on Information Systems 7, 205–229 (1989)

A Hybrid Approach for Authenticating MPEG-2 Streaming Data

Meiqin Wang[1] , Lin Li[1], S.M. Yiu[2], Lucas C.K. Hui[2], C.F. Chong[2], K.P. Chow[2], W.W. Tsang[2], H.W. Chan[2], and K.H. Pun[2]

[1] Laboratory of Cryptographic Technology and Information Security, Ministry of Education, Shandong University
Jinan, 250100, China
{mqwang, lli}@sdu.edu.cn*
[2] Department of Computer Science
The University of Hong Kong
Pokfulam Road, Hong Kong SAR, China
{smyiu, hui, chong, chow, tsang, hwchan, pun}@cs.hku.hk**

Abstract. There are two main approaches for authenticating digital streams, namely the *tree chaining* [1] and the *hash chaining* [2,3]. Both approaches have their disadvantages. Hash chaining is superior for low communication overhead, however, it is not resilient to packet loss and it has a longer verification delay. On the other hand, tree chaining is more robust even if packet loss occurs and with shorter verification delay, but the communication overhead is too high to be tolerable, especially in online applications. In this paper, we try to combine the two techniques and propose a hybrid scheme for authenticating MPEG-2 streaming data, which are still used by real application systems, by taking advantage of the characteristics of MPEG frames. The hybrid approach is shown to be more effective as compared to the other two approaches.

1 Introduction

The bandwidth of Internet has increased tremendously, which enables the digital streams (e.g. video clips in the CCTV) to be delivered to the receivers (e.g. the control room) in an online manner. In some applications, such as CCTV monitoring, the receiving data needs to be authenticated to prevent malicious attackers from injecting harmful data into the streams. Sometimes, non-repudiation may need to be guaranteed as well. A straight-forward solution to this problem is to use digital signature. However, standard digital signature schemes are not specially designed for continuous data streams. If one tries to sign each frame

* Supported by NFSC Key Project No.90604036, National Outstanding Young Scientist. No.60525201, National 973 Project No.2007CB807902, and Open Project Program of National Key Laboratory of Switching Technology and Telecommunications Networks Beijing University of Posts and Telecommunication.

** Supported in part by a grant from the Research Grants Council of the HKSAR, China (Project No. HKU/7144/03E), and a grant from the Innovation and Technology Commission of the HKSAR, China (Project No. UIM/145).

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 203–212, 2007.
© Springer-Verlag Berlin Heidelberg 2007

in the stream individually (say use the RSA algorithm), the computation is too intensive and may result in an unviewable video. A more feasible approach is to authenticate digital stream by *amortizing* a single signing operation over multiple packets. Another issue for signing digital streams is the possibility of packet loss. Usually a connectionless and unreliable transport layer protocol UDP is used in real-time applications for lower communication overhead. So, not all the frames are guaranteed to be delivered successfully nor the lost frames will be retransmitted. A practical solution should be robust even with packet loss.

By chaining up multiple packets to reduce the number of digital signing operations, there are two common existing approaches for authenticating digital streams: *tree chaining* and *hash chaining*. The basic idea of hash chaining is as follows. Every m packets are grouped into a block. For the second packet, the corresponding message digest is computed upon the concatenation of the packet and the message digest of the first packet. The message digests of the subsequent packets are computed similarly. And we only sign the message digest obtained from the last packet. To verify the whole block, we verify all the message digests and finally verify the signature on the last message digest received. The expensive signing and verification is done only once for the whole block. The verification of all other packets in the same block can be done using the inexpensive hashing function. This approach has low communication overhead, however, the receiver needs to receive all preceding packets in the same block in order to verify the current received packet. In other words, the scheme does not work if packet loss occurs. In [2], Gennaro et al. reduced the impact of packet loss by appending hash value to multiple locations. Golle and Modadugu reduced the effect of the bursty packet loss by appending the hash value of each packet in some important locations [3], but they still cannot resist the loss of the signed packet.

Tree chaining was proposed in [1]. Roughly speaking, every m packets are used to build a binary authentication tree with the message digests of these n packets as leaves. Each parent is the message digest of its two children. Only the root is signed. The signature is called the block signature of these packets. To make the scheme robust with packet loss, they make sure that each packet can be verified individually. So, each packet carries the information of the block signature and the siblings of each node in the packet's path to the root. It is clear that loss in packets will not affect the authentication of other received packets. However, the disadvantage of this scheme is on the heavy communication overhead that each packet should carry.

In this paper, we consider the authentication of an important type of streaming data, the MPEG-2 frames. Note that although MPEG-4 standard exists, MPEG-2 still have a lot of applications such as those in existing CCTV systems. We combine the above two approaches and propose a hybrid method to authenticate MPEG-2 streaming data. Our proposed scheme is shown to be effective than the other existing approaches. As a remark, there also exist approaches for authenticating MPEG-4 streaming data (e.g. [11]), those methods are dedicated for MPEG-4 and are too complicated for MPEG-2 data. In the rest of

the paper, we only consider MPEG-2 data, so we simply refer it as MPEG data. The rest of the paper is organized as follows. In Section 2, we review the two existing approaches for authenticating digital streams. Section 3 introduces the characteristics of MPEG stream. Section 4 introduces our scheme using MPEG frames as an example. Section 5 compares our scheme with the existing schemes. Conclusions are given in Section 6.

2 Existing Schemes to Authenticate Digital Stream

2.1 Tree Chaining

In the tree chaining authenticating scheme [1], Wong et al. group m packets into a block and build a binary authentication tree using the message digests of these m packets as leaves. An example is given in Fig 1, eight packets are in a block and D_i represents the message digest of the i-th packet. A parent in the tree is the message digest of the concentenation of its two children. For example, the parent of the leaves D_1 and D_2 is $D_{1-2} = H(D_1, D_2)$, where $H()$ is the message digest function. The root, D_{1-8} in the example, is the block digest. Only the block digest is signed. In this scheme, in order for each packet to be verified independently, the authentication information of each packet consists of the block signature, the packet position of the block and siblings of each node in the packet's path to the root.

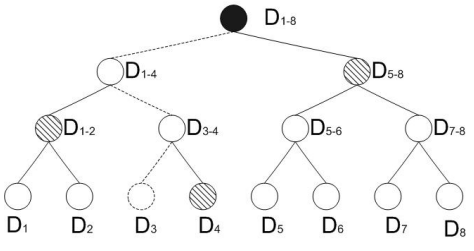


Fig. 1. Tree Chaining

This scheme is resistant to packet loss as each packet can be verified independently, however, the amount of authentication information carried by each packet is huge. This communication overhead may make the scheme infeasible for some real-time critical applications.

2.2 Hash Chaining

The basic idea of the hash chaining technique is as follows [2]. For a block of m packets, we only perform the expensive signing and verification once as follows. For the second packet P_2, its message digest is computed on the concatenation

of P_2 and D_1 where D_1 is the message digest of P_1. The generation of the hash chain is described in the following. We only sign on the last message digest D_m.

$$\begin{aligned} D_1 &= H(P_1) \\ D_2 &= H(P_2||D_1) \\ &\vdots \\ D_m &= H(P_m||D_{m-1}) \end{aligned}$$

If D_m has been verified to be correct, all the packets in this block should be correct. Note that the sender only needs to create one digital signature for the whole block. However, the basic scheme is not resistant to packet loss as it requires all preceding packets to verify the current received packet. But the communication overhead is lower than that of the tree chaining technique. In order to provide some resistance to the effect of packet loss, Gennaro and Rohatgi proposed a scheme called *EMSS* [2]. It appends the hash value of each packet in multiple locations. So, if some of the packets are lost, the required hash values of the preceding packets may still be available. And Golle and Modadugu proposed the use of *augmented chain* based on *EMSS*, which has the ability to resist the bursty packet loss by appending the hash value of the packet to important locations [3]. Once a packet is lost, the hash values of the preceding packets hash values in the hash chain appended to the lost packet still can be authenticated by the other appended packet. So the resistance to burst losses can be realized by additional hash connection in the hash chain. The advantages of this scheme are low communication overhead, low computational overhead and resistant to the burst loss of a given length. The drawback of the scheme is that the delay of the verification and non-resistance to the loss of a signed packet.

3 Characteristics of MPEG

Before we describe our hybrid scheme, we first describe the characteristics of MPEG frames as we will demonstrate our scheme using MPEG as an example.

In MPEG, there are several types of frames. An I-frame (Intra-frame) is a still image. A P-frame (Predicted frame) is predicted from the most recently decoded I-frame or P-frame. B-frame (Bidirectional frame) is interpolated from the closest two I- or P-frames. A MPEG sequence begins with an I-frame, and ends with the preceding frame of the next I-frame. Such a sequence is called a Group of Pictures (GoP). An important feature of common MPEG encoders is the manner in which frame types are generated. Although not required by the standard, typical encoders use a fixed GoP pattern when compressing a video sequence. We divide the possible GoP patterns into three types as follows.

1. There are only I-frames in the GoP, the Group of Pictures(GoP) pattern is represented as I.
2. There are several P-frames between two I-frames, the GoP pattern is $IPP\cdots P$.

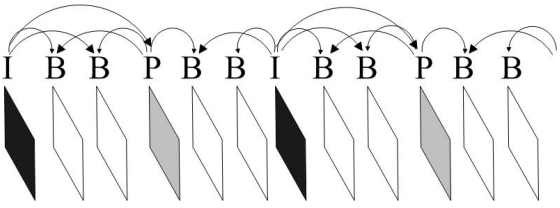


Fig. 2. Dependency Structure for a GoP pattern in an MPEG sequence(N=6, M=3)

3. There are several B-frames between I-frame and P-frame, the GoP pattern is $IBB \cdots PBB$.

Usually the GoP pattern is quite "regular" in the sense that the number of B-frames between two reference frames (I or P) is fixed. Such a GoP pattern can be characterized by two parameters: the I-to-I frame distance (N), and the I-to-P frame distance (M). In this paper, for ease presentation, we focus on a "regular" GoP pattern $IBBPBBPBB$, which is a typical MPEG sequence. It should be emphasized that MPEG-compliant sequences need not use a "regular" GoP pattern, and not even a repetitive one. However many sequences are being coded with regular GoP patterns in order to simplify the code design. As described in [9], the encoding and transmission orders of an MPEG sequence exhibit similar periodic behavior in the GoP pattern except for the first few frames. It is reasonable to assume that a stream consists of replications of a given GoP pattern.

Fig 2 shows the dependency structure among the three types of frames. From the figure, one can see that the loss of I-frame results in the non-viewable GoP and has a much worst impact on the video quality among the three types of frames. On the other hand, the loss of P-frame has a worse impact than the loss of B-frame. In [4], authors have taken into account the frame priority information to minimize the effect of frame losses on the quality of the video. Our proposed authenticating method for MPEG stream data will consider the different importance of the three types of frames. There are three kinds of packet losses.

1. The loss of an I-frame will result in all other frames in the same GoP unviewable.
2. The loss of a P-frame will result in the dependent B-frames unviewable.
3. The loss of a B-frame only affects this particular frame.

In the next section, we will show how to integrate the hash chaining and the tree chaining to authenticate MPEG streaming data. Note that although we employ the Group of Pictures(GoP) pattern $IBBPBBPBB$ as an example, the proposed strategy is general and can be applied to any GoP pattern.

4 The Proposed Scheme

In this section, we propose a hybrid scheme by integrating the ideas of tree chaining and hash chaining.

In the hash chaining scheme, if the hash value of each packet is appended to two positions, the computational overhead will increased greatly. In our proposed scheme, we determine how to append the hash value according to the priority of different packets. As presented in Section 3, I-frame is the most important, and B-frame is the least important. So the hash value of I-frame should be appended to multiple locations, but not the hash value of P-frame or B-frame.

In practical applications, the block size should be set appropriately. If we utilize the tree chaining to sign these streams, the scheme can provide the resistance to the packet losses for the individual verification of each frame, but it will result in the high communication overhead which is unacceptable. If we sign these streams based on hash chaining scheme proposed by Golle and Modadugu in [3], the computational overhead and communication overhead will be reduced and the scheme is resistant to bursts of up to a given length, but the large block size will incur the impracticable verification delay. This is the main shortcoming of hash chaining. In fact, if the block size is limited to a small value, and there is a signature to a block, the size of the signature will become relatively large when compared with a hash value.

For example, for an 1024-bit RSA signature and SHA-1 hash function, the size of the signature is more than 100 bytes. Assuming that the block size is 16 packets, tree chaining authentication (we use binary tree) will result in about 100 bytes signature and the hash value in overhead is 80 bytes; hash chaining authentication will incur 100 bytes signature and the hash value in overhead is 20 bytes. In the example, the hash chaining technique is superior to the tree chaining technique. On the other hand, if we select the block size to be 64 packets. The tree chaining will result in about 100 bytes signature and the hash value in overhead is 120 bytes. For the hash chaining, it will incur about 100 bytes signature and the hash value in overhead is still 20 bytes. So from the viewpoint of the communication overhead, larger block size favours hash chaining. Unfortunately, in hash chaining, the verification delay in the second example is four times longer than the verification delay in the first example. In the second example, in order to reduce the verification delay, the block size can be chosen to be 16, but it will produce about 400 bytes signature and four-time signature and verification which cannot be accepted in some practical applications.

4.1 The Scheme Description

The overview of the scheme is shown in Fig 3. The frames are divided into groups, the hash chaining method is used to verify each group. (Note that we just labelling the frames as I-frame, P-frame, or B-frame arbitrarily for illustration purpose. The scheme is applicable to any combination of frames.) Assume that the upper bound of verification delay is the time to receive m packets, we will set the group size to m. Suppose that there are n packets in all($n >> m, m|n$),

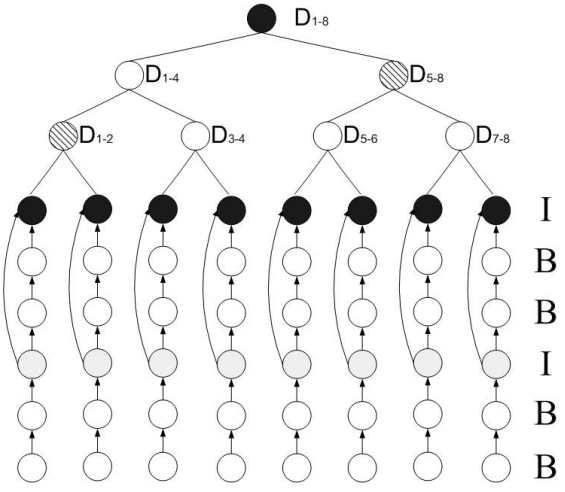


Fig. 3. Hybrid Authentication Technique

we will sign one time for n packets in order to reduce the time of signature. We locate each m packets on a hash chain, and the first packet hash is regarded as a leaf node in a binary tree. So there are n/m leaf node in the tree. Based on tree chaining, we only sign the root node. For each m packets to be individually verified, in each group the preceding $(m-1)$ packets except the first one are distributed without additional authentication information, but the first packet needs to carry the group's authentication information which consists of the block signature, the position of the group of packets, and the siblings of each node in the group packet's path to the root.

A simple hash chain for each group cannot be resistant to packet loss in the group. So we will take the measures proposed by Golle and Modadugu in [3]. Based on the different priority of m packets, the hash values of high priority packets are located to more positions to provide the resistance to the packet loss.

To verify a group of packet individually, the receiver needs to receive the necessary packets in the same group and calculates the hash value of the hash chain, then authenticates the group's path to the root.

In our proposed scheme, the trade-off between verification delay and communication overhead can be controlled by changing the block size n and the group size m. Increasing m will incur the long verification delay and low communication overhead.

5 Comparison with Other Schemes

We compare our scheme with the tree chaining and the hash chaining methods. Note that for hash chaining, we use the augmented chain approach as described in [3] for evaluation purpose.

Table 1. Comparison of Three Authentication Schemes

	sender delay	receiver delay	computation overhead	communication overhead	loss
tree chaining	n	1	$2n-1,1$	$s+1+(h+1)\log_2^n$	any
hash chaining	p	n	$n+1,1$	variable	burst
our scheme	n	m	$n/m-1+n$,1	variable	burst

Let n be the number of packets in a block, m be the number of packets in a group, p be the number of packets inserted by using augmented chain, h be the size of a hash value, s be the size of a signature. To compare the schemes, we use the following five parameters [3,7].

- sender delay: The delay on the sender side (in terms of the number of data packets) before the first packet in the block can be transmitted.
- receiver delay: The delay on the receiver side (in terms of the number of data packets) before verification of the first packet in the block is possible.
- computation overhead: The number of hashed values and signatures computed by the sender per block.
- communication overhead: The number of hashed values and signatures computed by the sender per block.
- loss: The type of loss that the scheme resists.

Table 1 shows the comparison of the schemes in terms of these parameters.

In Table 1, the sender delay of our scheme is greater than that of hash chaining, but the sender delay is not a critical requirement in most stream data applications as long as the delay is reasonable. The receiver delay in our scheme is less than that in hash chaining, which is an advantage of our method. Certainly the receiver delay for hash chaining can be reduced by decreasing n, but it will decrease the signing rate (rate for signing of packets) and it may not be very appropriate in most of the applications.

From the same table, the communication overhead varies for hash chaining, because the communication overhead in the signature packet is different from that in the unsigned packet. In our scheme, for m packets in each group, only the first packet of each group will be transmitted with additional overhead including the block signature , the position information and the sibling node's hash value of the path to the root node, but the other packets will be transmitted only with the position information, so the communication overhead in our scheme also vary and which is lower than that in tree chaining.

In terms of the loss for three scheme, there is little effect of packet loss in our scheme because each m packet is verified independently. Therefore any group failed to be verified cannot influence the verification of other received groups. At the same time, in each group, the hash value of P-frame and I-frame has two times to be appended to the hash chain, so the loss of them will not break the hash chain, but for there are only two concatenation B-Frame in each hash chain,

the loss of a B-frame only affects the verification of the preceding B-frame, but the failure to display B-frame has little distortion to the whole video. However the necessary requirement in our scheme is that the leaf node cannot be lost, which is the hash value of I-frame and hash of other node in the chain. The packet contains the block signature, the hash of the path etc. If the packet is lost, other packets in the group cannot be verified.

The main contribution of our scheme is proposing the new hybrid authentication method, which can provide that the tradeoff between the receiver delay and the communication overhead can be adjusted by selecting different m and n. Our scheme is efficient in the application where the packet loss is very serious and the verification delay is very limited. If the fraction of the signature time to the block size is fixed, our scheme can achieve the lower verification delay by moderately increasing communication overhead.

We have also performed some real experiments to evaluate the performance of our schemes. We analyzed 50 avi files and divide these files into three types: (1) High quality, real-scene video files - they have more changes between consecutive frames, so they tend to use only I-frames. (2) Normal quality files - they tend to use moderately fewer I-frames (1 I-frame per 4-20 frames). (3) Computer-made video files (games and animations) - they have fewer changes between consecutive frames, so they tend to use very few I-frames (1 I-frame per about 200 frames).

Note that the size of a packet is about 10240 bytes for all cases, and the transmission speed is about 384KB/s (that is, 3Mbps). These parameters do not affect the performance of the scheme, the reason we set them up is to have some concrete values for comparison.

Based on the experimental results, we have the folloiwng remark. For Type 1 and Type 2 files, we have set two reasonable values of m for the hybrid scheme. We find that, compared with hash chaining, the hybrid scheme greatly reduces the receiver delay while adding little overhead. Also, the overhead for the hybrid scheme is much less than that of tree chaining. Although sender delay of the hybrid scheme is high, it is not a concern in most cases. The hybrid scheme is the best among the three schemes in performance.

On the other hand, as Type 3 files contain very few I-frames and the GoP size is very large, the m value for the hybrid scheme will need to be large enough so that each hash chain can start with a key frame. However, both tree chaining and hash chaining have no such restriction. The advantage in receiver delay of the hybrid scheme over hash chaining is much less when compared with the cases for Type 1 and Type 2 files. Therefore the hybrid scheme, although still the best choice for Type 3 files, is less efficient for the delay than the cases for Type 1 and 2 files.

To conclude, the hybrid scheme, in general, has similar overhead as the hash chaining, but much less delay. When compared with the tree chaining, the hybrid scheme has much less overhead but more delay. However, this delay is reasonable and acceptable. From the perspective of receiver delay and computation and communication overhead, the hybrid scheme seems to be the most efficient scheme among the three. However, the scheme is slightly less efficient in terms of

receiver delay when the GoP size is very large. The detailed experimental results will be presented in the full paper.

6 Conclusions

In this paper, we integrate two typical authenticating schemes, tree chaining and hash chaining, to form a hybrid authentication scheme for MPEG streaming data. In fact, the scheme can be applied to other multimedia data formats that have the GoP (Group of Pictures) concept.

Based on the analysis and the experimentsf, we show that our proposed scheme is more effective than the other existing schemes. Actually, a variation of this scheme is being applied to a practical system and a US patent (Appn No 11/410,004) is being filed and is pending for approval.

Acknowledgement. The authors would like to thank Man Ki Li, Yuk Wah Li, Lai Yin Man and Wing Pan Pang for their assistance in the experimental work.

References

1. Wong, C.K., Lam, S.S.: Digital signatures for Flows and Multicasts. IEEE/ACM Transactions on Networking 7(4), 502–513 (1999)
2. Gennaro, R., Rohatgi, P.: How to Sign Digital Streams, Advances in Cryptology. In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 180–197. Springer, Heidelberg (1997)
3. Golle, P., Modadugu, N.: Authenticating Streamed Data in the Presence of Random Packet Loss. In: Proceedings of the Network and Distributed System Security Symposium, pp.13–22 (2001)
4. Chebroluc, K.: Multi-Access Services in Heterogeneous Wireless Networks, Dissertation (2004) `http://home.iitk.ac.in/chebrolu/docs/thesis.pdf`
5. Safavi-Naini, R., Wang, Y., Wysocki, B.J.: On a Method to Authenticate and Verify Digital Streams. Journal of Telecommunications and Information Technology 2/2002, 45–52 (2002)
6. Miner, S., Staddon, J.: Graph-based Authentication of Digital Streams. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, pp. 232–246 (May 2001)
7. Park, J., Chong, E., Siegel, H.: Efficient Multicast Packet Authentication Using Signature Amortization, In: Proceedings of IEEE Symposium on Security and Privacy, Oakland, CA pp. 227–240 (2002)
8. Merkle, R.: A Certified Digital Signature, Advances in Cryptology(Crypto'89), August pp. 218–238 (1989)
9. Krunz, M., Zhao, W., Matta, I.: Scheduling and Bandwidth Allocation for the Distribution of Archived Video in VOD Systems. Telecommunications System Journal 9(3/4), 335–355 (1998)
10. Nahrstedt, K.: An Architecture for End-to-End Quality of Service Provision and Its Experimental Validation, Dissertation (1995)
11. Wu, Y., Deng, R.H.: Scalable Authentication of MPEG-4 Streams. IEEE Transaction on Multimedia 8(1), 152–161 (2006)

HOS-Based Image Super-Resolution Reconstruction

Jianping Qiao and Ju Liu

School of Information Science and Engineering, Shandong University,
Jinan 250100, Shandong, China
{jpqiao, juliu}@sdu.edu.cn

Abstract. In this paper a novel high-order statistics (HOS) based regularized algorithm for image super-resolution reconstruction is proposed. In this method, the image is divided into various regions according to the local forth order statistics. The segmentation label is then used to determine the weighted operator of the regularization term. In this way, different regularization terms are applied depending on local characteristics and structures of the image. The proposed image achieves anisotropic diffusion for edge pixels and isotropic diffusion for flat pixels. Experimental results demonstrate that the proposed method performs better than the conventional methods and has high PSNR and MSSIM with sharper edges.

1 Introduction

With the development of visual communication and signal processing, there is a high demand for high-resolution (HR) images such as video surveillance, remote sensing, medical imaging, HDTV and so on. But because of the physical limitations and high cost, it is hard to get the HR image. One promising approach is to use image processing technique which is called Super-resolution (SR) image reconstruction. It refers to obtain a single HR image by combining several low-resolution (LR) images.

Numerous SR methods have been proposed in the literatures. Generally, these methods can be categorized into two classes: one is reconstruction-based method such as iterative back-projection (IBP), maximum a prior (MAP), projection onto convex set (POCS), etc [1]; the other is learning-based method [2-3]. Although the learning-based method breaks the magnification limitations, the performance of this kind of algorithms strongly depends on the training images. They are more effective for some special images such as face image super-resolution. In this paper we will focus on the reconstruction-based image super-resolution. In order to solve the ill-posed inverse problem, regularization is used to ensure the solution to be existing and stable. The most common priors have potential functions that are quadratic in the pixel values. But this estimation tends to be overly smooth and the sharp edges are not recovered. To achieve further improvement, Huber MRF prior is proposed as the potential functions [4]. Meanwhile, SR restoration can be considered as a second-generation problem of image recovery, therefore some methods of image recovery can be extended to solve SR problem. In [5] a segmentation-based regularization method for single image restoration is proposed in which regularization term apples the smoothness

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 213 – 222, 2007.
© Springer-Verlag Berlin Heidelberg 2007

constraint on constant areas. In [6] a region-based super-resolution algorithm is proposed in which different filters are used according to the type of region. But in these methods the computation cost is high and also the segmentation information is not fully used.

In this paper we propose a novel regularized super-resolution algorithm based on high-order statistics (HOS). In this method, the forth-order cumulant (kurtosis) of every pixel is computed by a sliding window. The image is segmented into different types of regions according to the local HOS. The weight function of the regularization term is determined by the segmentation label. The proposed method achieves anisotropic diffusion for edge pixel and isotropic diffusion for pixel in smooth region. Experimental results demonstrate that the method is robust to additive Gaussian noise because kurtosis is zero for Gaussian random variable.

This paper is organized as follows: the mathematical model of super-resolution problem is introduced in section 2. In section 3 the proposed HOS-based algorithm is presented. Experimental results are reported in section 4. Section 5 concludes this paper.

2 Mathematics Model

At present, there are two super-resolution models: the warping-blurring model and the blurring-warping model. The former model coincides with the imaging physics but it is usable only if the motion among SR images is known a priori, and the latter model is more appropriate when the motion has to be estimated [7]. In the following, the warping-blurring model is taken as an example to discuss the problem. The relationship between the low-resolution images and the high resolution image can be formulated as:

$$\underline{Y}_k = \mathbf{D}_k \mathbf{H}_k \mathbf{F}_k \underline{X} + \underline{E}_k \tag{1}$$

where $\underline{Y}_k$ is the $M_1 M_2 \times 1$ lexicographically ordered vector containing pixels from the k th LR frame, $\underline{X}$ is the $N_1 N_2 \times 1$ lexicographically ordered vector containing pixels from SR image , $\underline{E}_k$ is the system noise with the size $M_1 M_2 \times 1$, $\boldsymbol{D}_k$ is the decimation matrix of size $M_1 M_2 \times N_1 N_2$, $\boldsymbol{H}_k$ is the blurring matrix of size $N_1 N_2 \times N_1 N_2$, known as the Point Spread Function (PSF), $\boldsymbol{F}_k$ is a geometric warp matrix of size $N_1 N_2 \times N_1 N_2$, $1 \le k \le K$ and K is the number of LR images.

The MAP/POCS hybrid reconstruction approach finds SR estimates by minimizing the following cost function while constraining the solution within certain sets:

$$\underline{\hat{X}} = \underset{\underline{X}}{Arg\min} \left[\sum_{k=1}^{K} \left\| D_k H_k F_k \underline{X} - \underline{Y}_k \right\|_2^2 + \lambda \left\| C \underline{X} \right\|_2^2 \right] \qquad subject\ to \left\{ \underline{\hat{X}} \in C_p \ 1 \le p \le P \right\} \tag{2}$$

where C is a generally a high-pass filter, $\|\cdot\|_2^2$ represents a l_2 norm, λ is the regularization parameter which controls the tradeoff between fidelity to the data and smoothness of the solution, C_p represents the additional constraints such as constraints on the output energy, phase, support and so on.

3 Super-Resolution Based on High-Order Statistics

Although the traditional algorithm given by Eq. (2) provides a stable solution, the SR result looks smoother and the sharp edges are blurred. This is because the reconstructed image is assumed to be global smooth and the local structure of each pixel is not considered. The method in [6] applies different regularization term for homogeneous and inhomogeneous regions. But in this method the image is segmented through eigenvalue decomposition which is sensitive to noise and the segmentation information is not fully utilized. In our method high-order statistics is adopted to extract the local characteristics which will be used to supervise the construction of the weighted operator in regularization term. In this way, the anisotropic diffusion is performed on edge pixel to obtain sharper details and in smooth region isotropic diffusion is performed to make the solution less noisy. We set up a good prior model which alters with the local structure and characteristics of the image.

According to the basic idea, the proposed detail procedure contains two stages, image segmentation and SR reconstruction.

3.1 Image Segmentation

The kurtosis of a random variable z is defined as the forth-order cumulant [8]:

$$kurt(z) = E\{z^4\} - 3(E\{z^2\})^2 \tag{3}$$

where $E\{\cdot\}$ is the expectation over the density of input vector z. If z has a Gaussian density, the forth moment $E\{z^4\}$ equals $3(E\{z^2\})^2$. Thus, kurtosis is zero for a Gaussian random variable. For most non-gaussian random variables such as image and speech signals, kurtosis is nonzero.

The detail procedure of image segmentation based on kurtosis is described as follows. Given an image g, the $2n+1$ by $2m+1$ neighborhood of pixel at $g(x, y)$ is :

$$\begin{pmatrix} g(x-n, y-m) & \cdots & g(x-n, y+m) \\ \cdots & g(x, y) & \cdots \\ g(x+n, y-m) & \cdots & g(x+n, y+m) \end{pmatrix} \tag{4}$$

The neighborhood is converted by column into a vector. For example, $(g(x-n, y-m)\ ,\ \ldots\ ,\ g(x+n, y-m)\ ,\ \ldots\ ,\ g(x, y)\ ,\ \ldots\ ,\ g(x-n, y+m)\ ,\ \ldots\ ,\ g(x+n, y+m))$. This vector becomes z in Eq. (3) and is used to compute the kurtosis

of the pixel $g(x, y)$. By sliding this neighborhood over all positions of the image, kurtosis image $k(g)$ is obtained. Note that the vector here is not zero mean. The mean value of the vector is not subtracted because this kind of criterion measures both local structure as well as luminance value of the image.

The kurtosis image is then quantized according to the given segmentation number. For simplicity, we use uniform quantizer. Step size of the quantizer which is denoted by Δ is decided by:

$$\Delta = (\max(k(g)) - \min(k(g))) / segnum \tag{5}$$

where $k(g)$ is the kurtosis image and *segnum* is the segmentation number. After quantization, every pixel corresponds to an integer or label in $[1, segnum]$. The segmentation label will be used to decide the weighted operator in regularization term in the next step.

3.2 Super-Resolution Reconstruction

Super-resolution reconstruction is an ill-posed problem. Inversion of the problem can be stabilized by regularization. Half-quadratic regularization was defined by Geman and Yang and the regularization term can be expressed as follows [9]:

$$J_2(\underline{X}) = \sum_{(n,m)\in D_{\underline{X}}} \left(\varphi\left(D^x_{n,m}(\underline{X})\right) + \varphi\left(D^y_{n,m}(\underline{X})\right)\right) \tag{6}$$

where

$$\begin{aligned} D^x_{n,m}(\underline{X}) &= \left(\underline{X}_{n,m+1} - \underline{X}_{n,m}\right)/\delta \\ D^y_{n,m}(\underline{X}) &= \left(\underline{X}_{n+1,m} - \underline{X}_{n,m}\right)/\delta \end{aligned} \tag{7}$$

and

$$\varphi(t) = \frac{t^2}{1+t^2} \tag{8}$$

where δ is a scaling parameter, it tunes the value of the gradient and detect the discontinuity. φ is the potential function.

If $\lim_{t\to 0} \varphi'(t)/(2t) = M > 0$, then

$$\frac{\partial J_2(\underline{X})}{\partial \underline{X}} = -2\left\{\lambda_E \underline{X}_{n,m+1} + \lambda_W \underline{X}_{n,m-1} + \lambda_N \underline{X}_{n+1,m} + \lambda_S \underline{X}_{n-1,m} - \lambda_\Sigma \underline{X}_{n,m}\right\} \tag{9}$$

The weight function in regularization term at (n, m) can be expressed as follows:

$$C_w = \begin{bmatrix} 0 & \lambda_N & 0 \\ \lambda_W & -\lambda_\Sigma & \lambda_E \\ 0 & \lambda_S & 0 \end{bmatrix} \tag{10}$$

In our method, the weight function is determined by segmentation labels which are computed in the first stage. Denote the segmentation label of pixel (n,m) be $s(n,m)$, then

$$\begin{aligned} \lambda_E &= \begin{cases} 1 & if\ s(n,m+1) \neq s(n,m) \\ 0 & if\ s(n,m+1) = s(n,m) \end{cases} \qquad \lambda_W = \begin{cases} 1 & if\ s(n,m-1) \neq s(n,m) \\ 0 & if\ s(n,m-1) = s(n,m) \end{cases} \\ \lambda_N &= \begin{cases} 1 & if\ s(n+1,m) \neq s(n,m) \\ 0 & if\ s(n+1,m) = s(n,m) \end{cases} \qquad \lambda_S = \begin{cases} 1 & if\ s(n-1,m) \neq s(n,m) \\ 0 & if\ s(n-1,m) = s(n,m) \end{cases} \\ \lambda_\Sigma &= \lambda_E + \lambda_W + \lambda_N + \lambda_S \end{aligned} \tag{11}$$

For a target pixel, if the segmentation label of its adjacent pixel is not equal to the label of the target pixel, which indicates that they belong to different regions, then the adjacent pixel should have no contribution to the estimation of the target pixel and the corresponding weighted coefficient should be zero. In this way, local adaptive regularization is achieved and different regularization term is performed according to the local characteristic and structure. Meanwhile, the kurtosis for Gaussian noise is zero, so the weight function in regularization term is robust to the noise.

The new cost function for SR reconstruction can be described as:

$$\underline{\hat{X}} = \underset{\underline{X}}{Arg\min} \left[\sum_{k=1}^{K} \left\| D_k H_k F_k \underline{X} - \underline{Y}_k \right\|_2^2 + \lambda \left\| C_W \underline{X} \right\|_2^2 \right] \quad subject\ to \left\{ \underline{\hat{X}} \in C_p\ 1 \le p \le P \right\} \tag{12}$$

The iterative optimization algorithm is described in Algorithm 1.

4 Experiments

We perform several experiments to validate the efficacy and robustness of the proposed algorithm. PSNR and mean structural similarity (MSSIM) [10] are used as the objective quality measurement of the reconstructed image. The traditional method is described in Eq. (2). The proposed regularized objective function in Eq. (12) is used in all experiments and optimized by the steps in Algorithm 1.

The original *Cameraman* image is warped and blurred by a 7×7 Gaussian with standard deviation of two, then decimated by factor of two in both the horizontal and vertical directions and additive noise with variance σ^2 =2 (corresponding to BSNR=32dB), four LR observations are used to reconstruct the HR image. We first

Algorithm 1. HOS-based Super-Resolution Reconstruction
Data: Given LR images with additive noise.
Result: Reconstructed HR image $\hat{X}$.
Initialize the estimated image by image interpolation such as bicubic interpolation etc.
Subpixel motion estimation.
While ~ stop **do**

Compute the fidelity term $\sum_{k=1}^{K} \left\| D_k H_k F_k \underline{X} - \underline{Y}_k \right\|_2^2$ in Eq.(12);

Do image segmentation to the HR estimation $\underline{\hat{X}}^k$ as described in section 3.1 and get the segmentation label;
Compute the weight function C_w by Eq.(11) and get the regularization term in Eq.(12);
The optimal update by conjugate gradient optimal algorithm is obtained as: $\hat{X}^{k+1} = \arg\min J(\hat{X}^k)$;

If $\frac{J(\hat{X}^{k+1}) - J(\hat{X}^k)}{J(\hat{X}^k)} \le \varepsilon$, stop. Otherwise,

$\hat{X}^k = \hat{X}^{k+1}$.

End

compute the subpixel motion estimation using a two-level Hierarchical Block Matching (HBM) motion estimation algorithm. The initial estimate of HR image is obtained by bicubicly interpolating one of the LR images to the HR grid. In the fist experiment the window size used for computing HOS for each pixel is 3×3, that is, the parameters in Eq.(4) $n = m = 1$; Segmentation number is 20; regularization parameter λ is determined by $\lambda = \sigma^2 / E^2$, where σ^2 is the variance of the noise, E^2 limits high-frequency power of HR image among some power range. The SR results are shown in Fig.1 (f). Some intermediate results such as segmentation label and segmentation result are shown in Fig.1 (c) and (d). For comparison, results of the traditional method described in Eq. (2) are also shown in Fig.1 (e). Evidently, compared to the traditional method, our method significantly removes noise in smooth regions and preserves more details. Meanwhile, the region-based method in [6] is also implemented with the same data, the SR results are comparative to our method, but it takes longer running time. Similarly, Fig. 2 shows the results of *guitar* image. It can be seen that the proposed method is more suitable for images without many thin textures. For example, the smaller characters at the bottom of the *guitar* image are not well recovered while the larger characters at the top are better recovered.

Fig. 3 shows the results of *Monarch* image. The PSNR and MSSIM measurement results are given in Fig. 4. It can be seen that the proposed method has better convergence performance and the image by our method has the higher PSNR and MSSIM.

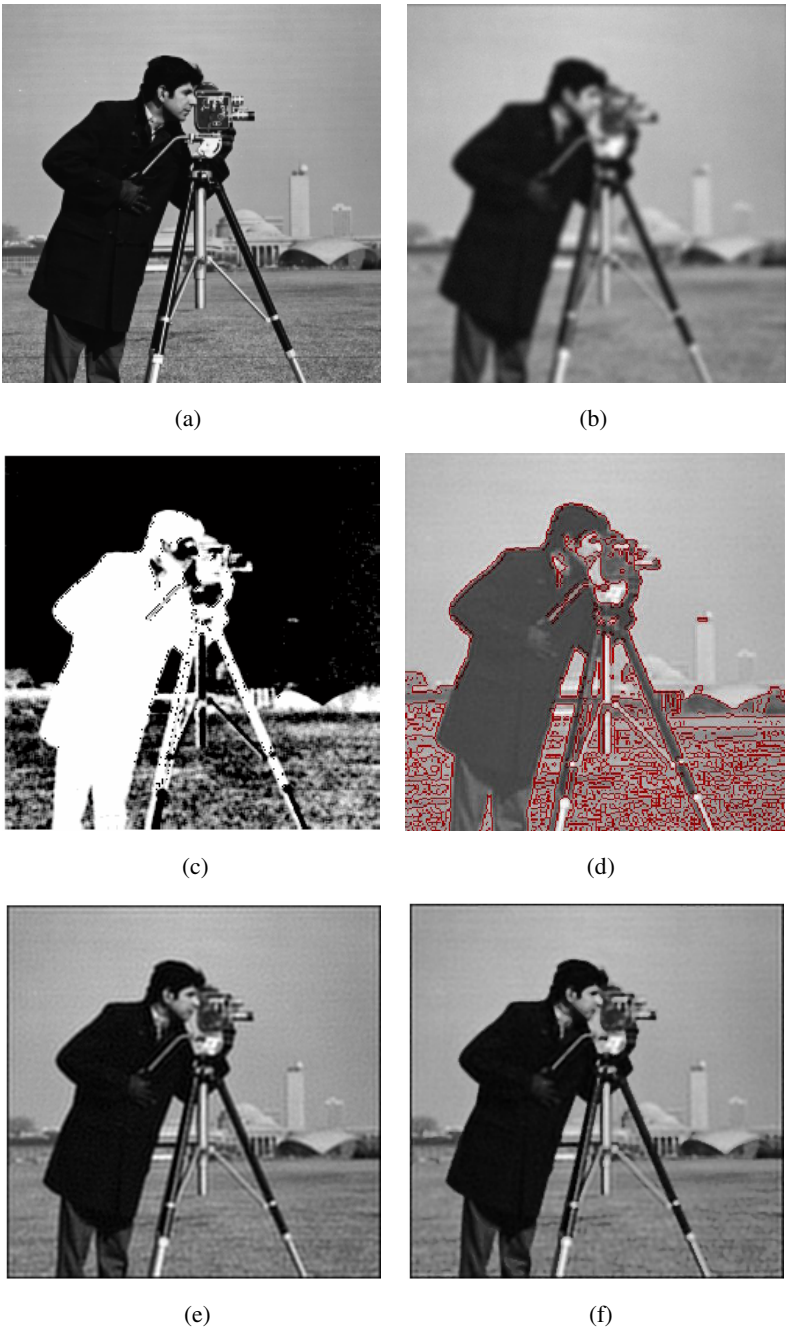


Fig. 1. (a) Original image; (b) bicubic interpolation; (c) segmentation label image in 5th iteration; (d) segmentation results in 5th iteration; (e) traditional regularized algorithm; (f) the final result of proposed algorithm preserves more details and reduces the noise more

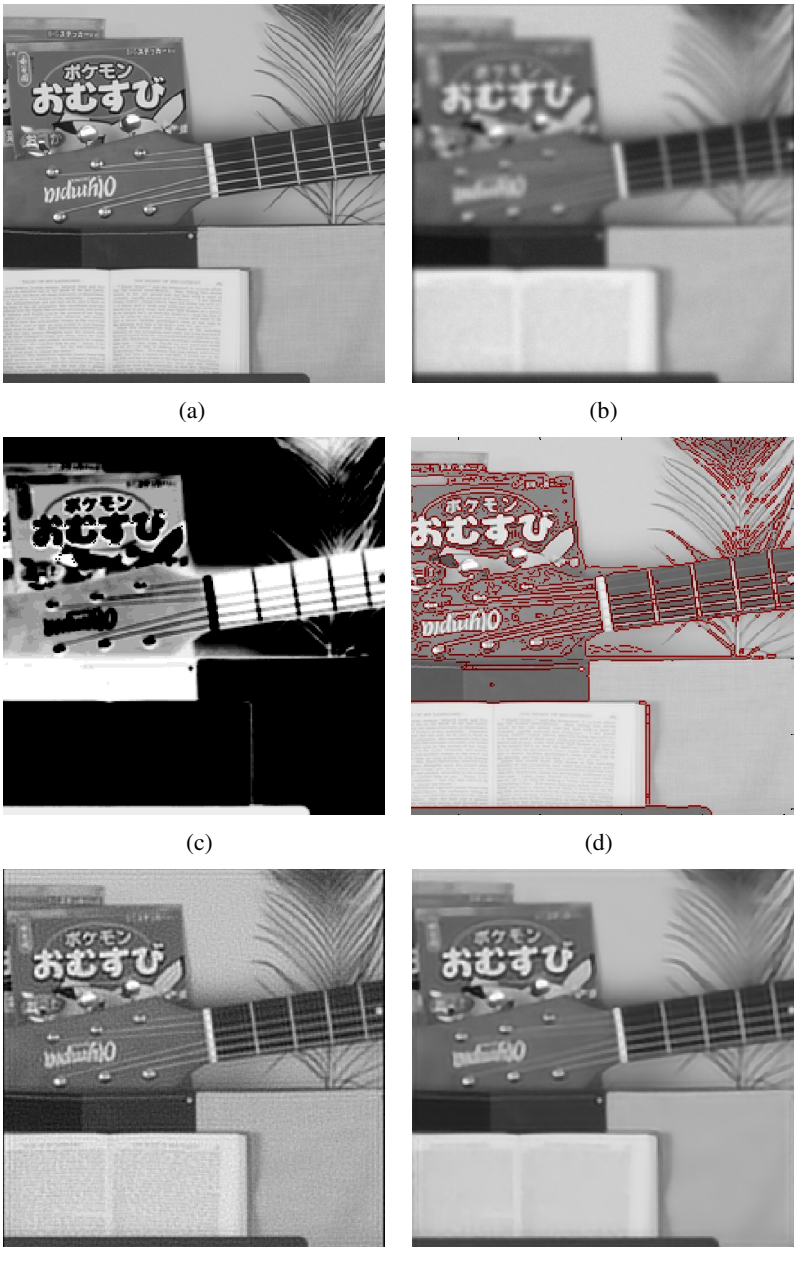
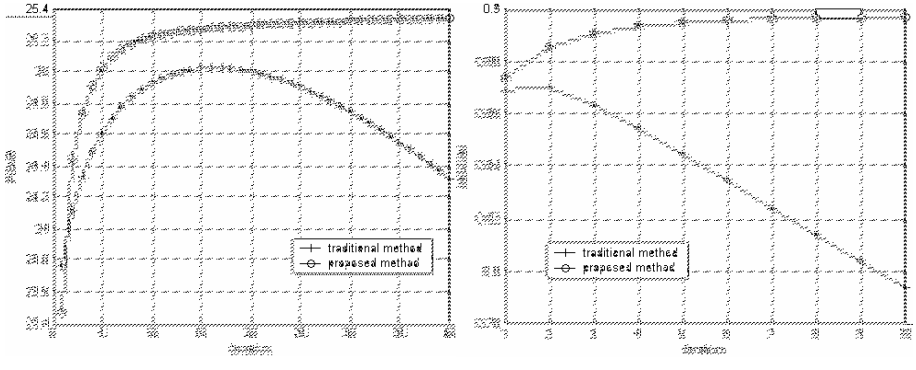


Fig. 2. (a) Original image; (b) bicubic interpolation; (c) segmentation label image in 3th iteration; (d) segmentation results in 3th iteration; (e) traditional regularized algorithm; (f) result of proposed algorithm

(a) Traditional regularized restoration (b) Proposed regularized restoration result

Fig. 3. Example of "Monarch" image



(a) PSNR vs. iteration (b) MSSIM vs. iteration

Fig. 4. Comparison of PSNR and MSSIM between traditional and proposed algorithm

5 Conclusion

In this paper a novel high-order statistics based super-resolution algorithm is presented. It makes use of the HOS to extract the local characteristics and structures. Then image segmentation labels are computed which will be used to decide the weight function in regularization term. By applying different regularization term to different types of regions, the SR results provide sharper edges and significantly reduce noise of flat region. Experimental results demonstrate that the proposed method performs better than traditional methods both in subjective and objective measurement.

Acknowledgement. The work was supported by Program for New Century Excellent Talents in University (NCET-05-0582), Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20050422017), Open Foundation of National Laboratory on Machine Perception, Peking University (0403) Open Foundation of National Laboratory of Networking and Switching, BUPT (07) and the Project-sponsored by SRF for ROCS, SEM ([2005]55).

References

1. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: A technical overview. IEEE Signal Processing Magazine 20, 21–36 (2003)
2. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(9), 1167–1183 (2002)
3. Ni, K.S., Kumar, S., Vasconcelos, N.: Single Image Superresolution Based on Support Vector Regression. In: IEEE International Conference on Acoustics, Speech, and Signal Processing 2, 601–604 (2006)
4. Capel, D., Zisserman, A.: Computer vision applied to super resolution. IEEE Signal Processing Magazine 20, 75–86 (2003)
5. Mignotte, M.: A segmentation-based regularization term for image deconvolution. IEEE Transactions on Image Processing 15(7), 1973–1984 (2006)
6. Choi, B., Ra, J.B.: Region-Based Super-Resolution Using Multiple Blurred and Noisy Undersampled Images. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 609–612 (2006)
7. Wang, Z., Qi, F.: On Ambiguities in Super-Resolution Modeling. IEEE signal processing letters 11(8), 678–681 (2004)
8. Hyvarinen, A., Karhunen, J., Oja, E.: Independent component analysis. John Wiley & Sons, Inc, West Sussex, England (2001)
9. Charbonnier, P., Blanc-Feraud, L., Aubert, G.: Deterministic edge-preserving regularization in computed imaging. IEEE Transactions on Image Processing 6(2), 298–311 (1997)
10. Wang, Z., Bovik, A.C., Sheikh, H.R.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)

Story Unit Segmentation with Friendly Acoustic Perception*

Longchuan Yan[1,3], Jun Du[2], Qingming Huang[3], and Shuqiang Jiang[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
[2] NEC Laboratories China, Beijing, 100084, China
[3] Graduate School of Chinese Academy of Sciences, Beijing, 100049, China
{lchyan, jdu, qmhuang, sqjiang}@jdl.ac.cn

Abstract. Automatic story unit segmentation is an essential technique for content based video retrieval and summarization. A good video story unit has complete content and natural boundary in visual and acoustic perception, respectively. In this paper, a method of acoustic perception friendly story unit segmentation for broadcast soccer video is proposed. The approach combines replay detection, view pattern and non-speech detection to segment story units. Firstly, a replay detection method is implemented to find the highlight events in soccer video. Secondly, based on positions of replay clips, an FSM (Fine State Machine) is used to obtain rough starting points of story units. Finally, audio boundary alignment is employed to locate natural audio boundaries for acoustic perception. The algorithm is tested on several broadcast soccer videos. The story units segmented by algorithms with and without audio alignment are compared in acoustic perception. The experimental results indicate the performance of the proposed algorithm is encouraging and effective.

Keyword: Video Processing, Story Unit Segmentation, Acoustic Perception, SVM.

1 Introduction

Story unit segmentation is to extract the events related video clips from a long video, which is a necessary component for video summarization, retrieval and indexing. The general story unit segmentation is still an open problem due to the limited generalization ability of pattern recognition algorithm and various video patterns. The pre-knowledge combined story unit segmentation algorithm can provide some good solutions in specific fields. As a major branch, the story unit segmentation in sports videos attracts much attention for its wide applications and tremendous commercial potentials. This paper utilizes the soccer video as representative example to illustrate the story unit segmentation issue.

* This work is supported by National High-Tech R&D Program (863 Program) under the grant No. 2006AA01Z117 and NEC Labs, China.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 223 – 231, 2007.
© Springer-Verlag Berlin Heidelberg 2007

In the last decade, many approaches have been proposed for story unit segmentation of various sports video. B. Li et al. employed audio energy, scene cut and grass ratio to locate starting points of live video to extract the story unit for soccer video [1]. In [2-3], the view type and its transition relationship were used to detect the goal and "attack" events for soccer video. N. Noboru et al. used the CC (Closed Caption) text to aid the story segmentation for football video [4]. J. Wang et al. used SVM classifier to detect the event for soccer video taken by main camera and used play position and time duration to determine the event boundaries [5]. C. Liu et al. employed unsupervised scene cluster and their transition matrix to extract the story units in table tennis, diving and archery [6]. X. Tong et al. tried to segment highlight play-units for sports video browsing by some rules [7]. N. Babaguchi et al. used the dominant color and the ratio of the number of vertical lines to that of horizontal lines to link the replay and the live scene [8]. The above algorithms combine domain knowledge and some image or audio features to segment story units for specific video types. They employ the visual content as the story unit segmentation criterion, such as grass ratio, scene cut, view pattern, and etc. From the visual content, the story unit has complete content and natural boundaries. However, the algorithms pay little attention to the characteristics of acoustic perception of story unit. The story unit may have incomplete content or unnatural boundaries in audio stream, which is unfriendly in acoustic perception. The acoustic perception friendly story unit has complete content and natural boundary in audio stream, which give audience better acoustic perception.

In this paper, a method of acoustic perception friendly story unit segmentation is proposed. The algorithm combines production pattern, view pattern, and non-speech detection to locate the highlight events and segment the story units. To obtain good human perception, the algorithm tries to segment the story units with complete content and natural boundaries in image and audio streams, respectively. The story unit used in [1] is employed to represent highlight event, which is composed of the replay and its corresponding live video. The diagram of our algorithm is shown in Fig. 1. Firstly, the algorithm uses the replay detection to locate the highlight events. The replay is a well-known production pattern to enhance highlight events in soccer video, thus the replay detection can locate almost all the highlight events. Secondly, the algorithm determines the starting point of a live video by using view pattern. The

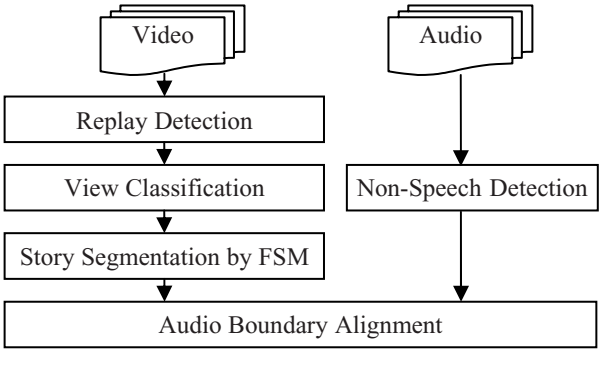


Fig. 1. Block diagram of story unit segmentation algorithm

grass ratio has been used to describe view type and story unit segmentation, but it is still a rough method. To obtain a fine view pattern description, the algorithm employs an SVM classifier to recognize the view type and uses an FSM to describe the view pattern. Lastly, the algorithm adjusts the audio boundaries of story units by selecting the non-speech points as final boundaries. The audio boundary alignments make the story unit more natural by keeping the integrity of speech. Compared with the grass ratio based method, the proposed algorithm is more robust by introducing the fine view classification and view pattern model. Furthermore, the algorithm can extract story units with the complete content and natural boundaries in image and audio streams, which can provide better visual-audio perception to audience.

The rest of the paper is organized as follows. Section 2 introduces story unit and its characteristics of acoustic perception. In Section 3, the view classification and non-speech detection are described, respectively. In Section 4, the proposed story segmentation algorithm is discussed in detail. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

2 Story Unit and Its Characteristics of Acoustic Perception

The story unit used in this paper is defined as a replay and its corresponding live video. For visual content, the story unit makes audience view event clearly because replay often show event in different angles or views and the live video can offer audience to see the whole process of event.

For acoustic perception, announcer's speech in audio stream contains much useful information about a match. The story unit with the complete content and natural boundaries is acoustic perception friendly. The integral content means the audio clip contains enough information about the content of event in story unit. For the story unit, the replay part gives enough time for the announcer to illustrate the event though his/her speech always lags behind the content of image sequence. The natural boundary means the story boundary is on non-speech point. If the segmentation algorithm does not consider the acoustic perception, some story boundaries may be on phonations, such as points A and C in Fig. 2. They are unnatural story boundaries for acoustic perception. The proposed algorithm tries to select non-speech points as story boundaries for good acoustic perception.

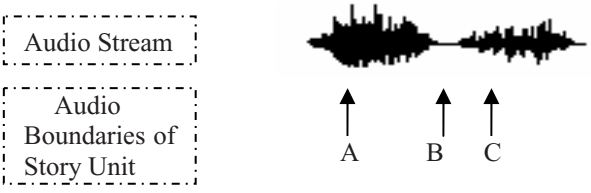


Fig. 2. Audio boundaries of story unit

3 Video and Audio Preprocessing

3.1 Replay Detection

In broadcast soccer video, the replay is often emphasized by adding two logos before and after the replay to attract the attention of audience. Detecting the replay through logo detection is an effective method to locate the highlight events. In this paper, our previous logo based replay detection method in [9] is implemented to detect the replay.

3.2 View Classification

Generally, there are four kinds of typical views in soccer video: namely global view, middle view, close-up view and outside view, as shown in Fig. 3. Here an SVM classifier is used to recognize the view type and local grass ratios are used as view classification features.

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

Fig. 3. Typical views in soccer video (global view (a), middle view (b), close-up view(c) and outside view (d))

3.2.1 Adaptive Playfield Color Detection

The playfield color varies in different lighting or playfield conditions. To calculate the grass ratio correctly, a dominant color detection method is employed to adaptively determine the playfield color of each video. Firstly, the algorithm selects 100 frames randomly from the video and picks up the peak value of the Hue histogram of each frame. Then, the algorithm removes the peak values if they are not in the range to represent the Hue of playfield color. Finally, the average value and its range are calculated over the left peak Hue values to modal the playfield color.

3.2.2 Feature Selection

Grass ratio of image has been used as feature of view classification [2, 10], but the grass ratio does not contain spatial distribution and size information of objects. To

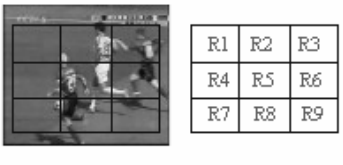


Fig. 4. Zones arrangement of local grass ratios

distinguish the views effectively, nine local grass ratios are introduced to describe the spatial distribution and size of objects. They are calculated by the number of grass pixels divided by that of all the pixels in a zone. The local zones arrangement is shown in Fig. 4.

3.2.3 View Classification

In this paper, the non-playfield images are regarded as the outside views if their global grass ratios are less than a predefined threshold. Other three types of views are recognized by an SVM classifier with RBF kernel function.

3.3 Non-speech Detection

Non-speech detection in an audio stream is a typical two-class recognition problem. We employ an SVM classifier to detect non-speech segments [11]. The feature extraction is a critical step of non-speech detection. The Shot-Time Energy, Zero-Crossing Rate, and 16-order Mel-frequency Cepstral Coefficients (MFCC) are extracted from each audio frame and combined as a feature vector. The length of the audio frame is 256 samples without overlapping. Before feature extraction, the audio signal is converted into a general format, which is 11,025 Hz, 16-bit and mono-channel.

To reduce abrupt misclassification of the SVM classifier, we use a 40-frame average filter to refine the classification results. In addition, we assume the minimal length of the non-speech segments is more than 40 frames (about one second). Otherwise, they are regarded as pauses between words in sentences.

4 Segmentation Method

In this section, we introduce the FSM based starting point detection of live video and audio boundary adjustment for friendly characteristics of acoustic perception.

4.1 Detection of Rough Starting Point of Live Video

In soccer videos, the view type transition has typical pattern in an event. We use an FSM to describe the view pattern to locate the rough starting points of live videos, as shown in Fig. 5. The view type V and view duration T are used to describe the story pattern in the FSM. There are three states in the FSM, namely E, L and B, which correspond to the ending candidate sub-clip, live story sub-clip and beginning candidate sub-clip, respectively.

In some cases, the live story sub-clip is too long relative to the length of replay. In other cases, the short zoom in or zoom out may affect the story unit segmentation. To avoid the two problems, two time constraints in the FSM are introduced. The length of the first shot in replay is used to calculate the time constraint T_c for constraining the search range of live video. It is more accurate than that of replay, because the replay may contain several shots to show an event. In addition, we use another time

constraint T_o to reduce the bad effect of short time zoom in or zoom out. In the proposed algorithm, the two time constraint thresholds in each replay are calculated by

$$T_c = \alpha \times T_Shot \tag{1}$$

$$T_o = \beta \times T_Shot \tag{2}$$

where *T_Shot* is the length of the first shot in a replay and α, β are the scale factors. In our experiments, α and β are set to 4 and 0.2, respectively.

For the detection of the starting points of live videos, view type and its duration are first identified, and then they are put into the FSM from the view type near the replay to the far. The FSM will stop when the starting point of live video is found.

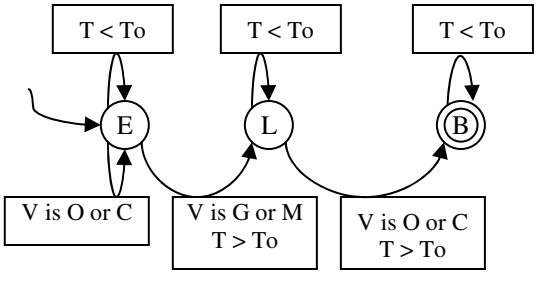


Fig. 5. FSM of view pattern in live video

4.2 Audio Boundary Alignment

The starting points of live videos produced by FSM and the ending points of replays are visual boundaries of story units. To obtain acoustic friendly boundaries of story units, the longest non-speech segment near the visual starting points or ending points is selected as the final story unit boundary. If there are no non-speech points near the boundaries, the visual boundaries are not adjusted any more.

5 Experimentation

For story unit segmentation, we use two broadcast soccer videos of FIFA 2006 to test the algorithm. The videos are compressed in MPEG-1 with 25 fps and frame resolution of 352×288. The middle experimental results on the view classification and the non-speech detection are first given below.

For view classification test, we collect 2,665 global view images, 846 middle view images and 534 close-up view images from different videos. About 1,400 images are used to train the classifier and others to test the classifier. The accuracy of view classification is from 88.6% to 99.0%. From experimental results, the classifier can

distinguish the three view types very well, though classification of the middle view is not as good as those of the other views.

For non-speech detection test, we manually label four audio sub-clips decompressed from soccer videos. The total length of audios is about 2,286 seconds. We select about 93 seconds audio as training set, and others as testing set. The average recall and precision are 79% and 87%, respectively. In the experiments, some transient pauses between words in sentences are the major obstacles for the high accurate non-speech detection, since it is difficult to label them exactly in training data set.

For the story unit segmentation, we use two metrics to evaluate the performance of the proposed algorithm, namely story integrity and natural audio boundary. The story integrity is that a story unit just contains one event. The natural audio boundary is that a story unit without speech abruption at its boundary. Since the two metrics are hard to compute, the subjective evaluation method is adopted. The evaluation result of story segmentation algorithm in story integrity is shown in Table 1. The #Story is the number of the whole stories in a video; the #P is the number of the segmented stories whose length is acceptable. The #S and #L are the numbers of the story units whose length are longer or shorter than their ideal length, respectively. Some key frames of segmented story units are shown in Fig. 6.

Table 1. Story Segmentation Results

Video name	# Story	# P	# S	# L
06_11_03.mpg	30	25	2	3
07_09_03.mpg	35	29	5	1

The numbers of natural audio boundaries story units produced by segmentation algorithms with and without audio boundary alignment are listed in Table 2. It can be seen that the audio boundary alignment can produce more story units with natural audio boundaries.

Table 2. Audio Boundary Evaluation

Video name	# Story	# natural story units with audio alignment	# natural story units without audio alignment
06_11_03.mpg	30	23	16
07_09_03.mpg	35	28	18

From the above two tables and Fig. 6, the algorithm can produce a promising performance. The proposed approach is based on some pattern recognition algorithms. When the view classification and non-speech detection are good, most story units can be extracted correctly. Otherwise, the story unit segmentation and boundary alignment are affected. Additionally, dynamic variation of view pattern in story unit may cause improper segmentations.

Fig. 6. Some Key Frames of Segmented Story Units

6 Conclusion

In this paper, an acoustic perception friendly story unit segmentation algorithm is proposed. The algorithm tries to segment the story units with complete content and natural boundaries in image and audio streams, respectively. For visual content, the view pattern and time constraints are used to segment the story units by an FSM. For audio stream, the audio boundary alignment is employed to avoid the speech abrupt cutting. From experiments, the algorithm can improve the acoustic perception of story units, especially in natural audio boundary. However, there is still much room to improve the story unit model and audio boundary alignment method. In future, we will seek some machine learning methods to model the story unit and take more attention to the audience perception of story unit.

References

1. Li, B., Errico, J.H., Pan, H., Sezan, I.: Bridging the semantic gap in sports video retrieval and summarization. Journal of Visual Communication & Image Representation, 393–424 (2004)
2. Ekin, A., Tekalo, A.M., Mehrotra, R.: Automatic Soccer Video Analysis and Summarization. IEEE transactions on Image Processing, vol. 12(7) (July 2003)
3. Ren, R., Jose, J.M.: Football Video Segmentation Based on Video Production Strategy. In: Proceedings of 27th European Conference on Information Retrieval, March 2005, Sandiago (2005)

4. Noboru, N., Babaguchi, N., Kitahashi, T.: Story Based Presentation for Broadcasted Sports Video and Automatic Story Segmentation. In: Proceedings of 2002 IEEE International Conference on Multimedia and Expo, August 2002, Switzerland (2002)
5. Wang, J., Xu, C., Chng, E., Wan, K., Tian, Q.: Automatic Replay Generation for Soccer Video Broadcasting. In: Proc. of 13th ACM International Conference on Multimedia (2004)
6. Liu, C., Huang, Q., Jiang, S., Zhang, W.: Extracting Story Units in Sports Video Based on Unsupervised Video Scene Clustering. In: Proc. of 2006 IEEE International Conference on Multimedia and Expo, Toronto, Canada, pp. 1613--1616 (2006)
7. Tong, X., u, Q., Zhang, Y., Lu, H.: Highlight Ranking for Sports Video Browsing. In: Proc. of 13th ACM International Conference on Multimedia, Singapore, November 2005 pp 519–522 (2005)
8. Babaguchi, N., Kawai, Y., Yasugi, Y., Kitahashi, T.: Linking Live and Replay Scenes in Broadcasted Sports Video. In: Proc. of ACM Multimedia 2000, Worshop on Multimedia Information Retrieval, Marina del Rey, pp. 205–208 (2000)
9. Zhao, Z., Jiang, S., Huang, Q., Zhu, G.: Highlight Summarization in Sports Video Based on Replay Detection. In: Proc. of 2006 IEEE International Conference on Multimedia and Expo, Toronto, Ontario, Canada, pp. 1613–1616 (2006)
10. Rjondronegoro, D., Chen, C., Pham, B.: Sports Video Summarization using Highlights and Play-Breaks. In: Proc. of 11th ACM International Conference on MIR (2003)
11. Lu, L., Zhang, H., Li, S.: Content-based audio classification and segmentation by using support vector machines. Multimedia Systems, pp. 482–492 (2003)

Decomposition in Hidden Markov Models for Activity Recognition

Weidong Zhang, Feng Chen, Wenli Xu, and Zisheng Cao

Department of Automation, Tsinghua University, Beijing, 100084, P.R.China
zwd03@mails.tsinghua.edu.cn
{chenfeng, xuwl}@mail.tsinghua.edu.cn
czc@mails.tsinghua.edu.cn

Abstract. Dynamic probabilistic networks have been widely used in activity recognition. However, few models are competent for long-term complex activities involving multi-person interactions. Based on the study of activity characteristics, this paper proposes a decomposed hidden Markov model (DHMM) to capture the structures of activity both in time and space. The model combines spatial decomposition and hierarchical abstraction to reduce the complexity of state space as well as the number of parameters greatly, with consequent computational benefits in efficiency and accuracy. Experiments on two-person interactions and individual activities demonstrate that DHMMs are more powerful than Coupled HMMs and Hierarchical HMMs.

1 Introduction

Activity recognition has been one of the most active topics in the high-level computer vision for its huge number of potential applications. As efficient models for learning sequential and temporal characteristics in data sequence, dynamic probabilistic networks, especially hidden Markov models (HMMs), have been widely used in recognition of individual activities and relative simple interacting activities. However, few models have been proposed for modeling long-term complex activities involving interactions of multiple persons performing multiple actions simultaneously, such as group activities lasting a long time.

Real scenes of interest often contain multiple persons and complex interactions between them. There are two kinds of inherent structures in these activities, temporal structure and spatial structure. First, from a perceptual psychology viewpoint [2], activities have hierarchically temporal structure. They can be viewed as compositions of actions having some orderly relations, which is named partonomic hierarchy reflecting "part of" relations between parts and sub-parts. Activities also can be viewed as existing at various levels of abstraction, which is named taxonomic hierarchy reflecting "kind of" relations. Second, activities often consist of multiple dynamic processes, and each of them has its own internal dynamics and is influenced by others, this reflects the spatial structural characteristic. Large dimension and asynchrony in interacting activity recognition make it more difficult. It is beneficial to model both of the structures, however,

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 232–241, 2007.
© Springer-Verlag Berlin Heidelberg 2007

existing models can only reflect one of them directly. One extension to traditional HMMs, such as Semi-HMMs [3] and hierarchical structure HMMs [10,11,12], involves combinatoric number of states to recognize interacting activities while it solves the long-term temporal dependency to some extent. The other extension, such as observation or state coupled HMMs [5,6,7,8], adopts a compositional representation of two or more variables for interacting multiple processes, but they can not represent the hierarchical structure of activities.

In this paper, we present a new method which can model and recognize long-term complex interacting activities. The proposed models, named decomposed hidden Markov models, decompose the states of traditional HMMs in multi-scale time and space. The compositional states bring conceptual advantages of parsimony and clarity, with consequent computational benefits in efficiency and accuracy [7]. In DHMMs, decomposition in space generates multiple coupled channels to model fixed-number multi-person interactions, a relation layer uncouples them for a more compact representation while keeping the causal-temporal influences, and higher multi-level abstractions capture the hierachical temporal structure of interacting activities. DHMMs have both of the advantages of Coupled HMMs (CHMMs) and Hierarchical HMMs (HHMMs), and are equivalent to them in some special cases as demonstrated later in this paper.

The layout of the paper is as follows. In Section 2 we review the related methods for activity recognition. Section 3 describes the proposed models and the corresponding inference and learning algorithms. Experimental results are given in Section 4, followed by the conclusions in Section 5.

2 Related Work

During the past two decades, there were significant researches on human activity recognition based on HMMs. Other methods can be found in a recent survey [1]. HMMs have been widely used in gesture, sign language and individual activity recognition. However, HMMs are inefficient to deal with long-term complex or/and multi-process activities.

To address the long-term limitation, hidden Semi-Markov models (HSMMs) and hierarchical structure HMMs were introduced. In HSMMs [3], each state is assumed to remain unchanged for some random duration of time and to emit a sequence of observations rather than a single one before its transition to a new state. In [4], HSMMs is adopted to analyze the extended time of multiple primitive events detected by Bayesian networks. Variable-Length Markov models [9] provide a more flexible representation of activities than fixed memory Markov models. Unlike HSMMs, Abstract HMMs [10], Hierarchical HMMs [12,15] and Layered HMMs [11] generalize HMMs by introducing a hierarchical structure to capture hierarchical characteristics of activities. HHMMs make each of the hidden states a HHMM on its own, AHMMs closely resemble the HHMMs but assume that internal states only depend on their immediate parents. In LHMMs, the activities are classified level by level using HMMs, and the inference is conducted from low level to high level.

To address multi-channel interactions, another kind of extensions was studied. Factorial HMMs [5] factor the state of HMMs into several distinct state variables evolving independently, and the processes are nominally coupled at the outputs. In Linked HMMs [7], two parallel channels evolving in lockstep are coupled by weights that connect their hidden state nodes. Brand [6] proposed Coupled HMMs (CHMMs) to capture causal-temporal influences between two hands. Dynamically Multi-Linked HMMs [8] provide a more flexible structure which consists of more optimized factorization of the state transition matrices and has less state connections. Observation Decomposed HMMs [14] decompose the original observation into a set of sub-observations to recognize multi-agent activities.

Motivated by these two kinds of extensions, we integrate both of the ideas in our DHMMs to represent the inherent hierarchical structure and causal-temporal relations between processes simultaneously.

3 Decomposed Hidden Markov Models

In this section, we will present the general DHMM structure and the corresponding inference and learning algorithms.

3.1 Model Structure and Analysis

According to the characteristics of activity, decomposing the states of standard HMMs in time and space and decomposing the observations in space generate the proposed decomposed HMMs. Two examples represented as dynamic Bayesian networks are shown in Figure 1. In DHMM, there is a *production* layer which consists of several independent state channels. Above it, DHMM has multiple abstract layers. We specially name the lowest abstract layer *relation* layer which has influences on all the production state channels. Observation is decomposed into a set of sub-observations, and each depends on only one production state channel. The model of Fig.1a has two production state channels and three layers of hierarchical abstraction. The model of Fig.1b has three production state channels and one abstraction layer. The state and observation at kth channel and time t are represented by q_t^k and O_t^k respectively, h_t^k denotes the state at abstract layer k and time t, f_t^d is an indicator variable controlling when the Markov chain at layer k enters an end state and return to the $(k-1)$th[1] layer.

To accommodate multiple channels of data, traditional HMMs have to be formulated with multivariate distribution on the outputs. However, if there are multiple processes generating these channels, interactions and asynchronous manner of these processes will lead the state space growing rapidly to become intractable. DHMMs decompose the state variable of HMMs in space into several coupled variables to capture the spatial structure of activities. Since coupled HMMs [7] are densely connected, we uncouple the influence relations by a relation layer,

[1] For our convenience, we set the index of the lowest layer (relation layer) of abstract layers 1 versus D in standard HHMMs [13].

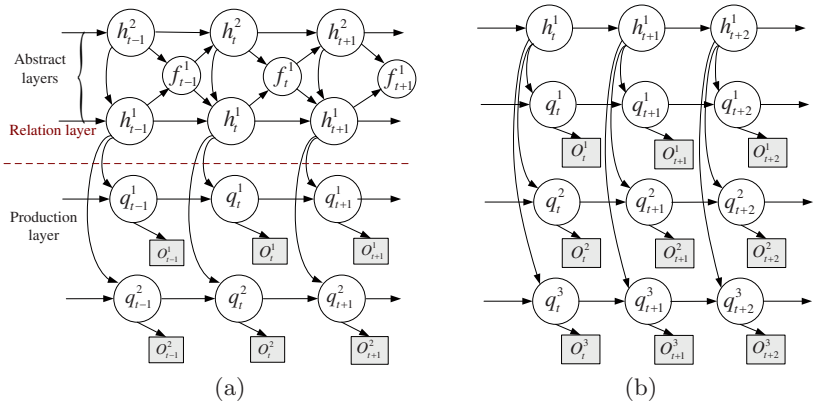


Fig. 1. DBN representation of decomposed hidden Markov model examples

that is, these variables are conditional independent. Observation is also decomposed and each of the sub-observations depends on only the corresponding state channel. The spatial decomposition in DHMMs brings several advantages. First, decomposition in space reduces the complexity of the model and relation uncoupling reduces the number of parameters further. Second, using one or several channels to represent the activities of one person provides more details about the internal dynamics of individual activities. Third, DHMMs avoid the problem of large feature space, feature selection of each channel becomes more flexible. DHMMs can model feature sets at different time scales. Additional duration state on the large-scale feature would make the model more accurate.

Above the production layer, DHMMs are structured multi-level stochastic processes to capture the hierarchical structure of activities. A high-level activity is decomposed into a sequence of simpler activities at low level and in turn, low-level activities are refined into a sequence of lower level activities until primitive activities are reached. There are two meanings corresponding to the taxonomic and partonomic characteristics of activities respectively. First, DHMMs model activities at multiple levels of abstraction to represent the taxonomic characteristic, and each of these layers contains one level activity semantic notion. Second, every state in high level of abstract layers consists of a sequence of low-level states, and each of them corresponds to the notion of a part of the activity, which can be seen as activity component decomposition. Transitions in the same layer indicate the orderly relations of activity parts. Through the multi-level structure, DHMMs present an intuitive representation of the hierarchical structure of activities, and then the long-term temporal dependency between the states is solved by the high-level abstraction. The states of relation layer can denote the relations among persons in multi-person interaction recognition and the relations between the global feature and the local feature of single person [17] in individual activity recognition.

DHMMs are similar, even equivalent, to CHMMs and HHMMs in some special cases, but more powerful than them in majority situation. When individual

activity recognition is concerned, merging the multiple channels of production layer in DHMMs would convert the models to HHMMs. To recognize simple interacting activities between two persons, we can remove abstract layers except the relation layer from DHMMs having two production state channels. From Bayesian networks viewpoint, the retained models have equivalent performance compared with CHMMs since we can recover the coupled relations between the channels provided the probabilities of relation layer states. However, only DHMMs are competent for recognizing long-term interacting activities.

3.2 Parameter Estimation in DHMMs

In a DHMM, the set of parameters, denoted as λ, includes three parts, initial parameters π, transition parameters A and observation distributions B. We assume K_h is the index of the highest abstract layer, K_q is the number of HMM chains at the production layer. N_k is the state number of the kth abstract layer, M_k is the state number of the kth channel in the production layer, L is the layer number including relation layer and production layer.

Initial parameters Π

π_i^{h,K_h}: the probability that the top activity begins at sub-activity $r_i^{K_h}$.

$\pi_{i|l}^{h,k}$: the probability that kth layer activities begin at r_i^k conditioned on the higher activity r_l^{k+1}.

$\pi_{i|l}^{q,k}$: the probability that kth person firstly do activity s_i^k conditioned on relation r_l^1.

State transition parameters A

$$\begin{cases} a_{i,j|l}^{h,k} = P(h_t^k = r_j^k | h_{t-1}^k = r_i^k, h_t^{k+1} = r_l^{k+1}) \\ a_{i,j|l}^{q,k} = P(q_t^k = s_j^k | q_{t-1}^k = s_i^k, h_t^1 = r_l^1) \end{cases} \tag{1}$$

$P(h_t^k = r_j^k | h_{t-1}^k = r_i^k, f_{t-1}^{k-1} = b, f_{t-1}^k = f, h_t^{k+1} = r_l^{k+1})$ and $P(f_t^k = 1 | h_t^k = r_i^k, h_t^{k+1} = r_l^{k+1}, f_t^{k-1} = b)$ are defined as same as [13].

Observation distributions B

$$b_{O_t^k|i} = P(O_t^k | q_t^k = s_i^k) = \sum\nolimits_{c=1}^{C^k} w_{i,c}^k \cdot N(O_t^k, \mu_{i,c}^k, \Sigma_{i,c}^k) \tag{2}$$

where $N(O_t^k, \mu_{i,c}^k, \Sigma_{i,c}^k)$ denotes a Gaussian distribution, C^k is the number of components of mixture Gaussian, $w_{i,c}^k$ is the corresponding weight.

In general, we assume a DHMM has one abstract layer (the relation layer), and two production state channels (same assumption in section 3.3, and it can be easily extended to other cases). We generalize EM learning algorithm to define Q function as $Q(\lambda, \bar{\lambda}) = \sum_{\{h,q\}} P(O, h, q|\lambda) \log P(O, h, q|\bar{\lambda})$ and we get

$$\pi_l^{h,1} = \frac{P(O, h_1^1 = r_l^1 | \lambda)}{P(O|\lambda)} = \sum\nolimits_{l'=1}^{N^1} \gamma_2^{h,1}(l, l') \tag{3}$$

$$\pi_{i|l}^{q,k} = \frac{P(O, q_1^k = s_i^k, h_1^1 = r_l^1|\lambda)}{P(O, h_1^1 = r_l^1|\lambda)} = \frac{\tilde{\gamma}_1^{q,k}(i,l)}{\sum_{l'=1}^{N^1} \gamma_2^{h,1}(l,l')} \tag{4}$$

$$a_{l',l}^{h,1} = \frac{\sum_{t=2}^{T} P(O, h_{t-1}^1 = r_{l'}^1, h_t^1 = r_l^1|\lambda)}{\sum_{t=2}^{T} P(O, h_{t-1}^1 = r_{l'}^1|\lambda)} = \frac{\sum_{t=2}^{T} \gamma_t^{h,1}(l',l)}{\sum_{t=2}^{T} \chi_{t-1}^{h,1}(l')} \tag{5}$$

$$a_{i,j|l}^{q,k} = \frac{\sum_{t=2}^{T} P(O, q_{t-1}^k = s_i^k, q_t^k = s_j^k, h_t^1 = r_l^1|\lambda)}{\sum_{t=2}^{T} P(O, q_{t-1}^k = s_i^k, h_t^1 = r_l^1|\lambda)} = \frac{\sum_{t=2}^{T} \xi_t^{q,k}(i,j,l)}{\sum_{t=2}^{T} \gamma_t^{q,k}(i,l)} \tag{6}$$

$$w_{i,c}^k = \frac{\sum_{t=1}^{T} \chi_t^{q,k}(i,c)}{\sum_{t=1}^{T} \chi_t^{q,k}(i)}, \quad \mu_{i,c}^k = \frac{\sum_{t=1}^{T} \chi_t^{q,k}(i,c) O_t^k}{\sum_{t=1}^{T} \chi_t^{q,k}(i,c)} \tag{7}$$

$$\Sigma_{i,c}^k = \frac{\sum_{t=1}^{T} \chi_t^{q,k}(i,c)(O_t^k - \mu_{i,c}^k)(O_t^k - \mu_{i,c}^k)^T}{\sum_{t=1}^{T} \chi_t^{q,k}(i,c)} \tag{8}$$

where $k = 1, 2$ and the details of the intermediate variables can be found in appendix.

3.3 Calculating the Likelihood

Given parameter set λ, observation sequence $O = \{o_{1:T}^1, o_{1:T}^2\}$, then the likelihood probability is $P(O|\lambda) = \sum_{\{h,q\}} P(O, h, q|\lambda)$. If a special state sequence $h_{1:T}^1$ and $q = \{q_{1:T}^1, q_{1:T}^2\}$ is given, then

$$P(O, h, q|\lambda) = \pi_{h_1^1}^{h,1} \cdot \prod_{t=2}^{T} a_{h_{t-1}^1, h_t^1}^{h,1} \cdot \prod_{k=1}^{2} \left\{ \pi_{q_1^k|h_1^1}^{q,k} b_{O_1^k|q_1^k} \prod_{t=2}^{T} a_{q_{t-1}^k, q_t^k|h_t^1}^{q,k} b_{O_t^k|q_t^k} \right\} \tag{9}$$

Similar to standard HMMs, it is difficult to compute the likelihood directly because of huge number of $\{h_{1:T}^1, q_{1:T}^1, q_{1:T}^2\}$, we generalize the forward algorithm and define forward operators $\alpha_t(i,j,l) = P(O_{1:t}, q_t^1 = s_i^1, q_t^2 = s_j^2, h_t^1 = r_l^1)$, we can solve it inductively as follows:

1) Initialization, $t = 1$:

$$\alpha_1(i,j,l) = \pi_l^{h,1} \pi_{i|l}^{q,1} \pi_{j|l}^{q,2} b_{O_1^1|i} b_{O_1^2|j}, \quad 1 \le i \le M^1, 1 \le j \le M^2, 1 \le l \le N^1 \tag{10}$$

2) Induction, $1 \le t \le T-1$:

$$\alpha_{t+1}(i',j',l') = \left(\sum_l \sum_i \sum_j a_{l,l'}^{h,1} a_{i,i'|l'}^{q,1} a_{j,j'|l'}^{q,2} \alpha_t(i,j,l) \right) b_{O_{t+1}^1|i'} b_{O_{t+1}^2|j'} \tag{11}$$

3) Termination, $t = T$:

$$P(O|\lambda) = \sum_l \sum_i \sum_j \alpha_T(i,j,l) \tag{12}$$

Similarly we can get generalized backward algorithm.

4 Experimental Results

This section presents the experimental results of the proposed method and the discussions on detail designing of activity models. Experiments are conducted on recognizing interacting activities between two persons in parking lot and sidewalk scenes. Two kinds of individual activities on a square are used to demonstrate the flexibility of the model in feature decomposition.

Five two-person interacting activities are selected as test examples:

- Act.1: Two people walk on the same path in the same direction with relatively constant distance.
- Act.2: Two people walk on the same path in the opposite direction.
- Act.3: Two people run on the same path in the opposite direction.
- Act.4: Two people walk in the opposite direction on the same path. They chat with each other when meet, and then go on separately.
- Act.5: Two people approach and meet, one puts an object on the ground and goes away, the other picks up this object and goes away.

The activity dataset contains five activities, each has around 30 and total 146 video clips lasting from 120 to 190 frames. For each activity, about half of its dataset is used for training. Fig.2 shows some key frames of the Act.5. As in our previous work [17], the feature vector $(\mathbf{p}, \mathbf{v}, \gamma, \alpha)$ of each person is selected, where $\mathbf{p} = (p_x, p_y)$ is the position, $\mathbf{v} = (v, \theta)$ is the magnitude and direction of velocity, $\gamma = w/h$ is the aspect ratio and α is the angle of inclination of human body. Here we add the acceleration $\mathbf{a} = (a_x, a_y)$ for its efficiency for representing details. All these features are extracted via our Video Intelligent Surveillance System (VISS) [16].

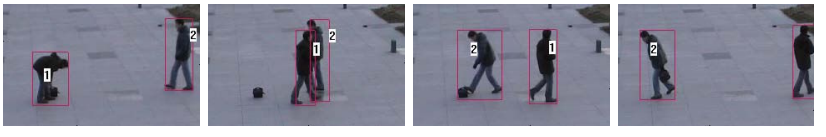


Fig. 2. Key frames of Act.5 (only the interesting region of original images is shown)

Different activities having different characteristics, DHMMs must be designed according to activities' temporal and spatial structures. Since this dataset only includes two-person interactions, the production layer consists two channels, and each corresponds to one person[2]. Generally, the more complex activity is, the more layers of abstraction and more states in each layer DHMM has. Firstly, we divide the interacting activities and individual activities into sequence of primitive activities[3]. The number of the primitive activities of each person defines the

[2] Spatial decomposition can not be totally decided by the person number. Sometimes several person share one channel or one person uses several channels.

[3] This notion cannot be explicitly defined, sometimes it depends on personal experience.

corresponding channel's state number in the production layer, and the number of primitive activities of interactions defines the state number of the relation layer. Secondly, using *bottom-up* strategy, we combine the primitive activities into sub-activities step by step, and each step defines a higher abstraction layer. In our experiments, $L = 2$ for the first four activities and $L = 3$ for Act.5, $(N_1 = M_1 = M_2 = 3)$ for the first three ones, $(N_1 = 3, M_1 = M_2 = 4)$ for Act.4 and $(N_2 = 2, N_1 = M_1 = M_2 = 4)$ for Act.5.

Comparison between the proposed DHMMs and traditional HMMs, Coupled HMMs, and Hierarchical HMMs is conducted with this dataset. When we use HMMs and HHMMs, we merge the feature vectors of two persons into a big one. Table 1 illustrates the recognition rates using DHMMs and the other models. The results show that traditional HMMs can obtain high recognition rate when the structure of activity is relative simple such as Act.1 and Act.2. CHMMs can recognize most of these interactions, HHMMs are inferior to CHMMs when this interacting activities dataset is concerned. Since DHMMs are nearly equivalent to CHMMs when $L = 2$, DHMMs obtain comparable results when activities 1-4 are concerned. However, for Act.5, our method is superior to other methods because of the efficient representation of temporal and spatial structures. The overall high errors at Act.3 are caused by the stronger noises in the observations.

Table 1. Comparison with other three models for interacting activity recognition, the number is recognition rate (%)

	HMMs	CHMMs	HHMMs	DHMMs
Act.1	100	100	100	100
Act.2	84.6	100	92.3	100
Act.3	84.2	84.2	84.2	84.2
Act.4	93.3	100	93.3	100
Act.5	63.6	81.8	81.8	90.9

Table 2. Comparison with HMMs for single person activity recognition, the number is the number of correctly recognized activities

	HMMs	DHMMs(1)	DHMMs(2)	total
Act.6	10	11	11	11
Act.7	10	10	12	12

Nearly any signal produced by human behavior can be beneficially decomposed into a group of interacting processes. We compare our method with HMMs for recognizing individual activities. This dataset includes two activities: (1)Act.6: one person walks in the scene; (2)Act.7: one person walks in the scene, picks up an object from ground and goes on. The same features as in interaction recognition are selected. We set the state number of HMMs 3 and 4 for Act.6

and Act.7 respectively, $L = 2$ for all DHMMs, ($N_1 = M_1 = M_2 = 3$) in DHHM for Act.6 and ($N_1 = 3, M_1 = M_2 = 4$) in DHMM for Act.7. The recognition results and comparison are shown in table 2. We decompose features into($\mathbf{p}, \mathbf{v}, \gamma$) and ($\mathbf{a}, \alpha$) for DHMMs(1) and decompose features into ($\mathbf{p}, \mathbf{v}, \mathbf{a}$) and ($\gamma, \alpha$) for DHMMs(2).

5 Conclusions and Discussions

This paper presents a novel method to recognize human activities, especially multi-person interacting activities with complex structures. We analyze the structural characteristics of activities from temporal and spatial points, and present a decomposition HMM structure to recognize long-term complex interacting activities. DHMMs decrease the dimension of the feature and reduce the complexity of network as well as the number of parameters greatly via space decomposition and relations uncoupling. DHMMs can not only model the interactions between persons but also represent the details of individual activities. However, the semantic abstraction and partonomic segmentation of activities are mainly based on personal experience, which affects the results greatly. This model cannot be used to multi-person activity recognition when the person number is unknown or variable, which will be our future work.

Acknowledgements

This work was supported by United Technologies Research Center (UTRC). The authors would like to thank the video team at UTRC for their pertinent and constructive discussion.

References

1. Moeslund, T.B., Hilton, A., Krger, V.: A survey of advances in vision-based human motion capture and analysis. International Journal on Computer Vision and Image Understanding 104(2-3), 90–127 (2006)
2. Zacks, J.Z., Tversky, B.: Event structure in perception and conception. Psychological Bulletin 127, 3–21 (2001)
3. Murphy, K.P.: Hidden Semi-Markov models (HSMMs).(2002) `http://www.ai.mit.edu/muphyk`
4. Hongeng, S., Nevatia, R.: Large-scale event detection using Semi-hidden Markov models. In: Proc. ICCV, pp. 1455–1462 (2003)
5. Ghahramani, Z.: An introduction to hidden Markov models and bayesian networks. Journal of Pattern Recognition and Artificial Intelligence 15(1), (2001) 9C42.
6. Brand, M., Oliver, N.M., Pentland, A.P.: Coupled hidden Markov models for complex action recognition. In: Proc. CVPR, 994–999 (1997)
7. Oliver, N.M., Rosario, B., Pentland, A.P.: A Bayesian computer vision system for modeling human interactions. IEEE Trans. PAMI 22, 831–843 (2000)
8. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: Proc. ICCV, pp. 742–749 (2003)

9. Galata, A., Johnson, N., Hogg, D.: Learing variable-length markov models of behavior. Comuter Vision and Image Understanding 81, 398–413 (2001)
10. Bui, H.H., Venkatesh, S., West, G.: Policy recognition in the abstract hidden Markov model. Journal of Artificial Intelligence Research 17, 451–499 (2002)
11. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision and Image Understanding 96(2), 163–180 (2004)
12. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden Markov model: analysis and applications. Machine Learning 32(1), 41–62 (1998)
13. Murphy, K.P, Paskin, M.: Linear time inference in hierarchical HMMs. In: Proc. NIPS, pp. 833–840 (2001)
14. Liu, X.H., Chua, C.S.: Multi-agent activity recognition using observation decomposed hidden Markov models. Image and Vision Computing 24, 166–175 (2006)
15. Nguyen, N.T., Phung, D.Q., Venkatesh, S., Bui, H.: Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In: Proc. CVPR, pp. 955–960 (2005)
16. Zhang, W.D, Chen, F., Xu, W.L., Zhang, E.W.: Real-time video intelligent surveillance system. In: Proc. ICME, pp. 1021–1024 (2006)
17. Du, Y.T., Chen, F., Xu, W.L., Li, Y.B.: Recognizing interaction activities using dynamic Bayesian network. In Proc. ICPR, pp. 618–621 (2006)

Appendix

Intermediate variables in parameter estimation:
The definitions of the intermediate variables are

$$\xi_t^{q,k}(i',i,l) = P(q_{t-1}^k = s_{i'}^k, q_t^k = s_i^k, h_t^1 = r_l^1 | O, \lambda) \qquad t = 2, ..., T \quad (13)$$

$$\gamma_t^{h,1}(l',l) = P(h_{t-1}^1 = r_{l'}^1, h_t^1 = r_l^1 | O, \lambda) \qquad t = 2, ..., T \quad (14)$$

$$\gamma_t^{q,k}(i',l) = P(q_{t-1}^k = s_{i'}^k, h_t^1 = r_l^1 | O, \lambda) = \sum\nolimits_i \xi_t^{q,k}(i',i,l), t = 2, ..., T \quad (15)$$

$$\tilde{\gamma}_t^{q,k}(i,l) = P(q_t^k = s_i^k, h_t^1 = r_l^1 | O, \lambda) \qquad t = 1, ..., T \quad (16)$$

$$\chi_t^{q,k}(i) = P(q_t^k = s_i^k | O, \lambda) = \sum\nolimits_l \tilde{\gamma}_t^{q,k}(i,l) \quad (17)$$

$$\chi_t^{h,1}(l) = P(h_t^1 = r_i^1 | O, \lambda) = \sum\nolimits_i \tilde{\gamma}_t^{q,k}(i,l) \quad (18)$$

$$\chi_t^{q,k}(i,c) = P(q_t^k = s_i^k, n_i^k = c | O, \lambda) = \chi_t^{q,k}(i) \frac{w_{i,c}^k b_{O_t^k|i,c}}{b_{O_t^k|i}} \quad t = 1, ..., T \quad (19)$$

where n_i^k is the index of Gaussian component and

$$\xi_t^{q,1}(i',i,l) = \frac{\sum_{l'} \sum_{j'} \sum_j \alpha_{t-1}(i',j',l') a_{l',l}^{h,1} a_{i',i|l}^{q,1} a_{j',j|l}^{q,2} b_{O_t^1|i} b_{O_t^2|j} \beta_t(i,j,l)}{P(O|\lambda)} \quad (20)$$

$$\gamma_t^{h,1}(l',l) = \frac{\sum_{i'} \sum_{j'} \sum_i \sum_j \alpha_{t-1}(i',j',l') a_{l',l}^{h,1} a_{i',i|l}^{q,1} a_{j',j|l}^{q,2} b_{O_t^1|i} b_{O_t^2|j} \beta_t(i,j,l)}{P(O|\lambda)} \quad (21)$$

$$\tilde{\gamma}_t^{q,1}(i,l) = \frac{\sum_j \alpha_t(i,j,l) \beta_t(i,j,l)}{P(O|\lambda)} \quad (22)$$

$\xi_t^{q,2}(j',j,l)$ has similar form as $\xi_t^{q,1}(i',i,l)$, and $\tilde{\gamma}_t^{q,2}(j,l)$ are similar to $\tilde{\gamma}_t^{q,1}(i,l)$.

Image Inpainting with Improved Exemplar-Based Approach

Qiang Chen, Yingxiang Zhang, and Yuncai Liu

Institute of Image Processing and Pattern Recognition
Department of Automation, Shanghai Jiao Tong University
800 Dong Chuan Road, Min Hang District, Shanghai 200240
The People's Republic of China
{chenqiang, frogren2002, whomliu}@sjtu.edu.cn

Abstract. Image Inpainting is widely used to restore a damaged paintings and photographs or remove unwanted objects. Criminisi's exemplar-based inpainting combined "texture synthesis" and PDE-based inpainting that showed great efficiency on removal of large objects. However, this approach has certain weakness such as high time cost, visual inconsistency in some cases. In this paper, we improve the Criminsi's exemplar-based inpainting based on Search Strategy and Color Transfer. This new method is also suitable to all exemplar-based methods and the results show its effectiveness.

Keywords: exemplar-based inpainting, color transfer, search strategy.

1 Introduction

Image Inpainting is a very useful image processing technique. It was first introduced by Bertalmio et al. [1]. Technically, it is used to filling in missing data in a target region of a single or video image. Nowadays, this technique has been used in many areas such as removing unwanted objects in an image, restoring the missing or damaged parts in the paints, even background creation. [2].

In previous work, there are two classes of algorithms to restore a digital image from a big hole or being damaged area (target region). "Texture synthesis" [3] algorithm grows a new image outward from an initial seed by non-parametric sampling. "Inpainting" techniques based on solving PDE [1][4][5] for filling in small image gaps. PDE-based inpainting methods try to use the isophote information to extend linear structure. However, this method actually uses the diffusion techniques; certain artifacts such as blur could be introduced. "Texture synthesis" generally copies image patches or pixels form original image that avoids the blur. But it is unable to keep the linear structure around the boundary of the target region.

Recently, a novel effective algorithm was proposed, Criminisi, etc. [6] use the sampling concept from Efros' approach and take the isophote into account. It gives higher priority to those points on the boundary of target region which contain linear information the most. After this work, Jiying Wu, etc. [7] makes a different view on the priority and proposed CIEI model compared with the Criminisi's AIEI model. However, the essences of these two approaches are the same, they try to keep the linear

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 242–251, 2007.
© Springer-Verlag Berlin Heidelberg 2007

structure and generate natural restoration simultaneously. These kinds of approaches are called as "Exemplar-based" inpainting. [8]

Although Criminisi's approach has achieved great success in many aspects, it has certain drawback. Since its greedy search for the similar patch, the method is very high time-cost. While it can not find the similar patch, it sometimes causes visual inconsistence because of its inability to generate new patch to fill in the target region. Our work is dedicated in solving these problems and we extend this approach to some other applications such as text erasion, image reproduction. Although it is based on the Criminisi's approach, the contribution is suitable to all the exemplar-based inpainting methods.

The remainder of this paper is organized as follows: Section 2 will describe the Criminisi's approach in detail; Section 3 will give deeper analysis about the drawback and give our resolution. Section 4 will extend the improved algorithm to some other application. The comparative experiments between the original and improved algorithm will also be presented in Section 4.

2 Exemplar-Based Inpainting

This Criminisi's Method presents a novel and efficient algorithm that combines the advantages of "texture synthesis" and PDE-based inpainting. The simultaneous propagation of texture and structure information is achieved by a single, efficient algorithm.

The core of this algorithm is an isophote-driven image-sampling process. Fig. 1 illustrates this point. The region to be filled, i.e., the *target* region is indicated by Ω, and its contour is denoted δΩ The contour evolves inward as the algorithm progresses, and so we also refer to it as the "fill front." The *source* region Φ which remains fixed throughout the algorithm, provides samples used in the filling process.

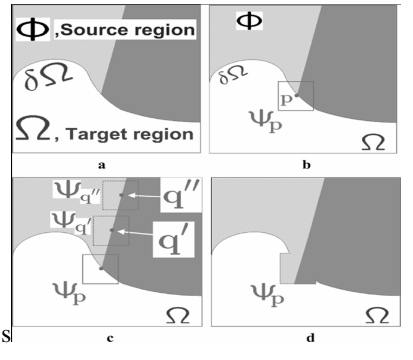


Fig. 1. Structure propagation by exemplar-based synthesis. (a) - (d) shows the process of single iteration of this algorithm.

Here is the description of a single iteration of this algorithm,

Step 1. For each pixel on the boundary, a patch is constructed with that pixel at the center. Each pixel p with a patch has a Patch Priority. It is the product of two elements:

a confidence term C (p), and a data term D (p). They are defined as followed. The confidence term describes how many existing pixels are there in this patch, and the data term describes how strong the isophote hitting the boundary is. The D (p) is especially important since it emphasize the linear structure to be synthesized first and maintain the image structure.

$$P(p) = C(p) * D(p) \tag{1}$$

$$C(p) = \frac{\sum_{q \in \Psi p \cap (I-\Omega)} C(q)}{|\Psi p|}, \quad D(p) = \frac{\left|\nabla I_p^{\perp} \cdot n_p\right|}{\alpha} \tag{2}$$

Step 2. After finding the prior point p, the algorithm searches the source image to find a patch that is the most similar to the target patch. Sum of Square Difference (SSD) is used to calculate the over all color distance between the target patch Ψp and the source patch Ψq. Finding the most similar patch toΨp, formally :

$$\Psi_{q'} = \arg \min_{\Psi_q \in \Phi} d(\Psi_p \Psi_q) \tag{3}$$

Step 3. When the most similar patch Ψq was found, the patch will be copied to the target patch Ψp to fill in the empty area. After filling the patch, the boundary list should be updated. This is a single iteration of the Criminisi's method. Eventually, the whole target region will be entirely filled, and the boundary would be empty.

The Criminisi's method is effective to fill in the big hole due to its isophote-driven process. But it has some flaws. First, the time cost of this example-based algorithm is very high because of the facts as followed, the original search in the Step 2 is a global one, it searches the entire source image to find the most similar patch. The SSD used to calculate the color distance is high-cost. Second, if there is no good corresponding patch to fill the hole, the algorithm finds patch with the lowest SSD value which sometimes causes visual inconsistency. For example, if we use the Criminisi's method to reproduce an image, if the outline color gradual changes and there are no corresponding patch to reproduce, chances are that we can see the intersected outline ambiguously. (Fig.2.)

Fig. 2. The left image is the original image. We use the Criminisi's approach to reproduce this image about 16 rows leftward. From the result (the right image), we can see an obvious intersected line.

3 Improvement of Criminisi's Approach

From the analysis above, we know that the high time-cost and the visual inconsistence are the main problems of the Criminisi's approach. Here, we implement search strategy to reduce the time cost while maintaining its effectiveness and color transfer to solve the visual inconsistence while increasing its performance.

3.1 Search Strategy

The most cost of these exemplar-base methods is spent on the procedure of searching most similar patch. And the original search is a global one that it can find the most similar patch to the target patch. But to most images, the similar patch is just around the target patch. So it provides huge space to raise the algorithm's effectiveness. Besides, the use of SSD as criterion to evaluate the similarity of two patches can be replaced by other color distance measurement which could reduce the time cost while remaining the function as criterion.

To reduce the cost in the searching procedure, we use a concept – Window. For each target patch, we set a window size manually denoted by a parameter, *WindowLength*. In the original Criminisi's algorithm, there is also a parameter set manually, *PatchSize*. Here we set

$$WindowLeng\ th = StepLenght\ *\ PatchSize\ .$$

The search range is restricted to a certain area which is centered by the current prior boundary point and is extended *WindowLength* rows upwardly, downwardly, leftward, rightward. So the actual window size is

$$(2 * WindowLeng\ th + 1) * (2 * WindowLeng\ th + 1)\ .$$

This improvement makes the search being the local one which can reduce the cost significantly. The cost decreases from N (n*m) to N(window size).(n, m is the rows and columns of the image). The effectiveness of this improvement is determined by the manual set of *StepLength*. *StepLength* denotes how many times of *PatchSize* the *WindowLength* should be. It is often easy for people to find the proper *StepLength*.

Besides, we can use L1 Euclidean distance norm instead of SSD (L2 Euclidean distance norm) to evaluate the similarity. L1 Euclidean distance norm can be described as follows, and the experiment shows that it does not reduce the effectiveness of finding similar patch while reduce the time cost.

$$LED = \sum \left| d_s - d_t \right| \tag{4}$$

3.2 Color Transfer

In the original Criminisi's approach, once the most similar patch Ψq is selected, the patch can not be changed and copies its pixel directly to the target patch Ψp. Although the similar patch is the best patch in the source image, it always causes some unnatural

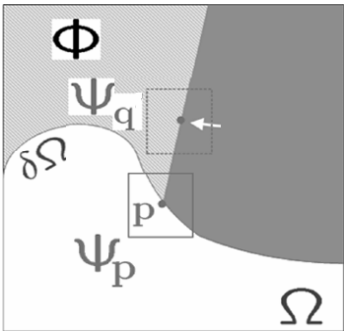


Fig. 3. Color transfer between patch Ψp and patch Ψq

discontinuity especially in the boundary points. When the target region has anomalous outline, sometimes people can not find this discontinuity, while the outline is a regular one, this drawback has been magnified by human visual system. Here we use a simple statistical analysis to transfer the similar patch according to the target patch Ψp.

We use $l*\alpha*\beta$ color space to transfer the image because of its most suitable view on human visual system. Firstly, we compute the mean and standard deviations for both the target patch Ψp (the non-white area of Ψp) and the selected patch Ψq. We subtract the mean from the selected data points:

$$\begin{aligned} l^* &= l - \bar{l} \\ \alpha^* &= \alpha - \bar{\alpha} \\ \beta^* &= \beta - \bar{\beta} \end{aligned} \tag{5}$$

Then we scale the selected these data points by the respective standard deviations and next, instead of adding the averages that we previously subtracted, we add the averages computed for the target patch. Finally, we get the transferred image patch $\Psi q'$. This part of method is called "Color Transfer". [9]. It maintains mostly the color information of surrounding target area.

$$\begin{aligned} l' &= \frac{\sigma_t^l}{\sigma_s^l} l^* + \bar{l}_t \\ \alpha' &= \frac{\sigma_t^l}{\sigma_s^l} \alpha^* + \bar{\alpha}_t \\ \beta' &= \frac{\sigma_t^l}{\sigma_s^l} \beta^* + \bar{\beta}_t \end{aligned} \tag{6}$$

4 Experiments

In this section, we extend our approaches to some applications. Also we compare our improved exemplar-based inpainting with the Criminisi's approach on the time cost and time inconsistency.

4.1 Some Applications

The exemplar-based inpainting has great applications on many areas besides the traditional ones. Image Reproduce uses the improved exemplar-based method to grow the image upward, downward, leftward or rightward while keep the linear structure of

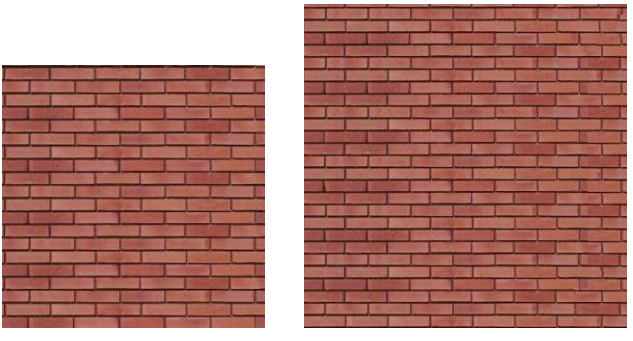


Fig. 4. The left image is the original texture. The right one is the result of image reproduce.

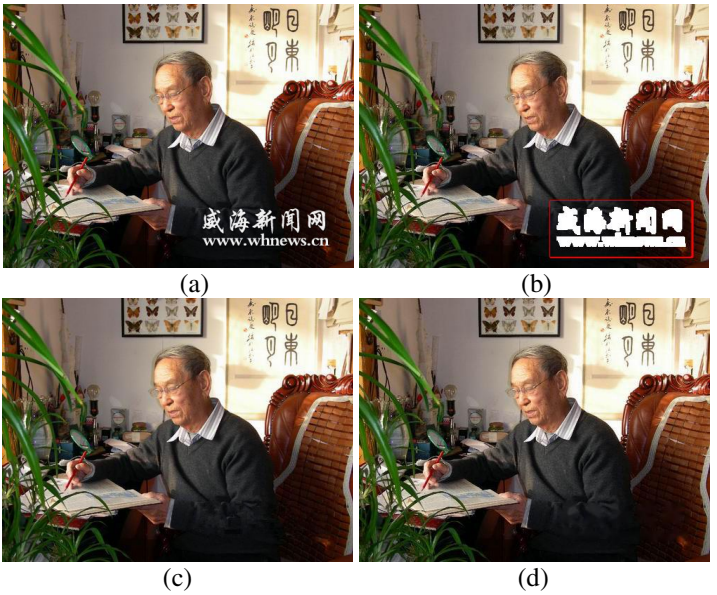


Fig. 5. (a) is original image; (b) is image with white mask. (c) is the result using improved exemplar-based inpainting, (d) is result using N-S inpainting.

the surrounding area. We also have made some experiments on image reproducing using our improved exemplar-based inpainting.

Another application of this inpainting should be "text erasion". The first step of text erasing is text detection. We suppose the text has been selected automatically and use the inpainting method to restore the text area. The effect of text erasing depends on the effect of text detection. If the detection is good, both the Criminisi's method and the N-S inpainting [4] can get a good effect. (Fig 8)

4.2 On Cost Comparison

As analysis above, we can reduce the time cost from N (n*m) to N(window size) using the improved exemplar-based method. Here are some experiments:

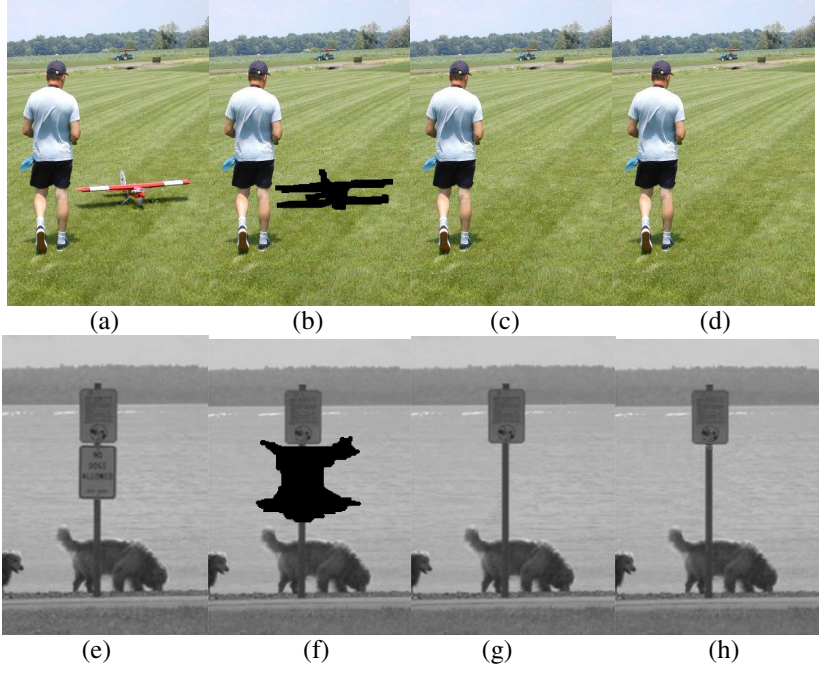


Fig. 6. (a), (e) are the original image; (b), (f) are the image with black mask; (c), (g) are the result of the original Criminisi's approach; d, h are the result of improved one.

Table 1. On the comparison of time cost

	Image Size	Parameter	Time Cost	Image
Criminisi	360*480	PatchSize = 13	68.06s	(c)
Improved	360*480	PatchSize=13, StepLenght=10	6.90s	(d)
Criminisi	190*282	PatchSize = 7	12.4840	(g)
Improved	190*282	PatchSize=7, StepLength=10	3.7190	(h)

From the data above, we can confidently draw the conclusion that the search strategy reduces the time cost of exemplar-based inpainting significantly. The first set of experiment reduces the time cost from 68.06s to 6.90s, almost only 10% left because of its high revolution and local search. The second set of experiment is from [5], we compare our result with Criminisi's result. The effectiveness is obvious and we can not tell one from these two result image. We also find the use of L1 Euclidean distance norm can reduce the time cost by 10-20%. The result shows great efficiency on image inpainting while maintaining the high performance.

4.3 On Visual Inconsistency

We use color transfer to eliminate the visual inconsistency, here are some results.

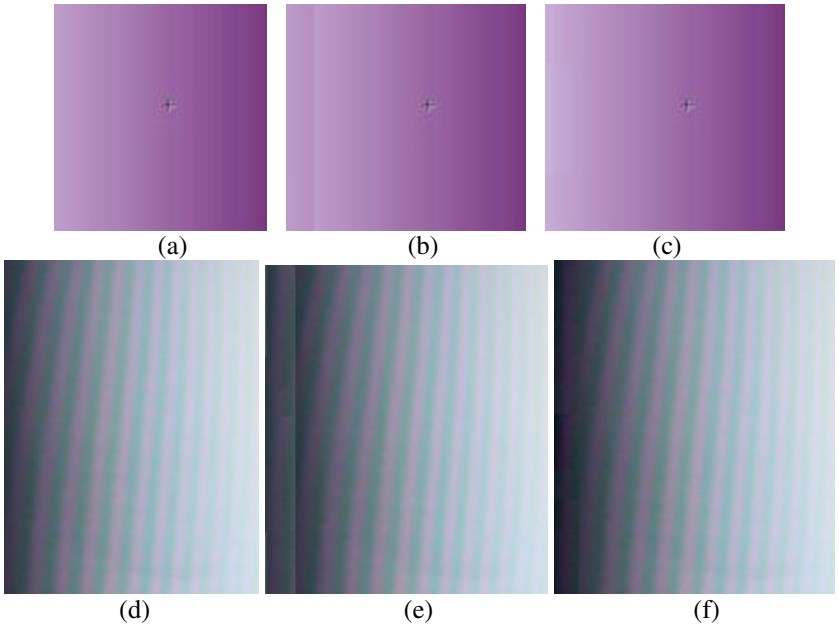


Fig. 7. Two sets of gradual color image. (a), (d) are the original image. (b), (e) are the results of Criminisi's reproducing 16 rows leftward. (c), (f) are the results of our improved one.

Image c, f is visually better than image b, e. Before copying the patch to the target patch in the iteration, the improved approach transfers the similar patch according to the statistical analysis of target patch. The results indicate the color transfer technique makes the patch more similar to the target patch. The "image reproduce" and gradual color maybe seems to be extreme, but these results are helpful to understand the effect of color transfer. Fig 6 shows that this technique is suitable to the image restoration. We can ambiguously see a round outline around the target region from c and the result using color transfer is much better. The introduction of color transfer changes the stable concept that the exemplar-based method is not effective when there is no good exemplar.

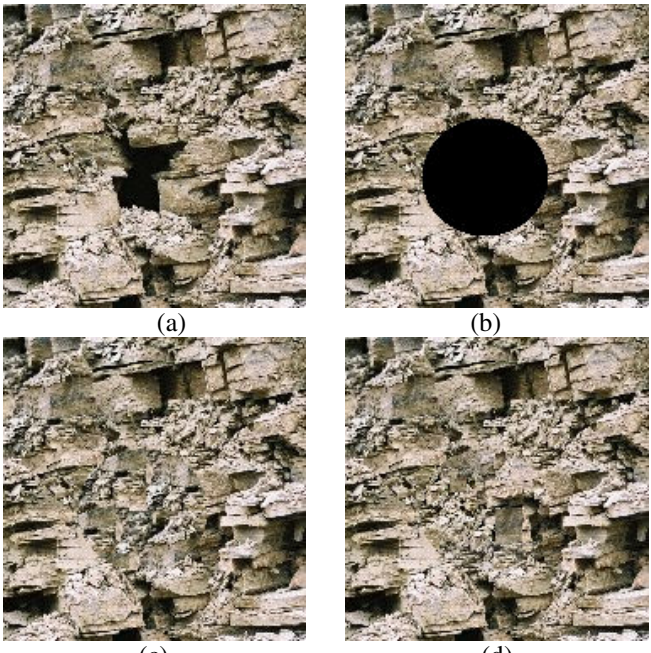


Fig. 8. (a) is the original image; (b) is the image with black mask. (c) is the result of Criminisi's method; (d) is the result using color transfer.

5 Conclusion

In this paper, we first introduce an effective inpainting approach, Criminisi's exemplar-based inpainting. Then we analyze the weakness of this method. Search strategy and color transfer is proposed to enhance the efficiency and effectiveness of the exemplar-based method. The introduction of color transfer breaks a new path to dealing with non-exemplar.

In the future, more concentration should be laid down on the structure transfer of the exemplar and it would be great if this exemplar-based inpainting can be used to maintain the curvilinear structure.

Acknowledgements. This research is supported by NSFC 60675017 and 973 Program 2006CB303103

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image Inpainting. In: SIGGRAPH 2000, Computer Graphics Proceedings (2000)
2. Chung, Y.-Y., et al.: Animating pictures with stochastic motion textures. In: SIGGRAPH 2005 (2005)

3. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: Proc. ICCV, pp. 1033–1038, Kerkyra, Greece (September 1999)
4. Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: Proc. Conf. Comp. Visio, Pattern Rec, pp. 355–362 (December 2001)
5. Chan, T.F., Shen, J.: Nontexture inpainting by curvature-driven diffusions. Jouranl of Visual Communication and Image Representation 12, 436–449 (2001)
6. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based inpainting. IEEE Transactions on Image Processing,(9) 1200–1212(2004)
7. Wu, J., Ruan, Q.: Object Removal by cross isophotes exemplar-based inpainting. In: the 18th International Conference on Pattern Recognition (ICPR'06) (2006)
8. Komodakis, N., Tziritas, G.: Image completion using global optimization. IEEE Comuter Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (2006)
9. Reinhard, E., Ashikhmin, M., et al.: Color transfer between images. In: IEEE Computer Graphics and Applications (2001)

Fingerprinting Codes for Live Pay-Television Broadcast Via Internet

Shuhui Hou[1], Tetsutaro Uehara[2], Yoshitaka Morimura[1], and Michihiko Minoh[2]

[1] Dept.of Intelligence Science and Technology, Graduate School of Informatices, Kyoto University, Japan
[2] Academic Center for Computing and Media Studies, Kyoto University, Japan

Abstract. In recent years, with the rapid growth of the Internet as well as the increasing demand for broadband services, live pay-television broadcasting via internet has become a promising business. To get this implemented, it is necessary to protect distributed contents from illegal copying and redistributing after they are accessed. Fingerprinting system is a useful tool for it. This paper shows that the anti-collusion code has advantages over other existing fingerprinting codes in terms of efficiency and effectivity for live pay-television broadcasting. Next, this paper presents how to achieve efficient and effective anti-collusion codes based on affine plane and unital, which are two known examples of balanced incomplete block design (BIBD). Meanwhile, performance evaluations of anti-collusion codes generated from affine plane and unital are conducted. Their practical explicit constructions are given last.

Keywords: fingerprinting code, anti-collusion code, BIBD, unital, affine plane.

1 Introduction

In the last few years, businesses offering digital contents (such as music and video) via internet have come to be established with the development of broadband networks. But it is not realized to serve live pay-television broadcasts via internet to a number of people simultaneously at the moment. The authors are developing a system to realize such internet-based pay-TV system under the assumption that the same datagram could be transmitted to a number of receivers simultaneously via IP-multicasting.

Broadcast encryption schemes are techniques that allow a center to transmit encrypted data to a large set of receivers so that only a particular subset of privileged receivers can decrypt it. Such schemes are useful to limit the access of digital contents but can not limit the illegal copying and redistributing of decrypted contents. Fingerprinting schemes are helpful techniques to protect distributed contents from illegal copying and redistributing after they are accessed.

Fingerprinting schemes embed unique user information (e.g., ID or fingerprints) into each user's copy as a watermark and if an illegal copy appears, user

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 252–261, 2007.
© Springer-Verlag Berlin Heidelberg 2007

information can be extracted to help trace or identify illegal users. The goal of digital fingerprinting is to deter or discourage people from illegally redistributing the digital data that they have legally purchased. The major challenge to fingerprinting is collusion attacks from illegal users. A collusion attack is a cost-effective attack, where colluders (illegal users) combine several copies with the same content but different fingerprints to try to remove the original fingerprints or frame innocent users.

The research on digital fingerprinting can be broadly divided into two main directions: designing fingerprinting codes and jointly considering the fingerprint coding and embedding. The former direction mainly focuses on the coding theories, in which different emphasis is placed on criteria such as attack assumptions, collusion size, codeword size, codeword length, ability to trace one or all pirates, etc ([1],[2],[3],[4],[5],[8]). In contrast, the latter mainly focuses on fingerprint embedding, as well as advanced detection ([7],[9]).

Fingerprinting codes for internet-based pay-TV system need to manage a large number of users (e.g., more than 10,000 users) from a business and economic point of view. To discourage users from illegal copying, high resistance against collusion attacks is required. It is desirable to shorten the code length (to squeeze more users into fewer bits) for easily implementing code. These three aspects should be considered synthetically. So far, there have been almost no evaluations for what fingerprinting codes adapt to internet-based pay-TV broadcasting.

In this paper, efficiency and effectivity are introduced for evaluating fingerprinting codes. In terms of efficiency and effectivity, anti-collusion code[9] is shown to be superior to other existing fingerprinting codes. Then, it is presented how to achieve efficient and effective anti-collusion codes by using unital and affine plane, which are two known examples of balanced incomplete block design (BIBD). Meanwhile, performances of the anti-collusion codes derived from unital and affine plane are evaluated from the above-mentioned three aspects. Last, fingerprinting codes for internet-based pay-TV system are given out.

2 Efficiency and Effectivity of Fingerprinting Codes

To evaluate what fingerprinting codes are appropriate for pay-TV broadcasting, code efficiency and code effectivity are defined as follows.

For a given fingerprinting code, let its length be l, its maximum tolerated collusion size be c and the number of users that it can support be n.

- The *efficiency* on a fingerprinting code for a given collusion size is referred to as $\beta = n/l$, which describes the number of users that can be supported by code length. Fingerprinting code with a higher β can support more users under the same collusion size.
- The *effectivity* on a fingerprinting code is defined by $\gamma = c/n$, which describes the resistance against collusion attacks. Fingerprinting code with higher γ is resistant to a larger size collusion.

The major existing fingerprinting codes are examined in terms of *efficiency* and *effectivity* in section 3.

3 Anti-collusion Code is Superior to Other Fingerprinting Codes

3.1 Attacking Assumptions: Marking Assuption and *Envelope*

An unrestricted, arbitrary attack is difficult to cope with, so restrictive assumptions need to be made concerning what collusion can be allowed and what collusion cannot be allowed, given a set of codewords. The most widely used assumption is called Marking Assumption[2].

1. Marking Assuption
   A fingerprint is regarded as a collection of marks. Marking Assumption states that: By colluding, users can detect a specific mark if it differs between their copies; otherwise a mark can be undetected.
   According to whether a detectable mark can be allowed to change into an unreadable state or not, and the range within which pirated codes can fall, there are four cases described by *envelope*[5].
2. *Envelope* under Marking Assumption
   'Under Marking Assumption' means that colluders can only alter those marks where at least two users' fingerprints differ in, namely, detectable marks.

   Consider a code C of length l over an alphabet Q where $|Q| = q$. $C \subseteq Q^l$ and is called $(l, n, q) - code$ when $|C| = n$. The elements of C are called codewords and each codeword has the form $x^j = (x_1^j, x_2^j, \cdots, x_l^j)$, where $x_i^j \in Q, 1 \leq i \leq l, j = 1, \cdots, n$. The length of C is l and the size of C is n. Codeword $x^j (j = 1, \cdots, n)$ is assigned uniquely to j-th user, so code size n is also the number of users.

   Let $U(C)$ denote the set of undetectable mark positions for C and symbol $*$, the unreadable or erased state. The range which the pirated codes out of C is defined as *envelope* as follows:
   - narrow-sense envelope $e(C)$ is the set

     $$e(C) = \{y \in Q^l | y_i \in \{x_i^1, \cdots, x_i^n\}, 1 \leq i \leq l\}$$

     Under the narrow-sense case, colluders are only allowed to choose each detectable mark from their corresponding marks to generate pirated codes.
   - expanded narrow-sense envelope $e^*(C)$ is the set

     $$e^*(C) = \{y \in \{Q \cup *\}^l | y_i \in \{x_i^1, \cdots, x_i^n\}, 1 \leq i \leq l\}$$

     Under the expanded narrow-sense case, the restriction is relaxed. Hence, the detectable marks are allowed to change into an unreadable state.
   - wide-sense envelope $E(C)$ is the set

     $$E(C) = \{y \in Q^l | y|_{U(C)} = x^1|_{U(C)} = \cdots = x^n|_{U(C)}\}$$

     The wide-sense case gives a larger range for colluders to generate pirated code. Where, detectable marks can be chosen from alphabet Q.

– expanded wide-sense envelope $E^*(C)$ is the set

$$E^*(C) = \{\{Q \cup *\}^l | y|_{U(C)} = x^1|_{U(C)} = \cdots = x^n|_{U(C)}\}$$

Based on $E(C)$, the expanded wide-sense case allows the detectable marks to be changed into an unreadable state.

It is clear that $e(C) \subseteq e^*(C) \subseteq E(C) \subseteq E^*(C)$. In addition, few fingerprinting codes are defined on the expanded narrow-sense *envelope* $e^*(C)$.

3.2 Major Existing Fingerprinting Codes

For the brevity of description, the abbreviations used in this paper are listed below.

FP: frameproof code; SFP: secure frameproof code;
IPP: identifiable parent property code; TA: traceability code;
SPC: separating code; SEC: secure code with $\epsilon-$error;
ACC: anti-collusion code.

Several surveys on existing fingerprinting codes are conducted. Table 1 shows a taxonomy according to the attack assumptions mentioned previously. Where, 'TA $\Rightarrow$ IPP $\Rightarrow$ SFP $\Rightarrow$ FP' means that a $c-$TA code is a $c-$IPP code, a $c-$IPP code is a $c-$SFP code and a $c-$SFP code is a $c-$FP code. Note that generally the converse is not true.

TA code is stronger than FP, SFP, and IPP codes in terms of tracing ability, i.e., TA code is more effective than FP, SFP, and IPP codes. However, it is not efficient and effective code for live pay-television broadcasting in comparison with ACC. The main reason lies in three aspects:

1. ACC is More Desirable to Protect Multimedia Data than TA
   – ACC traces or identifies colluders by utilizing the fact that the common undetectable mark positions among any group of colluders are unique. The detectable marks can be allowed to change into any state or unreadable state, which provides the resistance to linear attacks or nonlinear attacks for multimedia data.

Table 1. Taxonomy: Research on Fingerprinting Codes Under Marking Assumption

	narrow-sense	wide-sense	expanded wide-sense
code designing	In [1], q-ary FP, SFP, IPP and TA code are defined, and their relationship is investigated, where TA $\Rightarrow$ IPP $\Rightarrow$ SFP $\Rightarrow$ FP; $q-$ary SPC is also studied in [5], which is equivalent to SFP.	In [3], binary FP and SFP code are defined, and their relationship is SFP $\Rightarrow$ FP.	In [2], binary FP and SEC are investigated. A secure code with $\epsilon-$error restricted in narrow-sense is an IPP code if $\epsilon = 0$.
joint coding-embedding	In [7], $q-$ary TA and IPP are defined and that TA codes generally offer better collusion resistance than IPP codes is demonstrated.		In [9], binary ACC is defined, but no explicit construction of efficient and effective ACC.

For multimedia data, components of the fingerprinting sequence are spread over the whole object rather than linked to a small area in the object. Different bits of fingerprinting code that are additively embedded in multimedia may not be easily identifiable and arbitrarily manipulated. Thus, from colluders' perspective, collusion attacks like averaging, interleaving, cut and paste and so on, are easy to operate instead of combining their fingerprinting bits.

- TA is defined under narrow-sense case and the detectable marks can only be changed by combinatorial methods, where, the pirated code stems from combination of colluders' fingerprinting bits. TA traces or identifies one colluder based on the fact that the hamming distance between the pirated code and the colluders codewords is smaller than others not in collusion. Therefore, TA only adapts to combinatorial attacks, which is far from the requirement for protecting multimedia data.

2. ACC is More Effective than TA
   ACC identifies all colluders, while TA only traces one colluder.
3. ACC is More Efficient than TA
   ACC is shorter in code length and larger in code size than TA in a certain sense. Compare the construction of $B(M,q)$[4] with the construction of incidence matrix of BIBD[9]. It is easy to see that bit-complement of incidence matrix of $(M, M-q+1, 1)-$BIBD has the same form as $B(M,q)$, that is, both of them only include 0/1 and the number of '1' is constant $q-1$. However their sizes are different:
   - $B(M,q)$ is the 0/1 matrix consisting of each possible column created by placing $q-1$ '1' and $M-q+1$ '0' among the M rows, which is $M \times L$ matrix, where
   $$L = \begin{pmatrix} M \\ q-1 \end{pmatrix}$$
   - The codematrix C_m which is bit-complement of incidence matrix of $(M, M-q+1, 1)-$BIBD, is also 0/1 matrix consisting of $q-1$ '1' and $M-q+1$ '0' in each column, but it is $M \times L^*$ matrix, where
   $$L^* = M(M-1)/(M-q+1)(M-q)$$

   $L \geq L^*$ holds because L is the total number of possible columns and L^* is the number of part colunms. C_m is submatrix of $B(M,q)$ (They agree at the number of rows but differ in the number of columns).

   In [4], the codewords of $q-$ary TA code are corresponding to the rows of $B(M,q)$, i.e., each row is corresponding to a user's fingerprints. So the code length is L and the number of users that can be supported is M.

   While, the codewords of ACC are corresponding to the columns of C_m, i.e., each column is corresponding to a user's fingerprints. So the code length is M and the number of users that can be supported is L^*, where $L^* \geq M$ is also known by Fisher's inequality[10].

   Based on the above analyses, $L \geq L^* \geq M$ holds. The *efficiency* on ACC is $\beta_1 = L^*/M \geq 1$ and the *efficiency* on TA which is constructed from $B(M,q)$[4] is $\beta_2 = M/L \leq 1$. It is clear that ACC is more efficient than TA.

Since ACC is more efficient and more effective than TA and can be adapted to multimedia data, the authors begin to consider how to generate ACC code.

4 How to Generate Efficient and Effective ACC Codes

4.1 Status Quo of ACC

In [9], an anti-collusion code is defined as follows:

Definition 1. *Let* $G = \{0, 1\}$. *A code* $\mathcal{C} = \{c_1, \cdots, c_n\}$ *of vectors belonging to* G^v *is called a* $K-$ *resilient AND anti-collusion code (AND-ACC) when any subset of* K *or fewer codevectors combined element-wise under AND is distinct from the element-wise AND of any other subset of* K *or fewer codevectors.*

A construction method of ACC is using BIBD.

Definition 2. *A* (v, k, λ) *balanced incomplete block design (BIBD) is a pair* $(\mathcal{X}, \mathcal{A})$, *where* $\mathcal{A}$ *is a collection of* k*-element subsets (blocks) of a* v*-element set* $\mathcal{X}$, *such that each pair of elements of* $\mathcal{X}$ *occur together in exactly* λ *blocks.*

A (v, k, λ)-BIBD has a total of $n = \lambda(v^2 - v)/(k^2 - k)$ blocks. Corresponding to it, there is the $v \times n$ incidence matrix $M = (m_{ij})$ defined by

$$m_{ij} = \begin{cases} 1, \textit{if the ith element belongs to the jth block}, \\ 0, \textit{otherwise}. \end{cases}$$

There exists conclusion:

Theorem 1. *Let* $(\mathcal{X}, \mathcal{A})$ *be a* $(v, k, 1)$*-BIBD and* M *the corresponding incidence matrix. If the codevectors are assigned as the bit complement of the columns of* M, *then the resulting scheme is a* $(k-1)$*-resilient AND-ACC.*

In [9], only two ACC codes from $(7, 3, 1)$-BIBD and $(16, 4, 1)$-BIBD, no practical construction are given. The significant advantages of ACC inspire the authors to generate efficient and effective ACC codes.

4.2 About BIBD

Low-density parity-check codes (LDPC) are constructed based on unital design in [12], which motivates the authors to generate ACC from known examples of BIBD, which is listed in Table 2. Hadamard design is not considered because ACC can be achieved only when $\lambda = 1$ (Theorem 1). Neither Fano plane nor Steiner triple system is under consideration since their parameters are either constant or small. Projective plane is excluded due to its symmetry ($v = n$), whatever the parameter m is, the *efficiency* $\beta = 1$ since the code length is equal to code size. Affine plane and unital show better *efficiency* and better *effectivity*, their performances on live pay-TV broadcasting are investigated below.

4.3 Unital and Affine Plane

Given a $(k-1)$-resilient ACC code (derived from $(v, k, 1)$-BIBD), let the number of users that can be supported be n. Then, the *efficiency* on such an ACC is $\beta = n/v$ and the *effectivity* on such an ACC is $\gamma = (k-1)/n$.

Affine plane and unital are investigated in terms of *efficiency* and *effectivity* (see Table 3): Unital has higher *efficiency* but lower *effectivity*, and affine plane has higher *effectivity* but lower *efficiency*. In other words, unital supports more users but exhibits weaker resistance, while affine plane exhibits stronger resistance but supports fewer users. Performances of affine plane and unital are also evaluated through the following concrete examples.

Assume that unital code (ACC code derived from unital) and affine plane code (ACC code derived from affine plane) support the same amount of users, for example, 10^4 users, their code length, collusion size are shown in Table 4 (a). Affine plane code exhibits very high resistance where its maximum tolerated collusion size is about 100. It is difficult to bring together 100 illegal users.

Assuming the collusion sizes of affine plane code and unital code as 101 (Table 4 (b)), unital code can support about 10^8 users. It is amazing ability to live multicasting.

Similarly, assume that affine plane code and unital code have the same implementing facility, for example, their code length are fixed to 10^4 bits (Table 4 (c)). Affine plane code performs better with respect to collusion resistance and unital code performs better with respect to the amount of users.

For live pay-television broadcasting system, unital code is preferable when higher *efficiency* is required and affine plane is preferable when higher *effectivity* is required.

5 Constructions of Unital Code and Affine Plane Code

5.1 Constructing Unital Code

The results about the existence of unital are: Unital is constructed from finite projective plane ($(m^2 + m + 1, m + 1, 1)$-BIBD) and the only known projective planes have orders that are prime powers (order $m = q^n$).

This paper adapts hermitian unital to construct ACC codes because it has deterministic construction.

Table 2. Existing Examples of BIBD

Block Design	(v, k, λ)	
affine plane	$(m^2, m, 1)$	$v = m^2, k = m, \lambda = 1$
Fano plane	$(7, 3, 1)$	$v = 7, k = 3, \lambda = 1$
Hadamard design	$(4m + 3, 2m + 1, m)$	$v = 4m + 3, k = 2m + 1, \lambda = m$
projective plane	$(m^2+m+1, m+1, 1)$	$v = m^2 + m + 1, k = m + 1, \lambda = 1$
Steiner triple system	$(v, 3, 1)$	$v \equiv 1$ *or* $3 (mod\ 6), k = 3, \lambda = 1$
unital	$(q^3 + 1, q + 1, 1)$	$v = q^3 + 1, k = q + 1, \lambda = 1$

Definition 3. *A hermitian unital in a projective plane of order $n = q^2$ is set of U of $q^3 + 1$ points that meets every line $q + 1$ points.*

The constructing procedure is as follows:

1. Constructing finite projective plane $PG(2, q^2)$

   (a) Define the points as follows:

   - Consider triples $X = (x, y, z)$ of elements from the finite field $GF(q^2)$, where (x, y, z) are not all zero. There exist $(q^2)^3 - 1$ triples.
   - Identify triples X and Y if $Y = \lambda X$ for some non-zero $\lambda \in GF(q^2)$, and say that X and Y are equivalent. Denote the equivalence class of X by $[X]$. Each equivalence class has $q^2 - 1$ members, corresponding to the $q^2 - 1$ possible non-zero values of λ, and so there are $[(q^2)^3 - 1]/(q^2 - 1) = (q^2)^2 + q^2 + 1$ different classes $[X]$, which is taken as the point of $PG(2, q^2)$.

   (b) Define the lines (or blocks) as follows:
   If $a = (a_0, a_1, a_2)$ is a triple of elements of $GF(q^2)$, not all zero, define the line $[\alpha]$ to be the set of all points $[X]$ such that $a_0x + a_1y + a_2z = 0$. There are $(q^2)^2 + q^2 + 1$ lines by an argument similar to the case for points in (a). On each line, there are $q^2 + 1$ points.
2. Constructing hermitian unital
   The hermitian unital is the set of points $(x, y, z) \in PG(2, q^2)$ satisfying $x^{q+1} + y^{q+1} + z^{q+1} = 0$.
   There are $q^3 + 1$ points in hermitian unital, which meets every line(described in (b)) in $q+1$ points. Hermitian unital can be denoted by pair $(\mathcal{H}, \mathcal{B})$, where $\mathcal{H}$ is the set of $q^3 + 1$ points and $\mathcal{B}$ is the collection of subsets consisting of $q + 1$ intersection points (referring to Definition 2).
3. Constructing unital code
   The hermitian unital $(\mathcal{H}, \mathcal{B})$ is a $(q^3 + 1, q + 1, 1)$-BIBD. Compute its corresponding incidence matrix M, take the bit complement of the columns of M as codewords and assign them to users. Thus, a q-resilient AND-ACC is achieved according to Theorem 1.
4. Implementing construction of unital code
   The above-mentioned hermitian unital is implemented by Magma[11]. Following procedure 3., unital codes of $q < 13$ can be obtained. In particular, when $q = 11$, an ACC code with code length 1332 , code size 13431 (i.e., the number of users is 13431), and the maximum tolerated collusion size 10 has been achieved.

Table 3. Evaluations of Affine Plane and Unital Under $k = m = q + 1$

Block Design $(v, k, 1)$ $n = v(v-1)/k(k-1)$	*Efficiency* $\beta = n/v$	*effectivity* $\gamma = (k-1)/n$
affine plane $(m^2, m, 1)$	$1 + 1/(q+1)$	$q/(q+1)(q+2)$
unital $(q^3 + 1, q + 1, 1)$	$q - 1 + 1/(q+1)$	$1/q(q^2 - q + 1)$

Table 4. Performance Evaluations of Affine Plane and Unital Code

Case	Block Design	Number of Users n	Code Length l	Collusion Size k
(a) under the same number of users	affine plane $(m^2, m, 1)$	$n = m(m+1)$ $n : \ 10^4$	$l = m^2$ $l : \sim 10^4$	$k = m - 1$ $k : \sim 100$
	unital $(q^3+1, q+1, 1)$	$n = q^2(q^2-q+1)$ $n : \ 10^4$	$l = q^3 + 1$ $l : \sim 10^3$	$k = q$ $k : \sim 10$
(b) under the same collusion size	affine plane $(m^2, m, 1)$	$n = m(m+1)$ $n : \sim 10^4$	$l = m^2$ $l : \sim 10^4$	$k = m - 1$ $k : \ 101$
	unital $(q^3+1, q+1, 1)$	$n = q^2(q^2-q+1)$ $n : \sim 10^8$	$l = q^3 + 1$ $l : \sim 10^6$	$k = q$ $k : \ 101$
(c) under the same code length	affine plane $(m^2, m, 1)$	$n = m(m+1)$ $n : \sim 10^4$	$l = m^2$ $l : \ 10^4$	$k = m - 1$ $k : \sim 100$
	unital $(q^3+1, q+1, 1)$	$n = q^2(q^2-q+1)$ $n : \sim 2 \times 10^5$	$l = q^3 + 1$ $l : \ 10^4$	$k = q$ $k : \sim 20$

5.2 Constructing Affine Plane Code

There exists an affine plane $(m^2, m, 1)$-BIBD of order m if and only if there exists a projective plane $(m^2 + m + 1, m + 1, 1)$-BIBD of order m. A finite projective plane $(m^2 + m + 1, m + 1, 1)$-BIBD exists when the order m is a prime power (i.e., $m = q^n$). Affine plane code is constructed as below:

1. Construct a finite projective plane $PG(2, q^2)$ by Magma[11].
2. Construct a finite affine plane $AG(2, q^2)$ by deleting a single line b_i of $PG(2, q^2)$ and all of the points of it, and altering the other lines to delete the point which is intersection of this line and line b_i.
3. Compute the corresponding incidence matrix M of $AG(2, q^2)$, then the bit complement of the columns of M are corresponding to users' codewords.

Affine plane code can be obtained when $q < 13$. Especially, when $q = 11$, an ACC code with code length 14641 , code size 14762 (i.e., be able to support 14762 users), and the maximum tolerated collusion size 120 can be achieved.

By using Magma, affine plane $AG(2, q^2)$ and its incidence matrix can be directly achieved when $q \leq 7$. For example, when $q = 7$, the source code is:

```
q:=7;
m:=q^2;
FAP:=FiniteAffinePlane(m);
IM:=IncidenceMatrix(FAP);
```

The main issue of the anti-collusion code is a practical construction problem. The authors have proposed their practical explicit constructions by using unital and affine plane, which is desirable from users' point of view.

6 Conclusions

We aim to achieve fingerprinting codes for internet-based pay-TV system. By conducting several surveys on fingerprinting codes, this paper has shown that

ACC code has advantages over other existing fingerprinting codes in terms of efficiency and effectivity. Then, it is presented how to generate efficient and effective ACC codes by using unital and affine plane. Last, fingerprinting codes derived from unital and affine plane have been given out. We consider these codes are applicable to internet-based pay-TV system. As a future work, we intend to implement the achieved ACC codes on internet-based pay-TV broadcasting and to examine their validity further.

References

1. Staddon, J.N., Stinson, D.R., Wei, R.: Combinatorial Properties of Frameproof and Traceability Codes. IEEE Transactions on Information Theory, vol.47(3) (2001)
2. Boneh, D., Shaw, J.: Collusion-Secure Fingerprinting for Digital Data, IEEE Transactions on Information Theory, 44(5) (1998)
3. Stinson, D.R., van Trung, T., Wei, R.: Secure Frameproof Codes, Key Distribution Patterns, Group Testing Algorithms and Related Structures. Journal of Statistical Planning and Inference 86 (2000)
4. Tina, L., Jacob, L., Mattias, S.: A Class of Traceability Codes, IEEE Transactions on Information Theory, 48(7) (2002)
5. Barg, A., Blakley, G.R., Kabatiansky, G.A.: Digital Fingerprinting Codes: Problem Statements, Constructions, Identification of Traitors. IEEE Transactions on Information Theory, vol. 49(4) (2003)
6. Wang, Z.J., Wu, M., Zhao, H., Trappe, W., Liu, K.J.R.: Anti-collusion forensics of multimedia fingerprinting using orthogonal modulation. IEEE Tran. Image Process 14(6), 804–821 (2005)
7. He, S., Wu, M.: Performance Study on Multimedia Fingerprinting Employing Traceability Codes. In: Barni, M., Cox, I., Kalker, T., Kim, H.J. (eds.) IWDW 2005. LNCS, vol. 3710, Springer, Heidelberg (2005)
8. Safavi-Naini, R., Wang, Y.: Collusion Secure q-ary Fingerprinting for Perceptual Content, Security and Privacy in Digital Rights Management (2002)
9. Trappe, W., Wu, M., Jane Wang, Z., Ray Liu, K.J.: Anti-collusion Fingerprinting for Multimedia. IEEE Transactions on Signal Processing 51(4) (2003)
10. Dinitz, J.H., Stinson, D.R.: Contemporary Design Theory: A Collection of Surveys. John Wiley and Sons, West Sussex, England (1992)
11. Key, J.D.: Some applications of Magma in designs and codes: Oval designs, Hermitian unitals and generalized Reed-Muller codes. J. Symbolic Computation 31(1/2), 37–53 (2001)
12. Johnson, S.J., Weller, S.R.: High-Rate LDPC Codes from Unital Design. In: Global Telecommunications Conference, In: GLOBECOM'03. IEEE (2003)

A New Type of Proxy Ring Signature Scheme with Revocable Anonymity and No Info Leaked

Chengyu Hu[1], Pengtao Liu[2], and Daxing Li[1]

[1] Institute of Network Security, Shandong University, P.R.China
hcy@mail.sdu.edu.cn, lidaxing@vip.sina.com
[2] Shandong Political-Legality Institute, P.R.China
lpt@mail.sdu.edu.cn

Abstract. In some real situations, we must apply proxy signature and ring signature both concurrently. In this paper, we present a new type of proxy ring signature scheme with revocable anonymity which allows the original signer to know exactly who the signer is and unlike other schemes, the original signer doesn't need to publish extra information except his original public key. It can play an important role in some real applications.

1 Introduction

The concept of proxy signature was first introduced by Mambo, Usuda, and Okamoto in 1996 [1]. The proxy signature schemes allow proxy signers to sign messages on behalf of an original signer. The concept of ring signatures was formalized in 2001 by Rivest, Shamir, and Tauman [2]. A ring signature is considered to be a simplified group signature which consists of only users without managers. It protects the anonymity of a signer since the verifier knows that the signature comes from a member of a ring, but doesn't know exactly who the signer is. And the proxy signatures can combine other special signatures to obtain some new types of proxy signatures[3,4,5,6].

2 Our Contribution

In this paper, we propose a new type of proxy ring signature with anonymity revocation and no information leaked, it plays an important role in the following scenario: In a company, the leader(original signer) delegates his signing capability to some proxies, called proxy signers set such as a public known directorate. These proxy signers can sign messages on behalf of the original signer while providing anonymity and convincing the verifier that the signer is in the directorate. At the same time, the leader can always reveal the identity of the signer. We can also call it "group ring signature" since it has a manager who can reveal the identity of the signer. And unlike other schemes, the original signer doesn't need to publish extra information except his original public key.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 262–266, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The rest of this paper is organized as follows. In Section 3, we describe some related works. In Section 4, we present our proxy ring signature scheme with revocable anonymity and no information leaked and analyze its correctness and security in Section 5.

3 Related Works

3.1 Ring Signature

A ring signature must satisfy the usual correctness and unforgeability property[7]. A fairly generated ring signature must be accepted as valid with higher probability; and it must be infeasible for any other user to generate a valid ring signature with the ring he does not belong to. Signature must be anonymous, so that no verifier should be able to guess the actual signer's identity. This notion was first formally introduced by Rivest, Shamir, and Tauman [2], and since then, this topic has been studied extensively, for instance[9,10,11,12].

3.2 Proxy Ring Signature

Some proxy ring signature schemes have been proposed in the last two years. But most of them[7,8] are from bilinear pairings and leak some extra information to the verifier in the delegating process. Meanwhile, they have no property of anonymity revocation.

4 Our Proxy Ring Signature Scheme

In this section, we describe our protocol.

4.1 The System Parameters

For the convenience of describing our work, we define the parameters as follows.

- p,q: two large prime numbers, $q|p-1$.
- g: an element of Z_p^*, its order is q.
- x_o, $x_1, \cdots, x_n$: the original signer A_o's secret key and the proxy signers U_i's secret key.
- $y_o \equiv g^{x_o} \ mod\ p$: A_o's public key.
- $y_i \equiv g^{x_i} \ mod\ p$: U_i's public key.
- $H(\cdot)$: a hash function, $H(\cdot) : \{0,1\}^* \rightarrow Z_q$.
- $H'(\cdot)$: a hash function, $H'(\cdot) : \{0,1\}^* \rightarrow Z_q$.

4.2 The Proxy Phase

- Commission Generation. For the user U_i, the original signer A_o randomly chooses $k_i \in_R Z_q$ and computes $\bar{s_i} = x_o g^{k_i} + k_i$, $\bar{r_i} = g^{k_i}$. Then A_o sends $\bar{s_i}, \bar{r_i}$ secretly to U_i, and keeps k_i private.
- Proxy verification. Each U_i checks $g^{\bar{s_i}} = y_o^{g^{k_i}} g^{k_i} = y_o^{\bar{r_i}} \bar{r_i}$. Then U_i computes $s_i = x_i + \bar{s_i}$ as his secret proxy signature key.

Note that in this phase, no extra information is leaked to the verifiers.

4.3 Signing Phase

Let player U_i be the real signer, the ring $B = (U_1, \cdots, U_n)$. On input a group size $n \in Z$, a message m, a public key set $y_N = (y_1, \cdots, y_n)$. The signer U_i does the following:

1. Select $d \in_R Z_q$, compute $h = H(m)$, $\sigma_i = h^{s_i - d}$. Let $A = \sigma_i^{\frac{1}{i}}$.
2. Select $\omega_i \in_R Z_q$, compute $a_i = g^{\omega_i}$, $b_i = h^{\omega_i}$.
3. For all $j \neq i$, pick up randomly $z_j, c_j, r_j \in_R Z_q$ and $r_j \neq y_o^{-1}$, and compute $a_j = g^{z_j} y_j^{c_j} y_o^{c_j} r_j^{c_j}$, $\sigma_j = A^j$, $b_j = h^{z_j} \sigma_j^{c_j}$.
4. Let $a_N = (a_1, \cdots, a_n)$, $b_N = (b_1, \cdots, b_n)$. Compute $V = g^{g^{-d}}$ and $c = H'(m, a_N, b_N, V)$.
5. Compute $c_i = c - \Sigma_{j \neq i} c_j$, $z_i = \omega_i - c_i s_i + c_i d$.
6. Compute $r_i = y_o^{(\bar{r_i} - 1)} \bar{r_i} g^{-d}$.

Let $z_N = (z_1, \cdots, z_n)$, $c_N = (c_1, \cdots, c_n)$, $r_N = (r_1, \cdots, r_n)$. Then the proxy ring signature is $\sigma = (m, A, z_N, c_N, r_N, V)$.

4.4 Verification

To verify signature $\sigma = (m, A, z_N, c_N, r_N, V)$ on message m, the verifier does the following:

1. Compute $h = H(m)$.
2. Compute $\sigma_i = A^i$, $a_i = g^{z_i} y_i^{c_i} y_o^{c_i} r_i^{c_i}$, $b_i = h^{z_i} \sigma_i^{c_i}$ for all i.
3. if $r_i = y_o^{-1}$, reject the signature.
4. Check $H'(m, a_N, b_N, V) = \sum_{i \in B} c_i$. If the check passed, then the verifier accepts σ as a valid proxy ring signature of m.

4.5 Open

To open the signature and reveal the identity of the signer, the original signer can check the equation

$$g^{r_i} = V^{y_o^{(\bar{r_i} - 1)} \bar{r_i}} \quad for\ all\ i \tag{1}$$

If for i, the check is passed, then U_i is the signer.

5 Analysis

5.1 Correctness

We prove the correctness of verification process as follow:

Proof. For the signer i, $\sigma_i = A^i = (\sigma_i^{\frac{1}{i}})^i$,

$$
\begin{aligned}
& g^{z_i} y_i^{c_i} y_o^{c_i} r_i^{c_i} \\
&= g^{\omega_i - c_i s_i + c_i d} y_i^{c_i} y_o^{c_i} r_i^{c_i} \\
&= g^{\omega_i} g^{-c_i s_i} g^{c_i d} g^{c_i x_i} g^{c_i x_o} g^{c_i (x_o g^{k_i} + k_i)} g^{-c_i x_o} g^{-c_i d} \\
&= g^{\omega_i} g^{-c_i s_i} g^{c_i s_i} g^{c_i d} g^{-c_i d} g^{c_i x_o} g^{-c_i x_o} \\
&= g^{\omega_i} = a_i.
\end{aligned}
$$

$$h^{z_i} \sigma_i^{c_i} = h^{\omega_i - c_i s_i + c_i d} \sigma_i^{c_i} = h^{\omega_i - c_i s_i + c_i d} h^{c_i s_i} h^{-c_i d} = b_i.$$

For all $j \neq i$, $g^{z_j} y_j^{c_j} y_o^{c_j} r_j^{c_j} = a_j$, $h^{z_j} \sigma_j^{c_j} = b_j$.

So, if the proxy ring signature is generated by a valid member in the ring, then the check $H'(m, a_N, b_N, V) = \sum_{i \in B} c_i$ is passed.

To open the signature, since the original signer has $\bar{r_i}$, so he can easily check the equation (1).

5.2 Security Analysis

Theorem 1. *The scheme proposed satisfies the unforgeability property.*

Let B be the ring. We prove that an adversary $A \notin$ the ring B can not forge a valid ring signature.

Proof. Suppose A can repeat the proxy phase and query the original signer for $\bar{s_k}, \bar{r_k}$ where $k \notin B$ to get some proxy signature key which doesn't belong the ring. To forge a ring signature, A executes the signing protocol. In step 3, A selects randomly $z_i, c_i, r_i \in_R Z_q$ for all $i \in B$ and computes $c = H'(m, a_N, b_N, V)$. Since z_i, c_i, r_i are taken randomly from Z_q, this implies that the probability of $c = \sum_{i \in B} c_i$ to be true is $\frac{1}{q}$. So we can say that the proposed scheme is non-forgeable.

Meanwhile, the original signer who doesn't has the proxy signature key of proxy signers in the ring can not forge a valid proxy ring signature either.

Theorem 2. *The scheme proposed is anonymous.*

Proof. For a signature $\sigma = (m, A, z_N, c_N, r_N, V)$ generated by user U_i, c_k, z_k, r_k $(k \neq i)$ is selected randomly, and c is the output of the random oracle H'. In addition, for any fixed message m and fixed set of public keys $y_1, \cdots, y_n$, we can see that $(z_1, \cdots, z_n)$ has exactly q^n possible solutions. Since the distribution of these possible solutions are independent and uniformly distributed, an adversary, even has all the private keys of ring members and unbound computing resources, has no advantage in identifying the signer over random guessing.

6 Conclusion

In this paper, we present a new type of proxy ring signature with anonymity revocation and no extra information leaked which allows the original signer to

know exactly who the signer is and the original signer doesn't need to publish extra information except his original public key. It can play an important role in some real applications. The proposed scheme is not from bilinear pairings and we prove its correctness and security.

References

1. Mambo, M., Usuda, K., Okamoto, E.: Proxy signature: Delegation of the power to sign messages. IEICE Trans. Fundamentals, E79-A(9), 1338–1353 (1996)
2. Rivest, R.L., Shamir, A., Tauman, Y.: How to leak a secret. In: Boyd, C. (ed.) ASIACRYPT 2001. LNCS, vol. 2248, pp. 552–565. Springer, Heidelberg (2001)
3. Lin, W.D., Jan, J.K.: A security personal learning tools using a proxy blind signature scheme. In: Proccedings of International Conference on Chinese Language Computing, Illinois, USA, July 2000, pp. 273–277 (2000)
4. Yi, L., Bai, G., Xiao, G.: Proxy multi-signature scheme: A new type of proxy signature scheme. Electronics Letters 36(6), 527–528 (2000)
5. Zhang, K.: Threshold proxy signature schemes. 1997 Information Security Workshop, Japan, September 1997, 191–197 (1997)
6. Tan, Z., Liu, Z., Tang, C.: Digital proxy blind signature schemes based on DLP and ECDLP, MM Research, No. 21, pp. 212–217 (Preprints 2002)
7. Awasthi, A.K., Lal, S.: ID-based Ring Signature and Proxy Ring Signature Schemes from Bilinear Pairings. Cryptology ePrint Archive, Report, 2004/ 184 (2004), available at `http://eprint.iacr.org`
8. Zhang, F., Naini, R.S., Lin, C.Y.: New proxy signature, proxy blind signature and proxy ring signature schemes from bilinear pairings. Cryptology ePrint Archive (2003), available at `http://eprint.iacr.org/2003/`
9. Bender, A., Katz, J., Morselli, R.: Ring signatures:stronger definitions, and constructions without random oracles. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 60–79. Springer, Heidelberg (2006)
10. Bresson, E., Stern, J., Szydlo, M.: Threshold ring signatures and applications to ad-hoc groups. In: Yung, M. (ed.) CRYPTO 2002. LNCS, vol. 2442, pp. 465–480. Springer, Heidelberg (2002)
11. Komano, Y., Ohta, K., Shimbo, A., Kawamura, S.: Toward the fair anonymous signatures: Deniable ring signatures. In: Pointcheval, D. (ed.) CT-RSA 2006. LNCS, vol. 3860, pp. 174–191. Springer, Heidelberg (2006)
12. Naor, M.: Deniable ring authentication. In: Yung, M. (ed.) CRYPTO 2002. LNCS, vol. 2442, pp. 481–498. Springer, Heidelberg (2002)

The Research of an Embedded Processor Element for Multimedia Domain

Lai Mingche, Guo Jianjun, Lv Yasuai, Dai Kui, and Wang Zhiying

School of Computer National University of Defense Technology, Chang Sha, China
mingchelai@chiplight.com.cn

Abstract. A novel embedded processor element basing on the Transport Triggered Architecture is presented in this paper. The processor element consisting of two powerful arithmetic clusters using the application specific instruction processor design methodology achieves higher performance and is especially good at exploiting the instruction level and data level parallelisms in the multimedia applications. To improve the efficiency, the processor also presents the decoupled stream memory system with the characteristics of the stream buffer proxy to support the cross-line indexed accesses and to enhance the memory bandwidth. Then, a heterogeneous multiprocessor SoC chip involving the embedded processor is fabricated using $0.13um$ CMOS process, and the SoC operates at $400MHz$ and consumes only around $690mW$. Experimental results show that the embedded processor element has good performance improvement for the multimedia applications.

1 Introduction

With the evolvement of many modern standards and the increasing magnitude of the computing rate, the embedded systems in the multimedia domain, e.g. the H.264 encoder and the MPEG4 encoder, are required to be high in performance and low in power consumption to meet the real-time and the energy-efficiency requirements. Through the research on the multimedia applications behaviors [1], especially the analysis to the kernels, the multimedia domain is characterized with the low-precision computations and a large amount of the instruction level and the data level parallelisms. Thus, many embedded digital signal processors nowadays, such as the Tricore serials of the Infineon Inc. and the TMS320C64XX series from Texas Instruments [2], follow the superscalar and the VILW architecture respectively to pursue high performance. However, the decoder logic, the issue logic, the register file ports and the data bypasses of these DSP cores become so complex that only a little proportion of the hardware resources may be used for the computations, which results to a low utilization of the transistor resources. Especially, with rapid development of the application requirements, the scalability and low energy consumption problems [4] of these architectures have been still puzzling the system designers.

As the increasing demands make the traditional processor architecture become helpless for the application requirements, the transport triggered architecture [3] is researched for some advantages including the simple distributed

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 267–276, 2007.
© Springer-Verlag Berlin Heidelberg 2007

control mechanism, the scalable and the custom-built architecture, the denser computing resources etc. Then, this paper implements an embedded TTA processor element which is composed of two arithmetic clusters, where, each cluster is configured with the optimized transport network, the abundance of the function units as well as some specific custom units. In the compiler aspect, with assistance of the scheduler, the processor can exploit the instruction level and data level parallelisms simultaneously. Then, this paper also studies the stream memory subsystem with the characteristic of the stream buffer proxy to support the cross-line indexed accesses and to enhance the memory bandwidth. The rest of this paper is organized as follows. Section 2 introduces the TTA. Section 3 presents the processor element architecture. Section 4 studies the stream memory system. Section 5 presents the implementation. Section 6 shows the experiment results. Finally, the paper concludes with section 7.

2 Transport Triggered Architecture

The general TTA structure is very simple, shown in Figure.1. The processor core is not only deployed with the powerful transport network to supply the enough bandwidth, but also configured with the abundance of the function units, the specific custom units and the register files which are connected by the sockets. The major difference [6] between TTA and VLIW architecture is the way that the operation is executed. In the TTA programming model, the program specifies only the data transports on the interconnection network. Operations only occur as a "side-effect". Thus, the fact that there is only one data transport template in the instructions results in the SIMT (Single Instruction Multiple Transports) model of TTA programming. As shown in Figure.1, four function units (FU) and one register file (RF) are connected to the network. Every FU or RF has one or more operator registers (OR), result registers (RR) but only one trigger register (TR). Data transferred to the TR triggers the FU to work, where the function units execute different operations indicated by the operation codes.

Then, the obvious advantages of the TTA during exploiting the parallelism are summarized as follows. Firstly, the function units and transport network of the TTA are configurable with the machine description files during the ASIP process [5]. Based on the flexibility and configurability of the TTA, the high

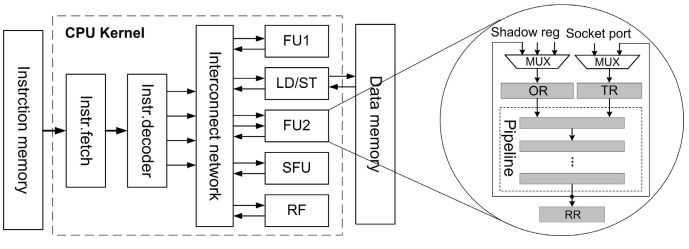


Fig. 1. General Structure of TTA

performance micro-architecture can be specially designed according to the specific applications. Secondly, the result registers of the FUs can be available to store the variable temporarily. The framework of the distributing registers not only reduces the pressure of the register files but also results in the higher performance. Thirdly, the decoupled architecture among the FUs, the RFs and the network is more suitable for exploiting their pipelines to make use of the hardware resources to the fullest extent. Fourthly, the transport triggered property of the TTA is more flexible and fine-grained than the traditional operation triggered ones, e.g. superscalar. The TTA programming model which allows the transfer parallelism among the several dependent operations is suited for the more efficient schedule in the instruction level as well as the data level.

3 Embedded Processor

3.1 Element Architecture

The whole embedded processor architecture is depicted in Figure.2. It composes of the processor control unit, the instruction cache, the stream memory subsystem, the DMA engine as well as two arithmetic clusters. The backbone of the processor is the control unit which is responsible to execute the external commands on the one hand and on the other hand to fetch and decode the microinstructions. Then, while the arithmetic clusters are busy in the multimedia computing, the DMA engine may be dedicated to the multimedia data transportation by executing the channel commands issued by the clusters to hide long memory latencies. Note that, in order to support the parallel operations in a single TTA instruction, the TTA instruction template includes eight dedicated fields, called move slots, for each bus to define data transport, while each move slot contains three domains: the guard domain, the destination domain and the source domain. Then, the processor also has three different external interfaces. The control interface for control and synchronization tasks which are handled by the control unit, the transfer interfaces for the processor instructions and the data exchanges between the local storage and the off-chip memory. Especially, The computations in the multimedia applications are accelerated in the clusters,

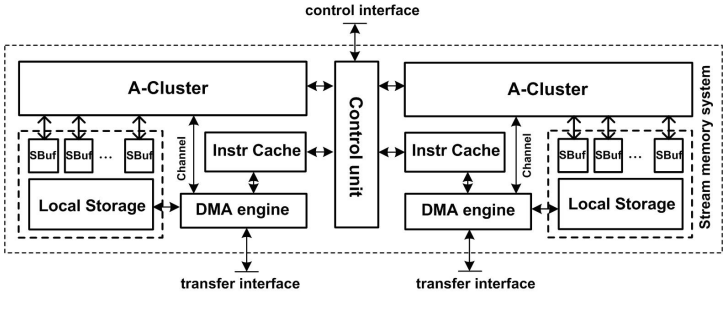


Fig. 2. Overview of the processor element Architecture

where eight 32bits buses and various function units connected to the network by the individual sockets, characterizing the interconnect-central structure with the transport trigger mechanism.

3.2 Application Specific Design Flow

In order to exploit the efficiency of the embedded processor kernels towards the multimedia computing, ASIP platform [5] of the TTA has been developed for the processor configuration exploration. The selected applications include the kernels from TI DSP library, e.g. FFT, IIR, IDCT, MAXDIX etc. and several multimedia applications including the MPEG4 decoder and the JPEG encoder. Firstly, the type and the number of major operations which are derived from the simulation results determine the type and number of the function units to meet the media application requirements. According to the type of the operations, this step quickly decides which function units to implement, then the number of function units is decided according to the individual operation proportion. As shown in the Figure.3, where, the operation proportions are derived from the sequential simulations of the target applications. Meanwhile, the average active registers gained though the trace analysis determines how many registers should be implemented to save the hardware cost while keeping high performance. Also, the parallel schedule [10] of the sequence code on the different bus configurations decides the number of the buses in the processor element. Figure.4 illustrates the acceleration ratios of the parallel simulations with the various bus configurations towards the sequence code simulation. It is very clear that the application performances increase with the bus number, while this trend tends to be saturated till the bus number reaches 8.

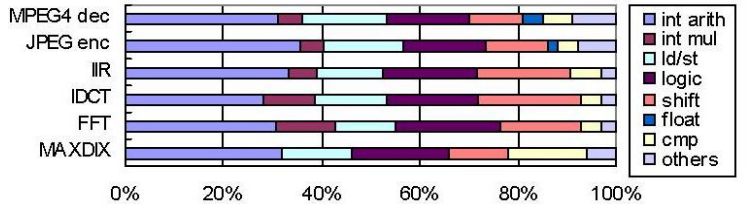


Fig. 3. Operation proportions from the sequential simulations of the applications

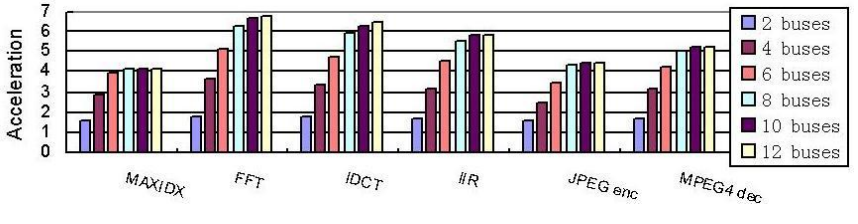


Fig. 4. Acceleration of the various bus configurations

Table 1. The configuration of each cluster

Function unit	Number	Description
Integer ALU	4	It performs the common operations including the arithmetical and logical ones. Then, it supports the sub-word parallelism.
Integer mult	1	It implements 2 cycle latency.Typically, it performs the 1-way 32x32-bit 2-way 16x16-bit or 4-way 8x8-bit multiplication.
Jump unit	1	It support the conditional jump and unconditional jump.
Compare unit	2	It makes if-conversion and conditional data flow possible.
Float unit	1	All operations are IEEE-754 compliant. The FPU is fully pipelined and a new operation is started every clock cycle.
load & store	3	It is configurable to supports the normal access and the stream way with the assistance of the stream descriptors.
Local RF	4	4 global register banks, deploys 176 local registers, consisting of 4 register sliding windows.

Secondly, the automatic methodology for the selection of custom instructions to augment the extensible the processor is used in order to maximize its efficiency for the target applications. With the accurately estimation of the performance benefits and hardware costs, the corresponding special function units are synthesized to the TTA framework. Then, due to the strict constraints from the application background, e.g. the area, an optimization towards the FU configuration and the transport network may be further applied for the higher performance with lower power consumption. Using the ASIP platform towards the target multimedia applications, the function units are listed as Table.1.

4 Memory Subsystem

As all known, many multimedia programs require a bandwidth-oriented memory system. However, the conventional cache-based memory hierarchies organized with the limited cache memories which lack the efficient prefetch mechanism and the raw bandwidth lead to a poor performance. In particular, memory bandwidth in the multimedia system can be reduced by capturing short-term producer-consumer locality in large local register files and long-term producer-consumer locality in a stream register file. These localities might not be captured by a reactive cache. Thus, the embedded processor introduces the stream register files to satisfy the memory requirements in these multimedia applications. The structure of the memory subsystem is composed of the DMA controller and the stream registers files as shown in the Figure.5, where, the image data are organized as the stream memory operations to transfer between the SRFs and the external memories with the assistance of the DMA controller. In this decoupled framework, the SRF keeps the arithmetic units busy by allowing the software to hide long memory latencies. Then, as shown in Figure.5, an exclusive management scheme is proposed to synchronize the stream transferring operations and the kernel accesses to the SRFs, which have been divided into some

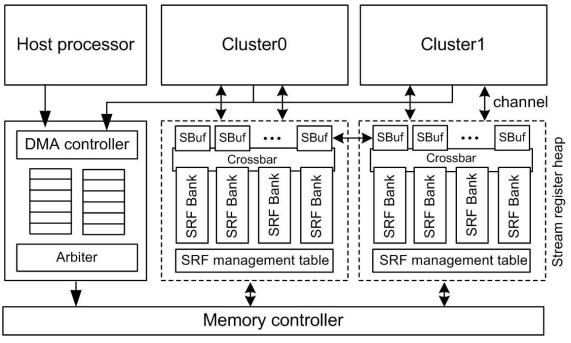


Fig. 5. Stream memory subsystem architecture

SRF pools by the logic address. Exactly, the host and the coprocessor complete to consult the SRF management tables whose items are used to be the atomic signals and record the states of the pools. Only when the returned state of the consulted pool is IDLE, the victor will lock the target pool at the moment. On the one hand, if the host processor or the channel succeeds in applying for the targeted pool, the pool state will switch to TRANS and a certain master will issue the API commands to the DMA controller to request the data transfers of the target pools. Then, the DMA controller deals with the commands using the FIFO selection. When the successive transfers are completed, the state of the target pool is set by the DMA controller to IDLE. On the other hand, when the load/store units succeed in applying for the target pool, the pool state switches to the CALCUL and the kernels just access the targeted pool. Then, when the arithmetic operations on this pool completes, the pool state is set back to IDLE.

Especially, the stream buffers between the SRFs and the clusters are introduced for the higher throughput of the stream channels as shown in Figure.5. Each cluster corresponds to its own stream register heap, which is deployed with 24KB register files and seven stream buffers which are connected with each other by the crossbar network. Note that, for the sake of the area and the power consumption, the SRF banks are organized to be single-ported, and thereby worsen the memory conflicts and limit the channel throughput. Thus, each SRF in the Figure.5 is divided into four banks for the accesses in parallel. Also, the stream buffers are proposed to match the access characteristics of the SRFs to the load/store operations of kernels. Multimedia data is scheduled through the stream buffers which are programmed using stream descriptors [9]. A compiler can also schedule the loading of a stream descriptor which is dependent on the run time values to access the time-multiplexed banks over several streams. Although stream descriptors are emphasized in this paper for the movement of image data, they can also be used to describe movement of non-image data in the system. In the special, with the introduction of the stream descriptors, the accesses to the operation registers of the load/store unit in the clusters are

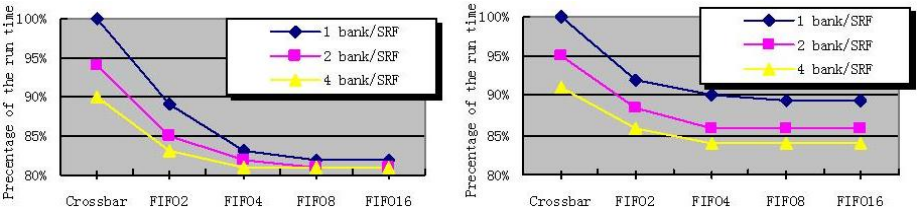


Fig. 6. Acceleration of the different configurations (left for FFT, right for JPEGENC)

reduced, then more transports may be scheduled on the saving buses to further increase the number of the triggered units. This paper selects two media applications including the FFT and JPEGENC to evaluate the performance of the memory subsystem. By adjusting the SRF structure and the FIFO size, Figure.6 illustrates the acceleration ratio of the different configurations. It is very obvious that the latter four configurations deployed with the FIFOs are priority to the crossbar structure but the acceleration effect trends to be saturated while the FIFO size increases.

In the special, some classes of the multimedia applications such as the 3D signal processing exhibit a lot of cross-lane indexed accesses which allow any cluster to access any SRF bank simultaneously. Although the statically scheduled inter-cluster transfers has been used to support data reordering patterns that can be analyzed at compile-time, it is inefficient for data-dependant access pattern since any static schedule must allocate communication resources assuming worst-case conflicts. Thus, Figure.7 (a) (b) show two approaches of indexed SRF access mechanism in Imagine [8]. The first approach in Figure.7(a) allows any cluster to access any SRF bank with dynamic conflict resolution with the assistance of the fully connected crossbar, which consumes a lots of hardware resources and weakens the advantage of the stream memory such as the scalability and the low latency. Moreover, the consistency problem exists among the FIFOs shown in the Figure.7 (a). The second approach in Figure.7 (b) introduces a reorder cache to solve the irregular stream accesses. It loads streams which need reorder from one SRF, and then store this stream reordered by the index back to another. This procedure must obligate the special local memory space for the communication which consumes the abundance of the hardware resources. Even more, this procedure may block the execution of the arithmetic clusters.and then store this stream reordered by the index back to another.

Then, the stream buffer proxy mechanism is proposed to support the cross-line indexed accesses. As illustrated in the Figure.7(c), each cluster is deployed with 5 stream buffers which are severed for the three local load/store units individually, while other 2 stream buffers act as the stream buffer proxy to support the accesses from the other cluster. Firstly, the programmer transfers a certain stream descriptor to the local load/store unit in the clusters. Secondly, once the stream descriptor analyzed to target to the other SRF, the local load/store unit will apply the remote proxy for the stream services. Note that, all the local

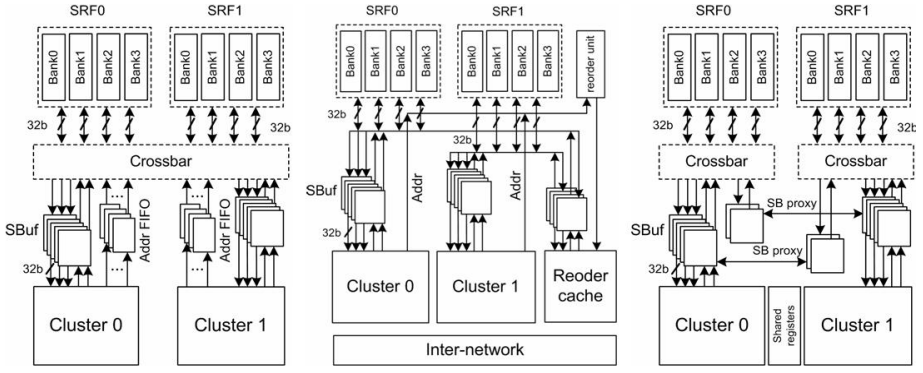


Fig. 7. Different architectures of the SRF and the stream buffers

load/store units complete to use the buffer proxy as an atomic resource, thus the victorious load/store unit need to lock the proxy buffer using the hardware primitive while accessing the remote SRF banks. Then, because all the data copies are distributed on the local stream buffers and the buffer proxies, the memory system structure in Figure.7(c) can ensure the memory consistency model only using the tiny hardware resources, and especially provides a friendly environment for the TTA-C compiler.

5 Processor Implementation

The heterogeneous multiprocessor SoC chip involving the embedded processor element is named with ATOM-II and fabricated using $0.13um$ eight-metal CMOS process with a area of $4.7 \times 4.7mm^2$. The SoC is composed of a single embedded processor element, a memory controller, a complete set of peripherals for the multimedia applications and the LEON3 processor deployed with 4KB instruction cache and 4KB data cache. This SoC can operate at $400MHz$ and consume about $690mW$. Then, the processor element is configured with 16KB instruction cache and 48KB on-chip memory with the area of nearly $8.9mm^2$.

6 Performance Evaluation

This section presents performance results for the described ATOM-II implementation. The selected benchmarks are part of next generation imaging application. The Table 2 presents the kernel benchmarks, which represent portions of the application that are mapped onto the cluster of the embedded processor element, assuming the 1-cycle memory access timing. The signal cluster performance is comparable to leading DSP processors with a similar number of function units, such as the DSP TMS320C64. To achieve the best performance, the assembly language programming and scheduling are performed to evaluate the performance

Table 2. Performance comparisons on kernel benchmarks (cycles)

Kernels	TI-C64	Single Cluster
FFT	1243	1027
FIR	1519	1377
MAXIDX	82	104
IIR	533	502

FFT: A 256-point complex FFT, 16 bits/input, normally ordered, bit reversed out.
FIR: A 128-sample, 32-taps, 16 bits/input complex FIR.
MAXIDX: status of the biggest element of the 128-element vectors, 16 bits/element
IIR: 128-sample , 16 bits/pixel

Table 3. Performance comparisons on application benchmarks (cycles)

Application	LEON	ATOM-II	Accelerated ratio
H264_dec	13.5×10^6	3.21×10^6	4.2
JPEG_enc	16.8×10^6	2.12×10^6	7.9
MPEG4_dec	3.52×10^6	0.85×10^6	4.1

H264_dec: H264 decoder, QVGA 240×320 frame
JPEG_enc: JPEG encoder, QVGA 240×320 frame
MPEG4_dec: MPEG4 decoder, QCIF 176×144 frame

on the simulator from the MOVE suit [10]. For TMS320C64, the applications are executed and evaluated in the CCS3.0, which is the software development environment which is developed by the Texas Instruments. Table 2 shows the simulator result of the TTA and the TMS320C64 [2] on four typical kernels. The metric is the execution time of the applications, representing by cycles. It is shown that the signal cluster has the better performance compared with the TMS320C64 except for the MAXIDX kernel for the frequent conditional jumps.

Secondly, this paper presents the evaluation to the entire SoC. All the application described using C language is selected from the multimedia computing domain. With the assistance of the FPAG prototype, we schedule the different computing tasks to the different clusters to accelerate the critical fractions of the application compiler by the TTA compiler. As the result shown in the Table.3, the accelerated ratio of the ATOM- to the LEON processor achieve 4.1 - 7.9.

7 Conclusions

This paper presents an embedded processor based on the TTA for the acceleration of the application in the multimedia domain. This paper presents the processor architecture and discusses function unit configuration. Through the application characteristics analysis, it is easy to explore the special function units and modify the architecture to adapt to different applications. Then, this

paper also studies the stream memory subsystem with the characteristic of the stream buffer proxy to increase the memory bandwidth, especially to support the cross-line indexed accesses without the consistency problem. Finally, a heterogeneous multiprocessor SoC chip involving the embedded processor element is fabricated. The experiment results show that the single cluster is prior to the traditional TI TMS320C64 DSP on the one hand, and on the other hand the embedded processor does the obvious acceleration to the target multimedia applications, and the average of the entire SoC is only $0.69W$.

References

1. Corbal, J., et al.: DLP+TLP processors for the next generation of multimedia workload. In: Proc. 7th Intl. Symp. on HPCA (2001)
2. TMS320C64 DSP library programmer's reference, Texas Instruments, USA (2003).
3. Corporaal, H., Janssen, J., Arnold, M.: Computation in the Context of Transport Triggered Architectures. International Journal of Parallel Programming 28(4), 401–427 (2000)
4. Dr.Gordon Cichon.: Introduction into Synchronous Transfer Architecture (STA) (February 2005)
5. Hong, Y., Li, S., Kui, D., Zhiying, W.: A TTA-based ASIP design methodology for embedded systems. The Journal of the Computer Research and Development, pp:752–758 (2006)
6. Hoogerbrugge, J., Corporaal, H.: Transport-triggering vs. operation-triggering. In: Compiler Construction conference CC-94 (1994)
7. Kapasi, U.J., Dally, W.J., Rixner, S.: Efficient Conditional Operations for Data-parallel Architectures. In: Proc. Intl. Symp. on. Microarchitecture, pp.159–170 (December 2000)
8. Jayasena, N., Erez, M., Ahn, J.H., Dally, W.J.: Stream Register Files with Indexed Access, In: Tenth International Symposium on High Performance Computer Architecture, February 2004, Madrid, Spain (2004)
9. Lopez-Lagunas, A., Chai, S.M.: Compiler Manipulation of Stream Descriptors for Data Access Optimization, In: Proceedings of the International Conference Workshops on Parallel Processing, January 2006, pp. 337–344 (2006)
10. Corporaal, H., Hoogerbrugge, J.: Code generation for Transport Triggered Architectures. In: Code Generation for Embedded Processors, Kluwer Academic Publishers, Boston, MA (1995)

Speeding Up Scalar Multiplication Using a New Signed Binary Representation for Integers

Bang-ju Wang[1,2], Huan-guo Zhang[1], Zhang-yi Wang[1], and Yu-hua Wang[1]

[1] School of Computer Science, Wuhan University, Wuhan 430079, China
[2] School of science, Huazhong Agricultural University, Wuhan 430070, China
wangbangju@mail.hzau.edu.cn

Abstract. Scalar multiplication dP and gP+hQ are important in encryption, decryption and signature in information security and wireless network. The speed of computation of scalar multiplication is significant for related applications. In this paper, a new signed binary representation (SBR) for integers called complementary code method (CC) is proposed, which has minimum weight and needs less memory. An efficient algorithm using CC method for computing dP is shown also. According to analyzing and comparing to the other methods, this algorithm is the better one in window methods and is the simplest for applying in software and hardware. By applying joint representation in computing gP+hQ, new algorithm using CC method has the least joint weight compared to other methods mentioned in this paper. So, the new SBR can efficiently speed up the computation of scalar multiplication dP and gP+hQ and can be widely used in secure communication for improving the speed of encryption and signature.

Keywords: elliptic curve cryptosystems (ECC); scalar multiplication; public key cryptosystem; signed binary representation (SBR).

1 Introduction

Public key cryptosystem solved the secure problem of key distribution soundly in symmetric cryptosystem [1] and realized digital signature and message authentication successfully in information system. ECC [2] (elliptic curve cryptography) and RSA [3] are typical cryptographies in public key cryptosystems and widely applied in secure communication. ECC is known to have high security because its security depends fully on discrete logarithm problem, which time complexity is exponential time. Furthermore, ECC has more advantages than RSA in short key, short signature, small software and low-wasted power of hardware, so the performance of ECC is significant in secure applications, such as smart cards, cellular phones, security software and hardware.

Scalar multiplication dP is the main operation in ECC, where d is an integer and P is a point on an elliptic curve. Computation of gP+hQ is widely used in digital signatures, where g and h are integers and P and Q are points on an elliptic curve. Performances of computation of dP and gP+hQ are the key to fast application of ECC.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 277–285, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The complexity of computing gP+hQ depending on the joint weight of the binary integers g and h. Similarly, the speed of computation of dP relies on the weight of the binary integer d. Katti [4] presented a new signed binary representation (SBR) for integers that increases the speed of computations in elliptic curve cryptosystems, which improved the speed of computation of dP and (gP+hQ). In [5] Xiaoyu proposed a left-to-right algorithm for computing an SBR of a pair of integers that has minimum joint weight as well as maximum average length of zero-column runs, and this SBR is modified method in [4]. Xiaoyu reduced the memory to perform the computation of gP+hQ by JSF binary representation of a pair of integers in [6]. Moon [7][8] proposed a more efficient and novel approach quad-and-add of a scalar multiplication method than existing double-and-add by applying redundant recoding which originates from radix-4 Booth's algorithm. The method results in about more than 30% performance enhancement than the algorithm using double-and-add method. A fast implementation efficient scalar multiplication based on the double-base chain representation of scalar using powers of 1/2 and 3 is proposed in [9], which using efficient point halving operation in the new double-base chain to achieve fast scalar multiplication. Solinas presented an optimal SBR Joint Sparse Form (JSF) method for a pair of integers to speed up the computation of ECC in [10].

In this paper, another new SBR is proposed to speed up scalar multiplication dP and gP+hQ.

This paper is organized as follows. ECC are firstly introduced in section 2, after that several methods of representation of integer are shown in section 3, and then in the next section, a new SBR for integers called complementary code method (CC) is proposed and analyzed. In section 5, an efficient algorithm used CC method for computing dP is presented and performance is analyzed. Moreover, another efficient algorithm using CC method for computing gP+hQ is given in section 6. Finally, conclusions are discussed in the last section.

2 ECC Algorithm

A supersingular elliptic curve $E(F_2^m)$ [11] over F_2^m is defined as:

Assume prime $p > 3$, and $4a^3 + 27b^2 \neq 0$ mod p, then $y^2+cy=x^3+ax+b$, a, b$\in$GF (F_2^m). If P and Q are points on elliptic curve, a integer t, computing Q=tP is easy. But if P and Q are known, computing t is very difficult. It is the discrete logarithm problem in elliptic curve. The basic operation is defined as follows.

1 P+∞ = ∞+ P=P for all P$\in$GF (F_2^m), $\infty$is called the point at infinity.

2 If P = *(x,y)* $\in$GF(F_2^m) , then (x,y)+(x,y+c)=∞. The point (x,y+c) is denoted –P

3 Let P= (x_1, y_1) and Q=(x_2, y_2), P, Q$\in$GF (F_2^m) and P$\neq$ $\pm$Q, then P+Q=(x_3,y_3) and

$$x_3 = \left(\frac{y_1 + y_2}{x_1 + x_2}\right)^2 + x_1 + x_2 \text{ and } y_3 = \left(\frac{y_1 + y_2}{x_1 + x_2}\right)(x_1 + x_3) + y_1 + c.$$

4 Let $P=(x_1, y_1)$, P≠-P, then P+P=2P=(x_3,y_3), where

$$x_3 = \left(\frac{x_1^2 + a}{c}\right)^2 \text{ and } y_3 = \left(\frac{x_1^2 + a}{c}\right)^2 (x_1 + x_3) + y_1 + c.$$

3 Methods of Representations for Computing dP

In this section, four methods to compute dP are described.

Ordinary method
Generally, an integer d is represented in binary format as:

$$d = (d_{L-1} \ldots d_1 d_0) = \sum_{i=0}^{L-1} d_i 2^i,$$

Where $d_i \in \{0, 1\}$, Then

$$P = \sum_{i=0}^{L-1} d_i 2^i P = d_0 P + 2d_1 P + 2^2 d_2 P + \ldots + 2^{L-1} d_{L-1} P$$

If the weight of d in the ordinary binary representation is wd_1, then the computation of dP by above equation need wd_1-1 additions and L doubles. so reducing weight is the key to improving the speed of scalar multiplication.

NAF method
The same integer d as above is represented in NAF [12] as:

$$d = (a_{K-1} \ldots a_1 a_0) = \sum_{i=0}^{K} a_i 2^i,$$

where $a_i \in \{-1, 0, 1\}, K = L \text{ or } L-1$. So

$$\begin{aligned} dP &= \sum_{i=0}^{K} a_i 2^i P = a_0 P + 2a_1 P + 2^2 a_2 P + \cdots + 2^{L-1} a_{L-1} P \\ &= a_0 P + 2(a_1 + 2(a_2 P + 2(a_3 P + \cdots .))) \end{aligned}$$

NAF method has below properties applied in ECC:
(1) No two consecutive digits are nonzero in NAF;
(2) NAF is unique and has minimal density of all signed digit representations;
(3) The expected weight of a NAF of length K is K/3;
(4) The computation of the negation of a point P=(x,y)(-P=(x,x+y) is free, so the cost of addition or subtraction is same.

If the weight of d in NAF method is wd_2, then $wd_2 \leq wd_1$, then NAF method requires less addition than Ordinary method. So computation of dP in NAF method is more effective than in Ordinary method, and the speed depends on the weight of the integer d.

KT method
Famous KT method is described in [13]. The transformation rule of this method is described as the follow example. If B=(1... b_i... 1) can be transformed into C (10... c_i... 1), where c_i=b_i-1. For example if B is (1,011,1) (note bi is 011) then it can be transformed into C=(10, -100, -1). In the example above the weight has been reduced from 4 to 3.

Katti method
For a given $d = (d_{L-1} \cdots d_1 d_0)$, d is represented [4] as follow:
$d = (d_{L-1} - 0)2^L + (d_{L-2} - d_{L-1})2^{L-1} + \cdots + (d_0 - d_1)2 - d_0$.
The detail of Katti method is described as Algorithm1:
(1) Convert the binary representation of d into X,

$$X = ((d_{L-1} - 0), (d_{L-2} - d_{L-1}), \cdots, (d_0 - d_1), (0 - d_0)).$$

(2) Convert X into Y by going from left to right and replacing 1, -1 by 0, 1 and -1, 1 by 0, -1.
(3) Convert Y into Z by going from left to right and replacing 1,0,-1 by 0, 1,1 and -1,0, 1 by 0, -1,-1.
For example, if d=25,722,562,047,811,804,959, the binary representation of d (65 bits) is:
(10110010011111000111001011110011000000100110001000101111100011111)
Z is obtained by applying Katti method algorithm1 as below:
Z=(10110010100001001001001100001010000001001100010001100001001000$\overline{1}$), where -1 is denoted by $\overline{1}$.

4 New Method to Represent Integer d

In this section, a new signed binary method to represent d called complementary code (CC) method will be introduced. This representation can be easily computed from the binary representation compared to the computation of NAF or Katti representation. Similarly, for a given integer d, and

$$d = (a_0 a_1 \cdots a_{K-1}) = \sum_{i=0}^{K} a_i 2^{K-i}$$

The main idea of CC method is to eliminate the consecutive '1s' in sequence $\{a_i\}$. It's known as that there has an equation: d=2^K-1-d_1, where d_1 is the complement of A. When there has consecutive '1' in sequence of d, making the transformation:d=2^K-1-d_1. If the representation of d_1 has less '0' than that of d, then the representation is meaningful. The detail process of transformation is described in algorithm2.
Algorithm2:
Step1: For computation dP, the binary representation of d is d=$(a_0 a_1 \cdots a_{K-1})$;
Step2: For all a_i (i=0,1,...,K-1), from left to right (from a_0 to a_{K-1}) do:
When a_i=1 and a_{i+1}=1, do following steps:

$a_{i-1} = a_i$;
a_m=-(a_m+1)&1 (m=i, i+1,i+2...K-1);
$a_{K-1} = a_{K-1}$-1;

When a_i=-1 and a_{i+1}=-1, do following steps:

$a_{i-1} = a_i$;
a_m=(a_m+1)&1 (m=i, i+1,i+2...K-1);
$a_{K-1} = a_{K-1}$+1;

Step3: d can be represented by this new method in D, and

$$D = (d_{L-1} \cdots d_1 d_0) = \sum_{i=0}^{L-1} d_i 2^i, \text{Where } d_i \in \{-1, 0, 1\}.$$

For example, a integer d=727=(1011010111), then d is transformed into D by new representation in algorithm2:

(1) d can be represented by binary method $d = (a_0 a_1 \cdots a_{K-1})$=(1011010111);

(2) Changing binary representation by Step2 above:

$(a_2 \times a_3 \neq 0) \rightarrow a_{1=1;}$

D=(1011010111) $(a_2 \times a_3 \neq 0) \rightarrow$ (1100-10-100-1), the weight is 5 and the length of sequence is 10.

d can be also rewritten by NAF method as d=(10-100-10-100-1), the weight is 5 and the length of sequence is 11.

So, for a given integer d, the representation in CC method has less length of sequence than in NAF method, and it has same weight as that of NAF method. Furthermore, the new method can transform the integer from left to right, so it doesn't need store memory. That's to say, CC method has same good efficiency as NAF method and can save memory, moreover, its transformation is simpler than NAF method.

5 Computing dP

Scalar multiplication dP is the key to secure application of ECC and is widely used in encryption, decryption and signature in network security and e-commerce. The weight of integer d determines the speed of computation when computing dP. So how to reduce the weight of integer d is significant. An efficient method by applying CC method representation for computing dP will be shown. The method of computing dP is described as algorithm3.

Algorithm 3:

Step 1: Transforming integer $d = (a_0 a_1 \cdots a_{K-1})$ into new binary representation E as algorithm2, that's

$$E = (d_{L-1} \cdots d_1 d_0) = \sum_{i=0}^{L-1} d_i 2^i, \text{ Where } d_i \in \{-1, 0, 1\}$$

Step 2: Convert E into D by going from left to right and replacing 0,1,0, -1 by 0,0, 1,1 and 0, -1,0, 1 by 0,0, -1, -1.

Step 3: Splitting the sequence $\{d_i\}$ into segments.

Step 4: Computing dP with window methods.

For example, if d=25,722,562,047,811,804,959, the binary representation of d (65 bits) is:

(1011001001111100011100101111001100000010011000100010111110001111 1)

All the results transformed by Algorithm3 and other three methods for computing dP are shown below, where -1 is denoted by $\overline{1}$.

The KT algorithm transforms d into d_{-KT} :

(1011001010000$\overline{1}$001000$\overline{11}$0$\overline{1}$0000$\overline{11}$0$\overline{1}$0000001001100010001100000$\overline{1}$0010000$\overline{1}$), So,d_{-KT} has 65-digit and the weight is 23.

d is represented in Katti method as d_{-Katti} :

(1011001010000$\overline{1}$00100$\overline{1}$00110000$\overline{11}$0$\overline{1}$0000001001100010001100000$\overline{1}$0010000$\overline{1}$), d_{-Katti} is expressed as 65-digit number and the weight is 22.

Then, d is represented in NAF as d_{-NAF} :

(10$\overline{1}$0$\overline{1}$001010000$\overline{1}$00100$\overline{1}$010$\overline{1}$000$\overline{1}$010$\overline{1}$00000010 10$\overline{1}$0001001 0$\overline{1}$0000$\overline{1}$0010000$\overline{1}$), d_{-NAF} is 66-digit number and the weight 22 is the same as of d_{-Katti}.

Now, d is represented in algorithm3:

E = (110$\overline{1}$001010000$\overline{1}$00100$\overline{1}$0011000$\overline{1}$010$\overline{1}$00000010 10$\overline{1}$0001000 1100000$\overline{1}$0010000$\overline{1}$), then d_{-CC} is gotten by step2 in algorithm3:

$d_{-CC} = D$

= (110$\overline{1}$001010000$\overline{1}$00100$\overline{1}$00110000$\overline{11}$0$\overline{1}$0000001001100010001100000$\overline{1}$0010000$\overline{1}$), CC method is expressed as 65-digit number and the weight is 22.

It's found that KT, Katti method, and algorithm2 all have the shortest bit-length of d, which are 65. Furthermore, Katti method, NAF method and algorithm2 all have the minimum weight 22. So, according to above representations for d, algorithm2 and Katti method all have the same weight and bit-length. However, the representation by algorithm2 can be gotten by one traveling from left to right, Katti will need three times. That's to say, algorithm2 is the best one for representation of d among these methods.

Window method is the important method for computing dP. Now, sequence in all representations will be split into segments for the window method. The window size is 4 and the segments are shown in the below:

d is split into segments by KT method as d_{-KT},

(1011 0010 1010 0000 $\overline{1}$001 0001 $\overline{11}$0 $\overline{1}$000 0 $\overline{11}$0 $\overline{1}$000 0001 001 1000 1000 1100 0000 $\overline{1}$001 0000 $\overline{1}$), it has 11 segments in window method.

d_{-Katti} is split into 12 segments also:

(1011 0010 1010 0000 $\overline{1}$001 00$\overline{1}$001 1000 0$\overline{11}$0 $\overline{1}$000 0001 001 1000 1000 1100 0000 $\overline{1}$001 0000 $\overline{1}$).

d_{-NAF}can be split into 14 segments:

(10$\overline{1}$0 $\overline{1}$001 0101 0000 $\overline{1}$001 00$\overline{1}$0 10$\overline{1}$0 $\overline{1}$000 $\overline{1}$010 $\overline{1}$000 0010 10$\overline{1}$0 0001 0010 $\overline{1}$000 $\overline{1}$001 0000 $\overline{1}$),

d_{-CC} is split into 12 segments as:

(110$\overline{1}$ 0010 1010 0000 $\overline{1}$001 00$\overline{1}$001 1000 0$\overline{11}$0 $\overline{1}$000 0001 001 1000 1000 1100 0000 $\overline{1}$001 0000 $\overline{1}$).

Table 1. The performation of ccomparison among four methods

Methods	Digital Length	Weight	Window	Transform Direction	AL(d)
Algorithm3	65	22	12	left→right	1.45
KT method	65	23	11	right→left	1.44
Katti method	65	22	12	left→right	1.44
NAF method	66	22	14	right→left	1.35

Now, the average length of zero runs in integer d will be considered, which is denoted by AL(d). AL(d) is defined as follow [4]:

$$d = \sum_{i=0}^{L-1} d_i 2^i, AL(d) = \frac{1}{L} \sum_{i=0}^{L-1} Z(i), \text{ where } z(-1) = 0.$$

Moreover, if $d_i = 0, z(i) = 1 + z(i\text{-}1)$, and if $d_i \neq 0$, $z(i) = 0$, $(0 \leq i \leq L\text{-}1)$

For different representations, the values of average length of zero runs for d_{KT}, d_{Katti}, $d_{NAF}$and d_{CC} are listed in below table 1.

So, in table 1, our methods algorithm3 and the other three methods are compared in digital length, weight, windows, transformation direction, and AL(d).

According to above results, Table 1 shows that algorithm3 is characteristic of the shortest digit length of sequence, the minimum weight, the maximum average length of zero runs and the simplest transformation. Moreover, it is found that the average length of zero runs for NAF method is the minimal value 1.35, and for KT method and Katti is 1.44. Obviously, new algorithm is better than NAF method in all aspects. That's, algorithm3 is the best one among these methods for the window method of computing dP. Additionally, the transformation of representations by algorithm3 is easy and simple, and it's simpler to implement in hardware or software. So we can say, our method algorithm3 is the most efficient method among the four methods in applying in computing scalar multiplication dP in ECC.

6 Computing gP+hQ

CC method can be applied in computing gP+hQ also. Computation of gP+hQ is very important in Digital signatures, where g and h are integers and P and Q are points on an elliptic curve. The joint weight of g and h is the key to the speed of computation of gP+hQ. For example, computing gP+hQ.

Suppose g=6699 and h=4846. The binary form of a and b is given[6] by:

$$\begin{bmatrix} g \\ h \end{bmatrix} = \begin{bmatrix} 1101000101011 \\ 1001011101110 \end{bmatrix}$$

When applying algorithm2 in computing gP+hQ, in order to keep minimum joint weight, detail processes will be modified as algorithm4.

Table 2. Comparison of joint weights in four different methods

Methods	Joint Representation	Joint weight
Our method	1101000101011 $100110001\overline{1}00\overline{1}0$	9
Binary method	1101000101011 1001011101110	10
NAF method	$10\overline{1}0100101\overline{0}10\overline{1}0\overline{1}$ $01010\overline{1}000\overline{1}00\overline{1}0$	11
Katti method	$100\overline{11}000101011$ $0100110001\overline{0}0\overline{1}0$	10

Algorithm4:

Step1: Transforming integer g and h into new binary representation G and H as algorithm2:

Step2: when doing step1, it's noted that the first two '1' and last two '1' may not be transformed to keep the same length of binary sequence when the length of integer g and h is different.

Because joint weight determines the speed of computing gP+hQ, so, how to decrease the joint weight is decisive in improving the speed of digital signature. Four methods in joint representation are compared by joint weight In Table 2, where –1 is denoted by$\overline{1}$.

Table 2 shows that the joint weight of our method is 9 and is the least one among these five methods. Its representation is different from the result of Xiaoyu [5] and that of the JSF [10], but they all result in minimum joint weight. So our method is efficient in applying in computation of gP+hQ.

7 Conclusions

A new SBR CC method (see in algorithm2) is proposed in this paper,which has minimum weight and can save memory. Then, efficient algorithms (algorithm3 and algorithm4) using new SBR for computing scalar multiplication are shown. According to the analyzing and comparing to famous KT method, Katti method and NAF method (see Table 1), the new method in this paper is the best method for window method and is the simplest applying in computation of dP. For improving the speed of computation of gP+hQ, algorithm4 has the least joint weight compared to the other methods (see Table 2). That's to say, the new SBR method shown in this paper is efficient to improving the speed of scalar multiplication dP and gP+hQ. Moreover, CC method can save memory and computation of transformation, then it can be easily applied in software and hardware. So, the new SBR is simple and efficient for speeding up scalar computation and can be efficiently used in encryption, decryption and signature in network security, e-commerce and wireless communication, etc.

Acknowledgements. This paper is supported by the National Natural Science Foundations of P.R.China under Grant No. 60373087, 90104005 and 60473023.

References

1. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory, IT-22, 644–654 (1976)
2. Koblitz, N.: Elliptic curve cryptosystems. Mathematics of computation 48, 203–209 (1987)
3. Rivest, R.L., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public-key Cryptosystems. Communications of the ACM 21, 120–126 (1978)
4. Katti, R.: Speeding up Elliptic Cryptosystems using a new Signed Binary Representation for Integers. In: Proceeding of the Euromicro Symposium on Digital System Design (DSD'02), Dortmund, DE, pp. 380–384 (2002)
5. Xiaoyu, R., Katti, R.: On the Signed-Binary Window Method. In: international conference of ISCAS 2005, Kobe, Japan, vol. 5, pp. 4501–4504 (2005)
6. Xiaoyu, R., Katti, R.: Left-to-Right Optimal Signed-Binary Representation of a Pair of integers. IEEE Transactions on Computers 54, 132–140 (2005)
7. Moon, S.: Elliptic Curve Scalar Point Multiplication Using Radix-4 Booth's Algorithm. In: International Symposium on Communications and Information Technologies, pp. 26–29 (2004)
8. Moon, S.: A Binary Redundant Scalar Point Multiplication in Secure Elliptic Curve Cryptosystems. International Journal of Network Security 3, 132–137 (2006)
9. Wong, K.W., Edward, C., Lee,W., Cheng, L.M., Liao, X.: Fast elliptic scalar multiplication using new double-base chain and point halving. Applied Mathematics and Computation 183, 1000–1007 (2006)
10. Solinas, J.A.: Low-Weight Binary Representations for Pairs of Integers. Technical Report, CORR 2001-41, Center for Applied Cryptographic Research, University of Waterloo, Canada (2001)
11. Hankerson, D., Menezes, A., Vanstone, S.: Guide to Elliptic Curve Cryptography. Huanguo Zhang, Publishing House of Electronics Industry, pp. 74–76 (2005)
12. Gordon, D.M.: A survey of fast exponentiation methods. Journal of Algorithms 27, 129–146 (1998)
13. Koyama, K., Tsuruoka, Y.: Speeding up elliptic cryptosystems by using a signed binary window method. In: Brickell, E.F. (ed.) CRYPTO 1992. LNCS, vol. 740, pp. 345–357. Springer, Heidelberg (1993)

Color-Texture Image Segmentation by Combining Region and Photometric Invariant Edge Information

Shengyang Yu[1], Yan Zhang[2], Yonggang Wang[1], and Jie Yang[1]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
whizstorm@sjtu.edu.cn
[2] Department of Computer Science, Nanjing University og Science and Technology, Nanjing, China

Abstract. An improved approach for JSEG algorithm is proposed for unsupervised color-texture image segmentation. The region and photometric invariant edge information are combined. A novel measure for color-texture homogeneity is defined by weighting the textural homogeneity measure with photometric invariant edge measure. Based on the map whose pixel values are values of the new measure, region growing-merging algorithm used in JSEG is then employed to segment the image. Finally, experiments on a variety of real color images demonstrate performance improvement due to the proposed method.

1 Introduction

Color image segmentation, decomposition of images into disjoint homogeneous regions, plays an important role in many computer vision applications such as image retrieval, object detection and recognition. Homogeneity is usually defined as similarity with respect to certain properties of pixel. However, it is difficult to determine homogeneous region in color images since natural scenes are rich in color and texture. Various methods have been proposed in the literature, and some of the recent works include, stochastic model-based approaches [1][2], morphological watershed based region growing [3], graph partitioning[4], and clustering method in color space [5]. However, there are few algorithms that can work well on a large variety of color image.

Deng and Manjunath [6] proposed a new approach called JSEG for unsupervised segmentation of color-texture regions in images. The basic idea of the JSEG method is to analyze the similarity of the color and their distributions separately, and the segmentation process is separated into two stages, color quantization and spatial segmentation. Although the robustness of the JSEG algorithm has been shown by experiments on a variety of real images, JSEG has several potential limitations. For example, it's difficult to set a suitable combination of parameter for quantization and region merging. Over-segmentation usually happens in the cases of spatially varying illumination. Several methods have been proposed recently to fix the inherent problems of JSEG. Wang et al. [7] applied adaptive mean-shift clustering for color quantization, and the clustering results were used to construct Gaussian mixture models for the soft class map, where each pixel was assigned with membership degree of every

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 286 – 294, 2007.
© Springer-Verlag Berlin Heidelberg 2007

class. Zheng et al. [8] introduced fuzzy mechanism into JSEG. However, these methods are complicated and bring some new problem.

In this paper, a new approach is presented to improve JSEG from the aspect of color-texture homogeneity measure. The photometric color invariant edge and texture information are combined, which results in a novel measure of color-texture homogeneity. The proposed measure takes account of both textural homogeneity and color discontinuity in local region. Experimental results on real color images show the robustness and the accuracy of this method. The rest of the paper is organized as follows. Section 2 reviews the JSEG method analyzing the limitations of JSEG. An improved color-texture segmentation method is proposed in Section 3 and some experimental results are given in Section 4. We conclude this paper in Section 5.

2 JSEG Algorithm

Due to the difficulty of analysis the similarity of the color and their distributions simultaneously, JSEG consists of two independent steps: color quantization and spatial segmentation. Firstly, colors in image are quantized to several representative classes, and a class-map of the image is constructed by replacing the pixel values with their corresponding color class labels. The class-map can be viewed as a special kind of texture composition. Then, spatial segmentation is performed on this class-map without considering the corresponding pixel color similarity.

A criterion for "good" segmentation using spatial data point in class-map is proposed. Let Z be the set of all N data points in a class-map. Let $z=(x,y), z\in Z$, and μ be the mean. Suppose Z is classified into C classes, $Z_i, i=1,\ldots,C$. Let μ_i be the mean of the data points of class Z_i.

$$S_T=\sum_{z\in Z}\|z-\mu\|^2 \tag{1}$$

And

$$S_w=\sum_{i=1}^{C}S_i=\sum_{i=1}^{C}\sum_{z\in Z_i}\|z-\mu_i\|^2 \tag{2}$$

S_w is the total variance. Define

$$J=(S_T-S_W)/S_W \tag{3}$$

J value can be calculated by applying J to a local area of the class-map, which can indicate whether that area is in the region interiors or near region boundaries. An example of class-maps and the corresponding J values is shown in Fig.1. Thus, a J-map whose pixel values correspond to these J values calculated over small windows centered at the pixels is constructed. The higher the local centered value is, the more likely that the corresponding pixel is near a region boundary. Finally, spatial segmentation based on region growing method is used on J-map which is constructed at different scales. For more detail information, please refer to [6].

Fig. 1. An example of two class-maps and their J values. o , * and + indicate three classes of data points respectively.

In JSEG method, the quantization parameter and region merging threshold are two important parameters, the former determines the minimum distance between two quantized colors, and the latter determines the maximum similarity between two adjoining regions in spatial segmentation. An appropriate quantization parameter can result in an accurate class-map, which will lead to an outstanding segmentation. Furthermore, when quantization parameter is small, over-segmentation is very obvious and a large region merging threshold is often needed. Contrarily, when quantization parameter is large, region merging threshold should be small. However, it's hard to identify a combination of a quantization parameter and a region merge threshold which will make the algorithm more efficient. In addition, JSEG calculates J values over the class-map, which describes the texture information but does not consider color information of the pixels. In other words, the measure J can only characterize the homogeneity of texture, but fails to represent the discontinuity of color. Undoubtedly, the limitation of J measure degrades the robustness and discrimination of JSEG.

3 The Improved Segmentation Method

In this section, the improved segmentation method is introduced in detail. First, in section 3.1, a color invariant edge is introduced. Then, in section 3.2, a novel measure is proposed. In section 3.3, spatial segmentation is briefly reviewed.

3.1 Color Invariant Edge

The absence of color discontinuity in J measure is the main cause of over-segmentation of JSEG. Edge detection can obtain the discontinuity of color. However, general edge detection methods are sensitive to noise, and have no invariance to specific imaging conditions. In real-world there are various physical phenomena which will trigger differential-based features and object reflectance changes. To provide robustness against imaging conditions (e.g. illumination, shading), photometric color invariants have been introduced [9][10][11] by modeling the physical process of color image formation.

In this paper, we focus on invariants under the imaging condition of equal energy illumination and matte, dull object. Geosebroek et al. [9] proved that within the Kubelka-Munk model [12], $C_\lambda = \frac{E_\lambda}{E}$ and $C_{\lambda\lambda} = \frac{E_{\lambda\lambda}}{E}$ are object reflectance properties independent of the viewpoint, surface orientation, illumination direction and illumination intensity.

$$E(\lambda, x) = i(x)\{(1-\rho_f(x))^2 R_\infty(\lambda, x) + \rho_f(x)\} \tag{4}$$

is the photometric reflectance model resulting from the Kubelka-Munk theory, where x denotes the position at the imaging plane and λ the wavelength. Further, the spatial component $i(x)$ denotes intensity variations and $\rho_f(x)$ the Fresnel reflectance at x. The material reflectivity is denoted by $R_\infty(\lambda, x)$. For matte, dull surfaces, the Fresnel coefficient can be considered neglectable, for which (4) reduces to the Lambertian model for diffuse body reflection

$$E(\lambda, x) = i(x) R_\infty(\lambda, x). \tag{5}$$

Differentiating E with respect to λ twice results in E_λ and $E_{\lambda\lambda}$. The spatial derivative of the intensity normalized spectral slope C_λ and curvature $C_{\lambda\lambda}$ are given by

$$C_{\lambda x} = \frac{E_{\lambda x} E - E_\lambda E_x}{E^2}, C_{\lambda\lambda x} = \frac{E_{\lambda\lambda x} E - E_{\lambda\lambda} E_x}{E^2}. \tag{6}$$

Then a total edge strength measure is given by

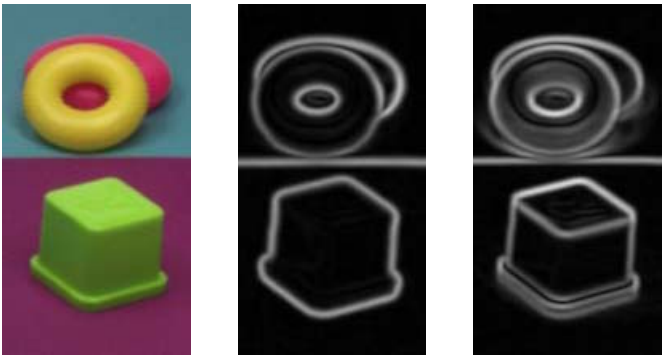
$$C_w = \sqrt{C_{\lambda x}^2 + C_{\lambda y}^2 + C_{\lambda\lambda x}^2 + C_{\lambda\lambda y}^2}. \tag{7}$$

Spectral differential quotients denoted by $\hat{E}$, $\hat{E}_\lambda$, and $\hat{E}_{\lambda\lambda}$ are obtained by a linear combination of given (RGB) sensitivities, the implementation is given by

$$\begin{bmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{8}$$

The spatial differential quotients denoted by $\hat{E}_x$, $\hat{E}_{\lambda x}$, and $\hat{E}_{\lambda\lambda x}$ are computed through Gaussian smoothed derivatives. The gradient magnitude $\hat{C}_w$ can be calculated according to formula (7). An example for color edge measures $\hat{C}_w$ and $\hat{E}_w$ is shown in Fig. 2. E_w is not invariant to any change in imaging conditions , which is given by

$$E_w = \sqrt{E_x^2 + E_y^2 + E_{\lambda x}^2 + E_{\lambda y}^2 + E_{\lambda\lambda x}^2 + E_{\lambda\lambda y}^2}. \tag{9}$$



(a) (b) (c)

Fig. 2. An example for total color edge strength measures. (a) original image; (b) $\hat{C}_w$; (c) $\hat{E}_w$

3.2 Construction of a Novel Map

In our method, the edge measure C_w is combined with the measure J in JSEG, which results in a new discriminative measure for color-texture image segmentation.

For a $M \times N$ color image, the $J(x, y)$ value for each pixel of local region is calculated according to Eqs. (1)-(3). Correspondingly, with the same scale, the magnitude $C_w(x, y)$ for each pixel of the color image is calculated according to Eq.(7). Then, $C_w(x, y)$ is normalized according to

$$R_C(x, y) = \frac{C_w(x, y) - C_w^{\min}}{C_w^{\max} - C_w^{\min}} . \tag{10}$$

where $C_w^{\min} = \min\{C_w(x, y)\}$ and $C_w^{\max} = \max\{C_w(x, y)\}$. The novel measure is given by

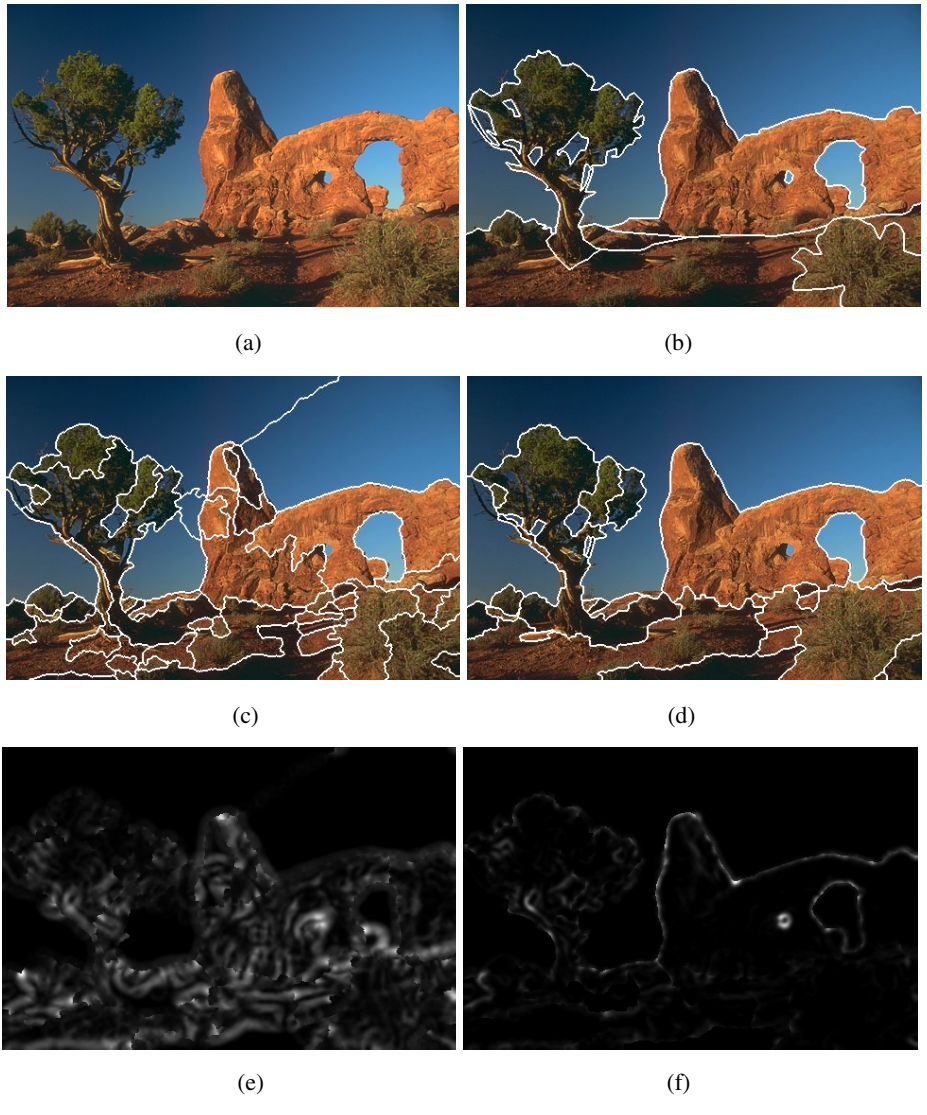
$$J_C(x, y) = R_C(x, y) * J(x, y) . \tag{11}$$

A novel map whose pixel values correspond to these J_C values calculated over local windows centered at the pixels is constructed.

Due to the introduction of photometric invariant edge information, the proposed measure J_C takes account of both the textural homogeneity and color discontinuity, and suppresses those pseudo edges derived from shading. The higher the local J_C value is, the more likely that the corresponding pixel is near a region boundary. Contrarily, a small local J_C value indicates corresponding pixel be in a region interior. Similarly, the characteristics of the novel map allow us to use a region growing method to segment the image.

3.3 Spatial Segmentation

JSEG provides a region growing based segmentation scheme to segment the image. The scheme consists of three main operations: seed area determination, region growing and region merging. The algorithm starts the segmentation of the J -image at a coarse initial scale. Then, it repeats the same process on the newly segmented regions at the next finer scale. The same scheme of spatial segmentation is employed in our work to segment the novel map constructed by using the measure J_C at multi-scale.



(a) (b)

(c) (d)

(e) (f)

Fig. 3. A segmentation example: (a) original image; (b) manual segmentation result; (c) JSEG; (d) our method; (e) J-map, and (f) the novel map at scale 2, respectively

4 Experimental Results

The improved algorithm is tested on 200 images chosen from Berkeley segmentation dataset [13], where the human labeled images are taken as ground truth. The proposed

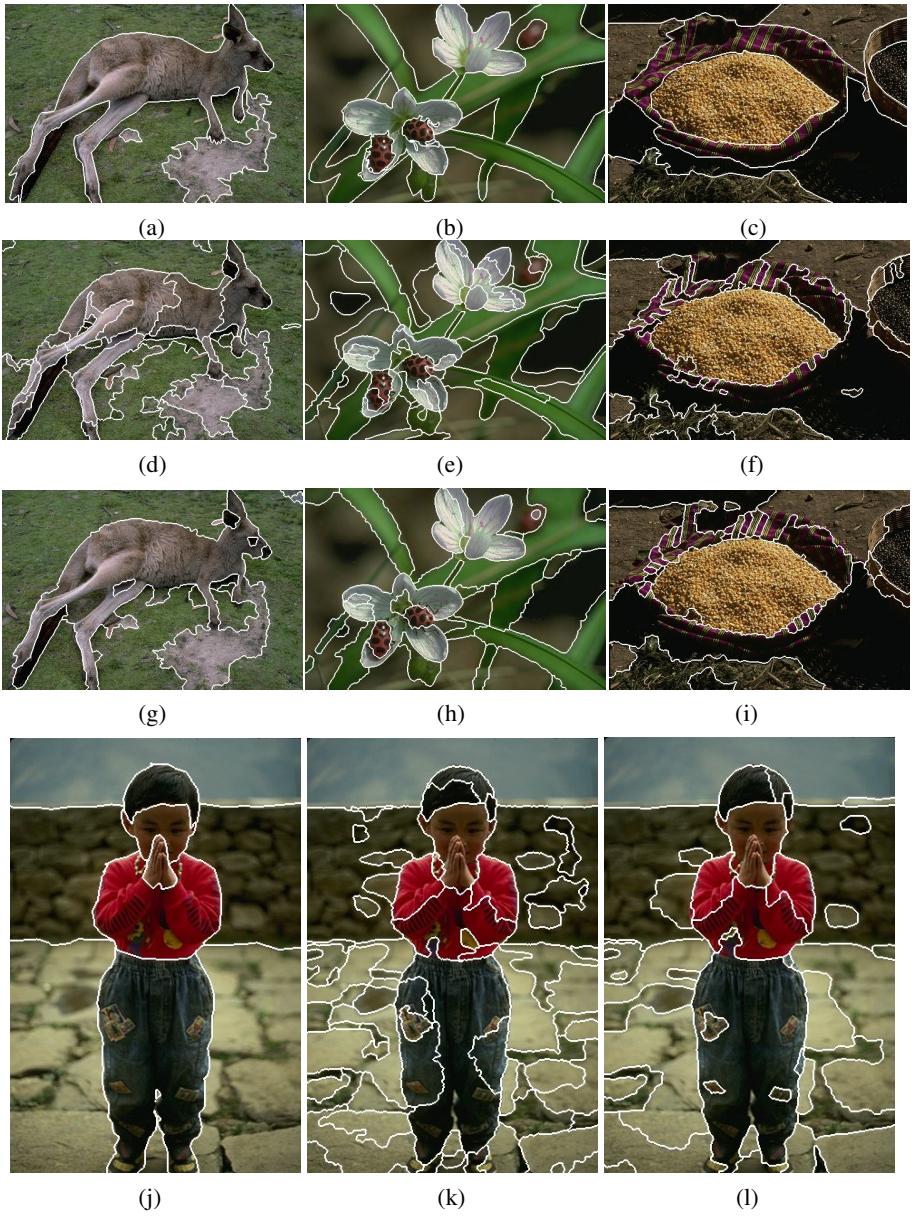


Fig. 4. (a, b, c, i) manual segmentation result; (d, e, ,f, ,k) Segmentation by JSEG; (g, h, i, l) segmentation by proposed method

algorithm is compared with JSEG. The results show that our method is more robust and more accurate.

Due to the size of all images from Berkeley segmentation dataset is 321×481, three scales are used in our test, which is consistent with JSEG method. The color quantization parameter is 200, and the region merging threshold is set to 0.4. The same parameters are used in all the experiments. Fig.3. gives a segmentation example. The hand segmented result, the segmented result of JSEG, the J -map and the novel map are shown.

Obviously, in Fig.3, the image is over-segmented by the JSEG method. Because the color of sky varies from blue to dark in a very smooth transition, the pixels in the region of sky are assigned with two or more kinds of label. Due to illumination, the surface of hill is also over-segmented. The proposed measure can enhance the J measure of describing color discontinuity discriminatively (see Fig.3(e) and Fig.3(f)). A more reasonable segmentation of our method is shown in Fig. 3(d). Fig. 4 shows four more examples.

Martin et al. [13] also presented two error measures to evaluate segmentation algorithms in an objective manner, and the Local Consistency Error (LCE) measure is used in our experiment to get the average accuracy of segmentations on the test images. The LCE measure of our method is 24.3% and that of JSEG is 36.1%.

The photometric color invariants provide robustness against imaging conditions, and the introduced color invariant edge measure can effectively describe color discontinuity. As expected, the proposed measure effectively characterizes textural homogeneity and color discontinuity synchronously.

5 Conclusion

In this work, an improved approach for JSEG algorithm is proposed for unsupervised color-texture image segmentation. A photometric invariant edge measure is employed to describe color discontinuity discriminatively. A new measure for color-texture homogeneity is defined by combining measure J and photometric invariant edge measure. Based on the novel map, region growing-merging algorithm is used to segment the image. The proposed algorithm can give effective segmentation results and commendably solve the problem of over-segmentation of JSEG.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 60675023).

References

1. Delignon, Y., Marzouki, A., Pieczynski, W.: Estimation of Generalized Mixtures and Its Application in Image Segmentation. IEEE Trans. Image Processing 6, 1364–1376 (1997)
2. Wang, J.P.: Stochastic Relaxation on Partitions with Connected Components and its Application to Image Segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence 6, 616–636 (1998)
3. Shafarenko, L., Petrou, M., Kittler, J.: Automatic Watershed Segmentation of Randomly Textured Color Images. IEEE Trans. Image Processing 11, 1530–1544 (1997)

4. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence 8, 888–905 (2000)
5. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 5, 603–619 (2002)
6. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE Trans. Pattern Analysis and Machine Intelligence 8, 800–810 (2001)
7. Wang, Y., Yang, J., Zhou, Y.: Unsupervised Color-Texture Segmentation. In: Campilho, A., Kamel, M. (eds.) ICIAR 2004. LNCS, vol. 3212, pp. 106–113. Springer, Berlin (2004)
8. Zheng, Y., Yang, J., Zhou, Y.: Unsupervised Segmentation on Image with JSEG Using Soft Class Map. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 197–202. Springer, Heidelberg (2004)
9. Geusebroek, J.M., van den Boomgaard, R., Seulders, A.W.M., Geerts, H.: Color Invariance. IEEE Trans. Pattern Analysis and Machine Intelligence 12, 1338–1350 (2001)
10. Gevers, T., Smeulders, A.: Color Based Object Recognition. Pattern Recognition 32, 453–464 (1999)
11. van de Weijer, J., Gevers, T., Geusebroek, J.M.: Edge and Corner Detection by Photometric Quasi-invariants. IEEE Trans. Pattern Analysis and Machine Intelligence 4, 625–630 (2005)
12. Wyszecki, G., Stiles, W.S.: Color Science: Concepts and Methods. Quantitative Data and Formulae. Wiley, New York (1982)
13. Martin, D., Fowlkes, C., Tal, D., et al.: A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics, In Werner, B. (ed.) Proc. Int. Conf. Computer Vision 2001, vol. 2, pp. 416–423. IEEE Computer Society Press, Los Alamitos, CA (2001)

Multimedia Authoring Tool for Real-Time Facial Animation

H.W. Byun

School of Media & Information, Sung Shin Woman University,
169-1 Dongsun-dong 2, Sungbuk-gu, Seoul, Korea
hyewon@sungshin.ac.kr

Abstract. We present a multimedia authoring tool for real-time facial animation based on multiple face models. In order to overcome the heavy burden of geometry data management on various multimedia applications, we employ wire curve [16] which is a simple, intuitive interface to local deformation of complex geometric objects such as human face models. Given multiple face models, we first extract wire curves and deformation parameters from a facial model. In runtime, given an input of expression parameters, only the extracted wire curves are blended to create new expression. Then, the blended output of the wire curves is applied to the reference model of neutral expression. The resulting animation preserves the characteristic features of the multiple face models as well real-time performance. Our method promotes local deformation and non-uniform blending by making use of the power of wire curve.

Keywords: real-time facial animation, multiple face models, local deformation, wire deformation.

1 Introduction

Creating lifelike characters for facial animation is an interesting and challenging task. Especially, high quality facial animation breathes emotion and personality into newly created characters. Existing facial animation techniques in various multimedia applications are hampered by a lot of obstacles, such as limited resources and the requirement of real-time speed. Our goal is to achieve realistic and expressive facial animation with real-time performance for multimedia authoring. We employ both multiple face models and wire curves.

Multiple face models, called templates, are widely used for facial animation of the conversational characters on desktop platforms.[1,3,6,12,13,14,18]. Each of these models reflects both a facial expression of different type and designer's insight to be a good guideline for animation. Those templates comprise a facial expression database from which we select appropriate face models to blend and to deform. However, a template consists of a set of vertices with few handles to control such geometric operations except the vertices themselves. It is almost impossible to keep all the vertices of multiple face models on mobile devices with

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 295–304, 2007.
© Springer-Verlag Berlin Heidelberg 2007

limited resources. Furthermore, to achieve smooth local deformation and non-uniform blending without any tools, one would have to deal with every vertex of face models involved in these operations.

"Wires"[16] have become a popular tool for facial animation, due to its abstract representation and capability of local control on facial features such as eyes and a mouth. The basic idea of wire deformation is to locally deform the vertices of a facial feature according to the displacements of wire curves from their references together with deformation parameters (see Section 2.) Therefore, by deforming a wire curve, the features near the curve are also deformed accordingly. These features may be refined further with deformation parameters.

In order to abstractly represent the templates, we present a method to extract a set of wire curves and deformation parameters from a template. The extracted wire curves have replaced the templates. This lowers a heavy burden of a lot of template geometry data on various multimedia application such as web and game with limited resources. In runtime, given the input of new expression parameter, the blended output of the extracted wire curves is applied to a reference template to create real-time facial animation. Our method not only provides handles for local deformation and non-uniform blending but also reduces the volume of the template database, by representing a template as a set of wire curves and deformation parameters characterizing its corresponding facial expression.

The remainder of the paper is organized as follows. We provide related work in Section 2. In Section 3, we give an introduction to the wire deformation technique. We present a formulation for extracting wire curves and deformation parameters in Section 4 and describe facial animation method based on templates in Section 5. In Section 6, we demonstrate how our technique can be used for facial animation. Finally, we conclude this paper in Section 7.

2 Related Work

Blending multiple face models with different expressions is popular for real-time facial animation [1,6,12,13,14]. Pighin et al.[12] captured face geometry and textures by fitting a generic face model to a number of photographs. Through transitions between captured face models of different expressions, they were able to generate expressive facial animation. Blanz et al.[1] and Pighin et al.[13] proposed an automatic face modeling technique by linearly blending a set of example faces from a database of scanned 3D face models. To avoid unlikely faces, they restricted the range of allowable faces with constraints derived from the example set. For local control on each facial feature, those approaches allow interactive segmentation of a face into a set of regions to assign a proper blending factor to every vertex. Joshi et al.[6] adopted an automatic segmentation method to learn the local deformation information directly from the multiple face models.

Other alternatives are based on deformation techniques. In free-form deformation(FFD)[15], control points of a parallelpiped-shaped lattice are manipulated to deform an object. Further extensions to FFD adopted lattices of

arbitrary topology instead of regular lattices[2,10]. For direct control, Hsu et al.[5] computed the displacements of the control points from the movements of points on the surface of an object. Thalmann et al.[8] employed FFD to simulate the muscle action on the skin surface of a human face. Kahler et al.[7] and Terzopoulos et al.[9] proposed a physically-based method for skin and muscle deformation to enhance the degree of realism over purely geometric techniques. Guenter et. al.[4] used facial motion data captured from real actors to deform face models. Marschner et al.[11] computed the displacements of control points for a specific face model from the movements of sample points on a face performer by solving a system of linear equations in the least squares sense.

Singh et al.[16] provided a more effective control metaphor based on wire deformation. A parametric curve called "wire curve" is used to define and directly control a salient deformable feature on a face model. Its basic idea is to locally deform geometry near the wire curve by manipulating the curve. Due to the capability of local control as well as direct manipulation, wire deformation is versatile for synthesizing facial expressions interactively.

3 Wire Curve

Singh *et. al.* [16] proposed wire deformation as a simple, intuitive interface to deforming complex geometric objects such as human face models. Wire deformation is defined by a tuple $< W, R, f, r, s >$, where W and R denote parametric curves called wire and reference curves, respectively, and f, r, and s are deformation parameters to be explained later. Initially, W and R are coincident. By deforming the wire curve W, it is displaced from R. For a point $\mathbf{p}$ on an object M to deform, let $\mathbf{p}_R$ be its nearest point on R, and $\mathbf{p}_W$ the point on W corresponding to $\mathbf{p}_R$. That is, $\mathbf{p}_R$ and $\mathbf{p}_W$ have the same curve parameter value. When W is deformed, the point $\mathbf{p}$ is moved to $\mathbf{p}'$ as follows.

$$\mathbf{p}' = \mathbf{p} + (s-1)(\mathbf{p} - \mathbf{p}_R)f(x) + (\mathbf{p}_W - \mathbf{p}_R)f(x). \tag{1}$$

Here, x is a function of R, $\mathbf{p}$, and the range parameter r. The x is proportional to the Euclidean distance from $\mathbf{p}_R$ to $\mathbf{p}$, that is normalized by r. The function f is a monotonically decreasing function of x and the s is a scaling parameter.

Multiple wire curves can be used to better control the shape of the object. Let $\Delta\mathbf{p}_j$ be the displacement of the point $\mathbf{p}$ when a wire curve W_j alone is applied. Given n wire curves W_j, $j = 0, 1, 2, \cdots, m$, the new position $\mathbf{p}'$ is obtained as follows:

$$\mathbf{p}' = \mathbf{p} + \frac{\sum_{j=0}^{m} \Delta\mathbf{p}_j f_j(x)^k}{\sum_{j=0}^{m} f_j(x)^k}. \tag{2}$$

Here, $f_j(x) = f(x(R_j, \mathbf{p}, r_j))$ where R_j and r_j are the reference curve and the range parameter corresponding to W_j. The localizing parameter k controls the influence of W_j and s_j on deformation.

4 Wire Extraction

Suppose that we use m pairs of wire and reference curves, denoted by (W_j, R_j), $j = 0, 1, 2, \cdots, m$, to characterize the geometry of a facial expression template T deformed from the base model M by displacing their vertex positions. Then, our problem is:

> Given M and T, determine the curve pairs, $(W_j, R_j), j = 0, 1, 2, \cdots, m$ and deformation parameters, f, r, s, and k such that T can be obtained from M through wire deformation by using those curve pairs and parameters.

Let $\mathbf{p}_i$ and $\mathbf{p}'_i, i = 0, 1, 2, \cdots, n$ be the vertices of M and its corresponding vertices of T. From equation (2), $\Delta\mathbf{p}_{ij}$ is the displacement of $\mathbf{p}_i$ when only W_j is applied, $f_{ij} = f(x(R_j, \mathbf{p}_i, r_j))$, and $\Delta\mathbf{p}_i = \mathbf{p}'_i - \mathbf{p}_i$. From Equation (1),

$$\Delta\mathbf{p}_{ij} = (s_j - 1)f_{ij}(x)(\mathbf{p}_i - \mathbf{p}_{iR_j}) + f_{ij}(x)(\mathbf{p}_{iW_j} - \mathbf{p}_{iR_j}). \tag{3}$$

where $\mathbf{p}_i$, s_j, $\mathbf{p}_{iR_j}$, and $\mathbf{p}_{iW_j}$ are a vertex $\mathbf{p}_i$ on M, the scaling parameter of the reference curve R_j, the point on R_j closest to $\mathbf{p}_i$, and the point on W_j corresponding to $\mathbf{p}_{iR_j}$. As stated in the previous section, W_j and R_j are parametric curves. In particular, we employ cubic B-splines to represent them, where $\mathbf{w}_{jl}$ and $\mathbf{r}_{jl}, l = 0, 1, 2, \cdots, t_j$ are the control points of wire and reference curves W_j and R_j, respectively, and $B_l, l = 0, 1, 2, ..., t_j$ are their basis functions.

From Equations (2) and (3),

$$\Delta\mathbf{p}_i = \sum_{j=0}^{m} \frac{f_{ij}(x)^k}{\sum_{j=0}^{m} f_{ij}(x)^k}[(s_j-1)f_{ij}(x)(\mathbf{p}_i-\mathbf{p}_{iR_j})+f_{ij}(x)\sum_{l=0}^{t_j} B_l(\mathbf{w}_{jl}-\mathbf{r}_{jl})]. \tag{4}$$

Given $\mathbf{p}_i$, we can determine $\mathbf{p}_{iR_j}$ and its curve parameter value on R_j and thus B_l can be evaluated. The only unknowns on the right-hand side of Equation (4) is the control points $\mathbf{w}_{jl}$'s of the wire curve W_j and their scaling factor s_j's.

We are going to solve Equation (4) for s_j and W_j. Let

$$w_{ij} = (f_{ij}(x)^{k+1})/(\sum_{j=0}^{m} f_{ij}(x)^k), \tag{5}$$

$$\mathbf{c}_i = \sum_{j=0}^{m} w_{ij}(\mathbf{p}_i - \mathbf{p}_{iR_j}) \text{ and } \mathbf{q}_{jl} = \mathbf{w}_{jl} - \mathbf{r}_{jl}, l = 0, 1, \cdots, t_j. \tag{6}$$

Then, Equation (4) becomes

$$\Delta\mathbf{p}_i = \mathbf{c}_i + \sum_{j=0}^{m} w_{ij}((\mathbf{p}_i - \mathbf{p}_{iR_j})s_j + \sum_{l=0}^{t_j} B_l\mathbf{q}_{jl}), i = 0, 1, 2, \cdots, n. \tag{7}$$

Here, $\mathbf{q}_{jl}$'s and s_j's are the only unknowns. We have a system of linear equations:

$$\mathbf{Bs} + \mathbf{Cq} = \mathbf{b}, \tag{8}$$

where

$$\mathbf{B} = \begin{bmatrix} w_{00}(\mathbf{p}_0 - \mathbf{p}_{0R_1}) & w_{01}(\mathbf{p}_0 - \mathbf{p}_{0R_2}) & \cdots & w_{0m}(\mathbf{p}_0 - \mathbf{p}_{0R_m}) \\ w_{10}(\mathbf{p}_1 - \mathbf{p}_{1R_1}) & w_{11}(\mathbf{p}_1 - \mathbf{p}_{1R_2}) & \cdots & w_{1m}(\mathbf{p}_1 - \mathbf{p}_{1R_m}) \\ \cdots & \cdots & \cdots & \cdots \\ w_{n0}(\mathbf{p}_n - \mathbf{p}_{nR_1}) & w_{n1}(\mathbf{p}_n - \mathbf{p}_{nR_2}) & \cdots & w_{nm}(\mathbf{p}_n - \mathbf{p}_{nR_m}) \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} w_{00}B_0 & w_{00}B_1 & \cdots & w_{00}B_{t_0} & w_{01}B_0 & w_{01}B_1 & \cdots & w_{01}B_{t_1} & \cdots & w_{0m}B_0 & w_{0m}B_1 & \cdots & w_{0m}B_{t_m} \\ w_{10}B_0 & w_{10}B_1 & \cdots & w_{10}B_{t_0} & w_{11}B_0 & w_{11}B_1 & \cdots & w_{11}B_{t_1} & \cdots & w_{1m}B_0 & w_{1m}B_1 & \cdots & w_{1m}B_{t_m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{n0}B_0 & w_{n0}B_1 & \cdots & w_{n0}B_{t_0} & w_{n1}B_0 & w_{n1}B_1 & \cdots & w_{n1}B_{t_1} & \cdots & w_{nm}B_0 & w_{nm}B_1 & \cdots & w_{nm}B_{t_m} \end{bmatrix},$$

$$\mathbf{q} = (\mathbf{q}_{00}\ \mathbf{q}_{01}\ \cdots\ \mathbf{q}_{0t_0}\ \mathbf{q}_{10}\ \mathbf{q}_{11}\ \cdots\ \mathbf{q}_{1t_1}\ \cdots\ \mathbf{q}_{m0}\ \mathbf{q}_{m1}\ \cdots\ \mathbf{q}_{mt_m})^T,$$

$$\mathbf{s} = (s_1\ s_2\ \cdots\ s_m)^T, \text{ and } \mathbf{b} = (\Delta\mathbf{p}_0 - \mathbf{c}_0\ \Delta\mathbf{p}_1 - \mathbf{c}_1\ \cdots\ \Delta\mathbf{p}_n - \mathbf{c}_n)^T.$$

We can further simplify Equation (8), juxtaposing matrices $\mathbf{B}$ and $\mathbf{C}$, that is, $\mathbf{A}\hat{\mathbf{q}} = \mathbf{b}$, where $\mathbf{A} = [\mathbf{B}|\mathbf{C}]$ and $\hat{\mathbf{q}} = (\mathbf{s}^T, \mathbf{q}^T)^T$. The system is over-constrained, since the number of vertices in the face model M is much greater than the total number of control points for all reference (or equivalently wire) curves and their scaling factors. Therefore,we need to compute the least squares solution, that is, $\hat{\mathbf{q}} = (\mathbf{A}^T\mathbf{A})^+\mathbf{A}^T\mathbf{b}$. Here, $(\mathbf{A}^T\mathbf{A})^+$ is the pseudo inverse of $(\mathbf{A}^T\mathbf{A})$ obtained from its singular value decomposition.

5 Facial Animation

To create facial animation, we adopt a novel example-based approach to blend the facial examples in accordance with their contributions to synthesize the resulting expression. The facial examples are represented by wire curves. With the advantages of this approach that preserve the characteristic features of examples and reflecting a designer's intention accurately, we present an example-based scheme for facial motion synthesis in conjunction with facial motion capture.

Given the displacements of feature points, our problem is to find the best blend of examples at each frame to resemble the facial motion of a performer. Based on multidimensional scattered data interpolation, this scheme consists of three steps: parameterization, weight computation, and runtime expression blending. Provided with the source examples and corresponding target examples, in the pre-processing, all the target examples are parameterized by using the corresponding source examples to apply scattered data interpolation. We provide a simple, elegant parameterization scheme for effective expression blending. Provided with the parameterized target examples, the next step is to compute the contribution of every target example to the new expression using cardinal basis functions. The final step is to blend the corresponding

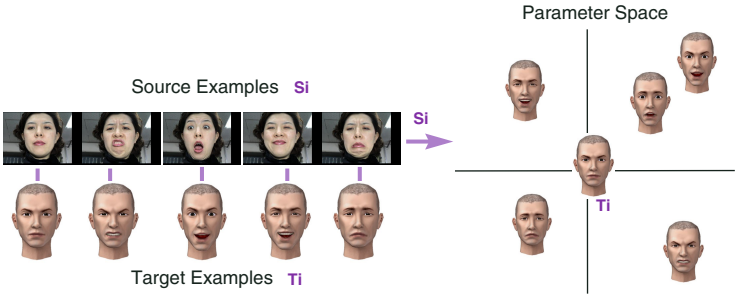


Fig. 1. The displacement vector of each source key-model $\mathbf{S}_i$ is used for parameterizing the corresponding target key-model $\mathbf{T}_i$

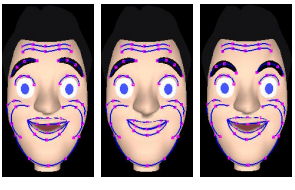


Fig. 2. The examples of high level user interface for local deformation

target examples in accordance with their contributions to synthesize the resulting expression.

We parameterize the target examples based on the displacements between the source examples. In the capture step, the displacements of feature points are extracted from a face performer. Concatenating these displacements, the displacement vector of each source example is formed to parameterize the corresponding target example. Most individual parameter components tend to be correlated to each other. Thus, based on PCA(principal component analysis), the dimensionality of the parameter space can be reduced by removing less significant basis vectors of the resulting eigenspace. As shown in Figure 1, the displacement vector $\mathbf{v}_i$ of source example $\mathbf{S}_i$ places each target example $\mathbf{T}_i$ in the N-dimensional parameter space, where N is the number of components, that is, three times the number of feature points.

With the target 3D face examples thus parameterized, our problem is how to blend the examples so as to resemble the input expression extracted from a face performer. Our problem is essentially one of scattered data interpolation, as we have very sparse target examples in a relatively high dimensional parameter space. To solve the problem, we predefines an weight function for each target example based on cardinal basis functions [17], which consist of linear and radial basis functions. The global shape of weight function is first approximated by linear basis functions, and then adjusted locally by radial basis functions to exactly interpolate the corresponding example.

6 Experimental Results

For our experiments, we have built a base face model and its templates of different expression types. The expression templates are derived by designers from the base model through displacing the vertices of the base model. We define 15 reference (and thus wire) curves lying on the base model. Each of reference and wire curves is a cubic B-spline and has four or more control points.

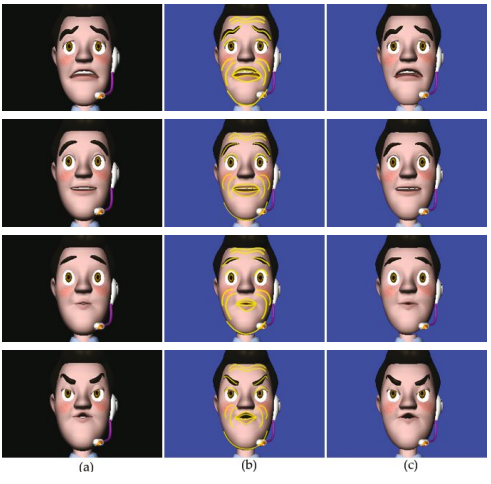
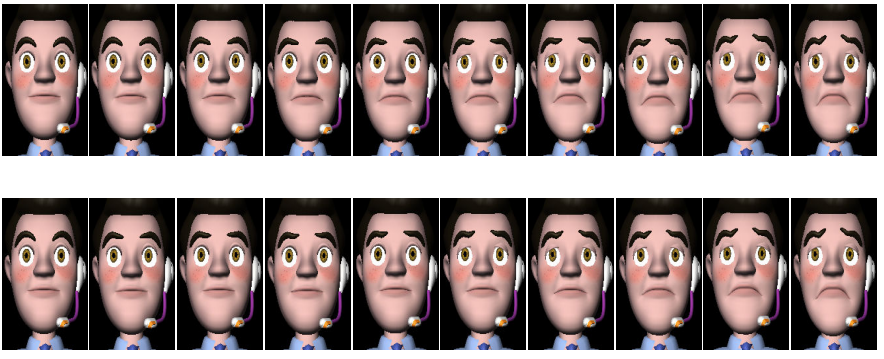


Fig. 3. The original face templates and their corresponding reconstructed face templates. (a) The original face templates with sad, happy, surprised, and angry expression (b) The extracted wire curves (c) The reconstructed face templates.

First, we show how well our wire extraction scheme works. In Figure 3, the original templates are arranged side by side with their corresponding reconstructed templates. We give the original templates in the first column (Figure 3 (a)) and the extracted wire curves on the reconstructed templates in the second column (Figure 3 (b)). The reconstructed templates are created by deforming the base model with the extracted curves and deformation parameters. In the third column (Figure 3 (c)), we get rid of the wire curves from the reconstructed templates in the second column to show them more clearly. We can observe that they are visually very similar.

Now, we exhibit the capability of wire curves for local deformation. With the wire curves and deformation parameters extracted, we can use them as a high level user interface to locally deform facial features such as eyes, lips, forehead, cheeks, and etc, instead of interactively manipulating every vertex involved in the deformation, individually. Figure 2 shows local deformation achieved with such wire curves: The left figure shows a smiling expression and its corresponding wire curve configuration, the expression in the middle is obtained by manipulating mainly the wire curves that characterize the lips, and the expression on the right is obtained by deforming mainly the wire curves on eyebrows.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 4. Upper row: uniform blending. lower row: non-uniform blending.

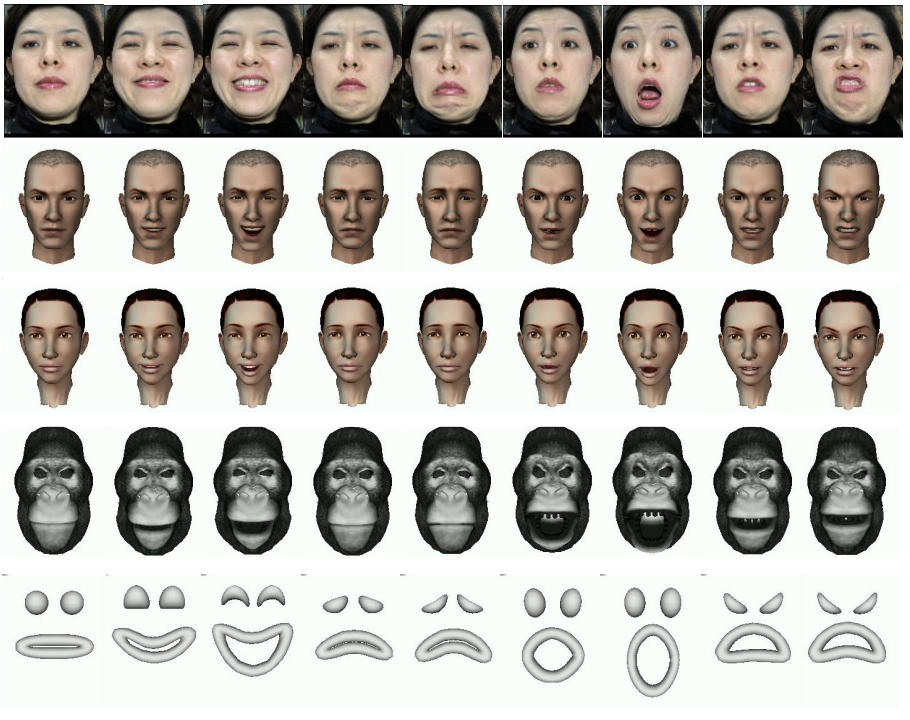


Fig. 5. Facial Animation

Finally, we demonstrate the ease of non-uniform blending with wire curves extracted. The upper row of Figure 4 shows the uniform blending of facial features,

that is, uniform feature interpolation between two templates represented by image (a) and image (j), respectively. Each facial feature of the former is transited to that of the later at the same speed. The lower row gives an image sequence due to their feature-wise non-uniform blending. For this non-uniform blending, we use different blending functions for eyes and the mouth. Feature-wise non-uniform blending can hardly be achieved efficiently without an effective user-interface such as the wire deformation scheme.

To demonstrate the final facial animation, we create several examples for various 3D face models. Figure 5 illustrates the facial animation as a result of deriving a 3D face model from the feature curve, extracted from a face performer. The first column shows the original video of the face performer. The video comprises a total sequence of 800 frames that are recorded at 30 frames per second. With thirteen target key-models, we made the facial animation for each 3D face model, shown in the next four rows. Four different styles(Man, Woman, Monkey, and Toy) of 3D face models were used to show the usefulness of our example-based approach. Each result of facial animation keeps the personalities of each face model and reflects the designer's original intention. We can observe that our approach works well even though the shape of the performer's face and 3D face model largely differ.

7 Conclusion

In this paper, we present a multimedia authoring tool for real-time facial animation. To create facial animation, we adopt a novel example-based approach to blend the wire curves of face examples in accordance with their contributions to synthesize the resulting expression. A set of wire curves and deformation parameters can be extracted from a face model regardless of its construction history. Given a pair of reference and face models with an identical topological structure, we formulate a system of linear equations of which the unknowns are the positions of control points of each wire curve and scaling parameters. This system is over-constrained since the number of vertices in the face models is much greater than that of unknowns. We extract the wire curves and parameters by solving the system for the least squares solution. The wire curves together with the scaling parameters, thus extracted, not only provide convenient handles for local geometry control but also facilitate non-uniform transitions among facial expression templates. The experimental results show the effectiveness of wire curve extraction and its usefulness.

Acknowledgement

This work was supported by the Sungshin Women's University Research Grant of 2006.

References

1. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH 1999 Conference Proceedings, pp. 187–194 (1999)
2. Chang, Y.-K., Rockwood, A.P.: A generalized de casteljau approach to 3d free-form deformation. In: SIGGRAPH 94, vol. 28, pp. 257–260 (1994)
3. Chuang, E., Bregler, C.: Performance driven facial animation using blendshape interpolation. Stanford University Computer Science Technical Report, CS-TR-2002-02 (2002)
4. Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F.: Making faces. In: SIGGRAPH 98 Conference Proceedings, pp. 55–67 (1998)
5. Hsu, W.M., Hugues, J.F., Kaufman, H.: Direct manipulation of free-form deformation. In: SIGGRAPH 92, pp. 177–184 (1992)
6. Joshi, P., Tien, W.C., Desbrun, M., Pighin, F.: Learning controls for blend shape based realistic facial animation. In: Eurographics/SIGGRAPH Symposium on Computer Animation (2003)
7. Kahler, K., Haber, J., Seidel, H.-P.: Reanimating the dead: Reconstruction of expressive faces from skull data. In: SIGGRAPH 2003 (2003)
8. Kalra, P., Mangili, A., Thalmann, N.M., Thalmann, D.: Simulation of facial muscle actions based on rational free form deformations. In: Eurographics 92 58, 59–69 (1992)
9. Lee, Y., Terzopoulos, D., Waters, K.: Realistic modeling for facial animation. In: SIGGRAPH 95 Conference Proceedings, pp. 55–62 (1995)
10. MacCracken, R., Joy, K.I.: Free-form deformations with lattices of arbitrary topology. In: SIGGRAPH 96 vol. 30, pp. 181–189 (1996)
11. Marschner, S., Guenter, B., Raghupathy, S.: Modeling and rendering for realistic facial animation. In: EUROGRAPHICS Rendering Workshop 2000, pp. 98–110 (2000)
12. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: SIGGRAPH 98 Conference Proceedings, pp. 75–84 (1998)
13. Pighin, F., Szeliski, R., Salesin, D.: Resynthesizing facial animation through 3d model-based tracking. In: International Conference on Computer Vision, pp. 143–150 (1999)
14. Pyun, H., Kim, Y., Chae, W., Kang, H.W., Shin, S.Y.: An example-based approach for facial expression cloning. In: Eurographics/SIGGRAPH Symposium on Computer Animation (2003)
15. Sederberg, T.W., Parry, S.R.: Free-form deformation of solid geometric models. In: SIGGRAPH 86 vol. 20, pp. 151–160 (1986)
16. Singh, K., Fiume, E.: Wires: A geometric deformation technique. In: SIGGRAPH 98 Conference Proceedings, pp. 299–308 (1998)
17. Sloan, P.J., Rose, C.F., Cohen, M.F.: Shape by example. In: Proceedings of 2001 Symposium on Interactive 3D Graphics, pp. 135–144 (2001)
18. Zhang, Q., Liu, Z., Guo, B., Shum, H.: Geometry-driven photorealistic facial expression synthesis. In: Eurographics/SIGGRAPH Symposium on Computer Animation (2003)

Object Re-detection Using SIFT and MPEG-7 Color Descriptors

Philipp Schügerl, Robert Sorschag, Werner Bailer, and Georg Thallinger

Institute of Information Systems and Information Management
JOANNEUM RESEARCH, Graz, Austria
{firstname.lastname}@joanneum.at
http://www.joanneum.at/iis/en

Abstract. Information about the occurrence of objects in videos and their interactions conveys an important part of the semantics of audiovisual content and can be used to narrow the semantic gap in video analysis, retrieval and summarization. Object re-detection, which aims at finding occurrences of specific objects in a single video or a collection of still images and videos, is an object identification problem and can thus be more satisfactorily solved than a general object recognition problem. As structural information and color information are often complementary, we propose a combined object re-detection approach using SIFT and MPEG-7 color descriptors extracted around the same interest points. We evaluate the approach on two different data sets and show that the MPEG-7 ColorLayout descriptor performs best of the tested color descriptors and that the joint approach yields better results than the use of SIFT or color descriptors only.

Keywords: object recognition, SIFT, interest points, MPEG-7, color descriptors, matching, video browsing.

1 Introduction

Information about the occurrences of objects in images and videos and their interactions conveys an important part of the semantics of audiovisual content. Thus the recognition of objects can help to narrow the semantic gap and is useful for various applications in the field of multimedia content analysis and retrieval. In this context, the term object includes both animate and inanimate objects, as well as settings ("background objects").

Objects can be recognized at different levels of specificity and we distinguish between object classification and identification. Over the last decades work on automatic object recognition of both types has been done from numerous directions with varying success. Object classification is only feasible if the objects of interest and/or the specific application domain are known a priori and when a limited number of object classes is used. In spite of these restrictions state of the art object classification techniques perform poor for real-world problems. For object identification the opposite is true, as a number of solutions for object identification problems lead to satisfying results nowadays.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 305–314, 2007.
© Springer-Verlag Berlin Heidelberg 2007

One object identification problem is object re-detection, which aims at finding occurrences of specific objects in a single video or a collection of still images and videos. In contrast to general object identification, the samples for learning an object are taken directly from the video on which the object re-detection is performed. Therefore this problem is based on the assumption that specific objects undergo only minor changes between all occurrences in the video, for instance, that a person wears the same clothes in all scenes. Of course this is not necessarily true for feature films but it holds for raw material of media productions ("rushes"), which is our focus in this work. Another advantage is that very similar imaging conditions exist between the learned sample and the other occurrences of an object because the same capture and recording devices, image compression methods, etc. are used. Nevertheless, object re-detection in real world images and videos is a difficult task due to varying lighting conditions and viewpoints under which objects are shown in real cluttered scenes.

There are a number of applications in multimedia content analysis, retrieval and authoring that benefit from the information that is extracted by object re-detection, including the following:

Tracking is an important analysis tool to learn about object motion and interactions in video. Object re-detection can be used to improve the tracking result by connecting parts of the object's trajectory, that have been split due to occlusion, shot boundaries or other distortions. Moreover, object re-detection makes it possible to track object occurrences over non-connected shots.

Content Structuring. Object occurrences are a salient feature of visual content that conveys much more semantics than low-level features. The information about occurrences of the same object can thus be used to cluster and structure a single media item or a collection of media items in order to support efficient browsing and retrieval. There are applications in audiovisual archiving and post-production for the organization or navigation of content by occurrences of the same foreground or background object, e.g. to find multiple takes of the same scene or scenes shot at the same location.

Support for Manual Annotation. The most high-level, reliable and semantically meaningful metadata are without doubt those added by human annotators. However, the drawback of manual annotation is its enormous effort (and thus cost), which can be reduced by better annotation tools that support the annotator. If information about object occurrences in the video is available, it can be used to apply the annotation which the user has made for a certain object in one shot to all other occurrences of the same object and thus significantly reduce the time needed for annotation.

The rest of the paper is organized as follows: Section 2 discusses the state of the art, in Section 3 we discuss the selection and combination of descriptors in our approach and Section 4 presents the details of our approach and the application in which it is used. In Section 5 the approach is evaluated on two data sets and Section 6 concludes the discussion of the results.

2 Related Work

2.1 Object Recognition Approaches

Many different approaches to object recognition exist, including neural network based approaches, graph matching, genetic algorithms and fuzzy systems. The major difference between these approaches is their different representation of objects which discriminates them into the so-called model-based and appearance-based approaches [14]. Model-based approaches use 3-dimensional models of the object shape to represent an object with geometric features such as lines, vertices and ellipses, while global or local photometric features are used for appearance-based approaches. In recent years methods using local features [8,10] have become most popular because they have solved a number of object recognition problems. With these features robustness to small perspective view changes as well as to partial occlusion is achievable and objects can be recognized anywhere in an image, with arbitrary size and rotated, without using a previous object segmentation step.

Local features are usually extracted from numerous image regions around *interest points*[14] and store visual information (color or texture) of these regions in *local descriptors*. A typical object recognition system that works with local features performs the recognition task in the following steps: (a) First some objects of interest are learned. Therefore local descriptors are extracted from images of these objects and stored in an object database. (b) To recognize objects in a test image, again local descriptors are extracted from this image and (c) matched against the descriptors in the object database. After the best matching descriptor pairs are found, (d) an optional verification step can be performed to decide whether an object appears in the test image or not.

2.2 Combining Local Descriptors

Although color is a very important cue for the identification of objects, most local feature approaches are based on textural information only. Existing color information in images is discarded because of the fact that invariance to different lighting conditions such as shadows or illumination changes is hard to achieve when using color features. Also, for many object recognition problems, color is not the most relevant feature, as it is often not discriminating for different objects in an image or for a class of objects. Nevertheless, cases exist in which only color information can distinguish between two features/objects (e.g. to distinguish a green from a red apple).

The following approaches reported in literature use a combination of color and texture features which makes them relevant in the context of our work. In [5] a combined color and shape histogram is used, outperforming the single descriptors by about 15% higher precision and recall values. In [1] a gradient based descriptor is combined with a color invariance model developed by Geusebroek [7]. In [17] the most popular texture descriptor nowadays, called SIFT (Scale Invariant Feature Transform) descriptor [8], is combined with photometric invariant color histograms. The approach yields a recall value of 21% for SIFT

and 36% for a hue color descriptor. A combined color descriptor based on these two descriptors reaches a recall of 34%. In [12] a combined SIFT and Luv color moments descriptor has been proposed. In contrast the authors of [6] have used an SVM to combine color histograms with a texture descriptor.

3 Descriptor Selection and Combination

3.1 Descriptor Selection

In order to develop an object re-detection system using a combined local feature approach we have selected one texture feature and one color feature that fulfills our requirements best.

Texture Descriptor. Texture descriptors capture the structure of image regions, for instance, the geometric configuration of edges. Examples of texture descriptors include the MPEG-7 EdgeHistogram, TextureBrowsing and HomogenousTexture descriptors, Gabor filter coefficients and SIFT. Extensive studies [10,11,16] have shown that SIFT descriptors outperform other texture descriptors for object recognition on various types of image data, including 3D objects and real world scenes. The SIFT descriptor is based on a multidimensional gradient histogram and has been proposed in [8]. Due to its good invariance to illumination changes as well as perspective changes, this texture descriptor has been successfully used in various works. One of them is the the Video Google system [15] which searches for and localizes all the occurrences of a user outlined object in a video. In fact this is an object re-detection problem and because of the good results of this work we have selected the SIFT descriptor for our system.

Color Descriptor. A large number of color descriptors for images and image regions has been proposed over the last 15 years. The approaches include among others histograms, sets of dominant colors and frequency domain methods. In our combined local feature approach, the color descriptors are extracted from image regions. Thus color descriptors designed for regions rather than for whole images are more likely to yield good results. MPEG-7 is a standard for the description of audiovisual content. It defines several description tools for visual features, including seven color descriptors [9] that make use of the mentioned state of the art methods. We have decided to use MPEG-7 color descriptors for our system after thorough evaluation of available approaches.

In order to select the color descriptor that is most appropriate to be combined with the SIFT descriptor, we have performed an additional evaluation of the three most promising MPEG-7 descriptors: the ColorLayout descriptor, the ColorStructure descriptor and the DominantColor descriptor. For this evaluation, the color descriptors have been extracted from image regions around Difference of Gaussian (DoG) points [8] and both the distance functions recommended in the MPEG-7 standard as well as other proposed alternatives were used in the matching stage. The evaluation has shown that the MPEG-7 ColorLayout descriptor is the most appropriate one for our purpose, the detailed evaluation results can be found in Section 5.

3.2 Descriptor Combination

Apart from the selection of the descriptors, their combination is an important issue. Two major ways of combining different local descriptors exist and they mainly differ in the stage at which the combination is performed.

Direct Combination. Instead of building a descriptor using only one type of visual information (color or texture), a single descriptor is built by combining two or more different descriptors. This can be done through a simple concatenation of the different descriptors or by more sophisticated techniques [12].

Indirect Combination. In this approach the different descriptors are not combined into a single descriptor but a combination during the matching or verification step is performed. Each descriptor from the test image is therefore matched individually against the descriptors in the object database of its own type. All detected matches, regardless of the descriptor type, are then combined for object recognition. It is reasonable to assume that this combination may lead to a higher number of matching descriptor pairs than with the use of a single descriptor type, and thus improvements can be achieved.

A number of approaches to combine different local descriptors are described in literature. Most of them [5,17,12] use direct descriptor combination but work has also been done on indirect descriptor combination [4,6]. Our object re-detection system is also based on an indirect descriptor combination because we assume that color and texture information often complement one another and that the most distinctive descriptors are not always found at the same interest points. There may be points that contain distinctive textural information while others contain distinctive color information. Therefore, in contrast to a direct descriptor combination, it is feasible to use different interest point detectors for different descriptor types. A further benefit of this approach is that if a single descriptor type (color or texture) can provide enough evidence for a recognition, the results of the other descriptor type are needed at most to confirm this recognition.

4 Implementation

4.1 Implemented Approach

To select image regions for the local descriptor extraction, we have used Difference of Gaussian (DoG) points which are located in a full-octave DoG pyramid at intensity extrema [8]. Each DoG point is described by position, scale and orientation information. The image regions for the SIFT descriptors are oriented, squared regions around these DoG points and the size of an image region depends on the scale of the DoG point. For efficiency reasons, we extract the SIFT descriptors directly from the Gaussian pyramid and discard DoG points which are close to an image border. Our SIFT descriptor has 128 real-valued dimensions (16 subregions, 8 orientations) extracted with a trilinear interpolation method. The MPEG-7 color descriptors are also computed from squared image regions

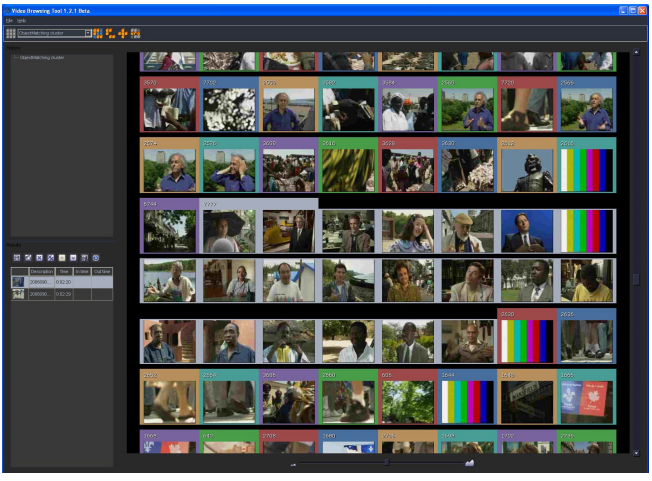


Fig. 1. Object re-detection results are used as a feature for clustering in our video browsing tool

around DoG points but no orientation information is used for their extraction. As descriptors should identify an image region for later matching, it is important that descriptors are as unique as possible for an image. By discarding SIFT and MPEG-7 features, which occur commonly across different objects, the number of descriptors is reduced while the distinctiveness of the object descriptions is improved. Therefore we have implemented the *descriptor filtering* method proposed in [13].

For descriptor matching we use a nearest neighbor distance ratio matching with an additional minimum threshold. The used thresholds and distance ratios have been selected independently for each descriptor type as a result of extensive experiments. As outlined in Section 3.2 we use two different descriptor sets as input for the verification process, one set with SIFT matches and one set with the color descriptor matches. In this verification step a *voting space* is used to decide whether an object appears in an image or not. Each matching descriptor votes for an object position and scale in a three dimensional (x-axis, y-axis and scale) space. By searching for clusters of votes, one or more object occurrences can be detected. Then the number of votes for a certain cluster is examined based on a threshold, which is defined relatively to the total number of votes.

4.2 Application

The approach presented in this paper is used for the re-detection of objects in video collections, audiovisual archives and post-production environments. In order to facilitate navigation and retrieval in not or sparsely annotated content collections, innovative approaches such as content based browsing tools are required. Clustering content by the occurrence of the same object can help to

organize the content collection. For example, in a production environment a lot of unedited material ("rushes") is available, mostly without any annotation. Clustering by occurrence of similar objects yields clusters containing all takes of the same scene and scenes containing similar objects. When the object of interest is the background, this can be used for clustering by the setting of a shot. Figure 1 shows the use of the presented algorithm in our video browsing tool [3]. The object similarities are determined during ingest of new video material into the collection and a cluster tree is built from the object similarities. The tree allows the user to navigate through the content by object similarities.

5 Evaluation

The evaluation both compares the performance of the proposed approach against single descriptor approaches as well as the performance of different color descriptors. The evaluation is performed on two different data sets (see Figure 2): A car data set is used to test the quality of the implemented system for general object re-detection, while a data set containing persons is used to test the performance for non-rigid objects. The comparison of different color descriptors shows that the MPEG-7 ColorLayout descriptor performs best of the tested ones. The average F1 measure on both datasets is 0.70 for the SIFT descriptor, 0.35 for DominantColor, 0.45 for ColorStructure and 0.7 for ColorLayout.

5.1 Car Data Set

In this test, a model of a car occurrence is learned from a single reference image. The model is then used to detect occurrences of the same car in other images. This is demonstrated by searching for a car in a set of 50 images. The data set contains 13 images which show the same car in different views, environments and under different lighting conditions, and 37 images showing other cars and random scenes.

The results for the car data set are shown on the left in Figures 3a and 3b. The combination of a ColorLayout descriptor (with high precision but low recall) with a SIFT descriptor led to a significantly better performance in this test. The improvement of the recall and precision values are supported by the use of high resolution images with no image artifacts, such as compression. The recall value is significantly improved as the texture descriptor responds well to the writing on the car's engine hood and the ColorLayout descriptor to the team colors all around the car. This way different perspective views of the object are detected by different descriptors.

5.2 Person Data Set

Although different algorithms for face recognition exist, the recognition of persons in images remains problematic in cases where a person's face is (partially) occluded or too small for the extraction of features used for face recognition. As

Fig. 2. Example images from the car and person data sets used for evaluation

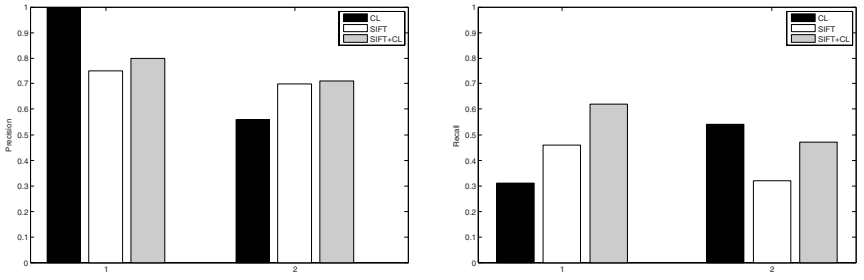
an alternative, it has been proposed to take the visual appearance of the person's body, such as the color and texture of the clothes, into account (cf. [2]). For this reason the developed object re-detection system is tested by searching for the occurrence of a person in a video. The test data set contains video scenes of a person walking inside a building and random shots taken from the TRECVID test data set, which includes different persons in various scenes. The data set contains 742 images.

The results for the person data set are shown on the right in Figures 3a and 3b. The combination of a SIFT descriptor with high precision but low recall and a ColorLayout descriptor with moderate precision and recall led to a descriptor with improved precision and average recall values. The reason for this are the additionally detected matches in comparison to SIFT because of color information and the reduced number of mismatches in comparison to the single ColorLayout descriptor. Object re-detection with SIFT alone doesn't work very well on this test set because persons are real 3D objects, while SIFT descriptors work best when they are extracted from planar surfaces.

The evaluation shows that SIFT performs always better under the influence of view and lighting changes. The color descriptors are more sensitive to these variations. The results also show that SIFT and the color descriptors are often complementary. The ColorLayout descriptor achieves the best results of the evaluated MPEG-7 color descriptors.

5.3 Discussion

The comparison of these results to that of other works is difficult due to the fact that different image sets are used by the different authors. The evaluation results of our approach show a slight increase of precision on both data sets, and a stronger increase of recall. The gain achieved in terms of recall is comparable to some other works which combine descriptors. Our approach has been evaluated on real-world images, similar to those that we would encounter in the intended application areas, such as media monitoring or audiovisual summarization and browsing. For these applications the increase in recall is especially important, as a user can easily discard false positives, while it is difficult to find occurrences of an object that have been missed due to false negatives.



(a) Precision (b) Recall

Fig. 3. Evaluation of (a) precision and (b) recall of the single descriptor approaches using SIFT and the MPEG-7 ColorLayout (CL) descriptor and the proposed combined approach on (1) the car and (2) the person data set

6 Conclusion

In this paper a system for re-detection of occurrences of objects in videos has been presented. Following a local feature based approach, a combination of a SIFT descriptor with a MPEG-7 ColorLayout descriptor has been introduced. The novelty of the approach is the use of a MPEG-7 color descriptor in combination with the SIFT descriptor. The performance of the proposed method has been evaluated by testing it using a car and a person data set.

We have shown that a combined approach using SIFT and ColorLayout descriptors outperforms the use of a single descriptor. In contrast to only using SIFT, especially the recall can be increased for re-detection of multi-colored 3D objects. Low quality images, large view and scale changes and blur influence the object re-detection results negatively.

Acknowledgments. This work has been funded partially under the 6th Framework Programme of the European Union within the IST project "IP-RACINE" (IST-2-511316, http://www.ipracine.org).

References

1. Abdel-Hakim, A., Farag, A.A.: Colored local invariant features for object description. In: Proceedings of the International Conference on Graphics, Vision and Image Processing, pp. 100–106 (2005)
2. Annesley, J., Orwell, J., Renno, J.P.: Evaluation of MPEG-7 color descriptors for visual surveillance retrieval. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation, October 2005 (2005)
3. Bailer, W., Schober, C., Thallinger, G.: Video content browsing based on iterative feature clustering for rushes exploitation. In: Proceedings of TRECVID Workshop November 2006, pp. 230–239, Gaithersburg, MD, USA (2006)
4. Carneiro, G., Lowe, D.G.: Sparse flexible models of local features. In: European Conference on Computer Vision, pp. 29–43, Graz, Austria May 2006 (2006)

5. Diplaros, A., Gevers, T., Patras, I.: Combining color and shape information for illumination-viewpoint invariant object recognition. IEEE Transactions on Image Processing 15(1), 1–11 (2006)
6. Eklundh, J.-O., Björkman, M.: Recognition of objects in the real world from a systems perspective. Künstliche Intelligenz (KI), Special Issue on Cognitive Vision, (April 2005)
7. Geusebroek, J.M., van den Boomgaard, R., Smeulders, A.W.M., Geerts, H.: Color invariance. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(12), 1338–1350 (2001)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., Yamada, A.: MPEG-7 color and texture descriptors. IEEE Trans. Circuits and Systems for Video Technology 11, 703–715 (2001)
10. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
11. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3D objects. In: IEEE Conference on Computer Vision and Pattern Recognition (2004)
12. Quelhas, P., Odobez, J.-M.: Natural scene image modeling using color and texture visterims. In: International Conference on Image and Video Retrieval (2006)
13. Schaffalitzky, F., Zisserman, A.: Automated scene matching in movies. In: Proceedings of the International Conference on Image and Video Retrieval, pp. 186–197. Springer-Verlag, Heidelberg (2002)
14. Schmid, C., Mohr, R.: Local greyvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(5), 530–535 (1997)
15. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, vol. 2 (2003)
16. Sorschag, R.: Object recognition for media monitoring with emphasis on didactical purpose. Master's thesis, Department for Information Technology, Alpen-Adria Universität Klagenfurt, Austria (2006)
17. van de Weijer, J., Schmid, C.: Coloring local feature extraction. In: Proceedings of the European Conference on Computer Vision, pp. 334–348 (2006)

Interactive Boosting for Image Classification

Yijuan Lu[1], Qi Tian[1], and Thomas S. Huan[2]

[1] Department of Computer Science, University of Texas at San Antonio, TX, USA
{lyijuan, qitian}@cs.utsa.edu
[2] Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL,USA
huang@ifp.uiuc.edu

Abstract. Traditional boosting method like adaboost, boosts a weak learning algorithm by updating the sample weights (the relative importance of the training samples) iteratively. In this paper, we propose to integrate feature re-weighting into boosting scheme, which not only weights the samples but also weights the feature elements iteratively. To avoid overfitting problem caused by feature re-weighting on a small training data set, we also incorporate relevance feedback into boosting and propose an interactive boosting called i.Boosting. It merges adaboost, feature re-weighting and relevance feedback into one framework and exploits the favorable attributes of these methods. In this paper, i.Boosting is implemented using Adaptive Discriminant Analysis (ADA) as base classifiers. It not only enhances but also combines a set of ADA classifiers into a more powerful one. A feature re-weighting method for ADA is also proposed and integrated in i.Boosting. Extensive experiments on UCI benchmark data sets, three facial image data sets and COREL color image data sets show the superior performance of i.Boosting over AdaBoost and other state-of-the-art projection-based classifiers.

Keywords: Relevance Feedback, Adaboost, Feature Re-weighting, Multiple Classifiers.

1 Introduction

Recent years have witnessed an explosion of digital images generated from different areas such as commerce, academia and medical institutes. The dramatic increase of images demands efficient indexing and retrieval methods, especially for a large image database. In image retrieval, an image is represented by its image feature vector as a data point in a high-dimensional space. Its dimension ranges from tens to hundreds. However, traditional statistical approaches have difficulties in modeling data directly in such a high dimensional space. Hence, dimension reduction technique plays a critical role in alleviating the high dimensionality problem.

A good dimension reduction method can map the high dimensional data space to a low dimensional space without loss of much useful information. However, any single dimension reduction method cannot find the optimal projection. Traditional techniques, such as Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2], cannot work well when the data distribution cannot be modeled

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 315–324, 2007.
© Springer-Verlag Berlin Heidelberg 2007

as Gaussian or mixture of Gaussians. A better solution is to boost a set of projections and corresponding classifiers first using boosting algorithms. Then combine these boosted classifiers using fusion in the projected space [3].

Here, boosting algorithms are designed to construct a "strong" classifier from a "weak" learning algorithm and present the superior result given by a thresholded linear combination of the weak classifiers. AdaBoost [4] is often regarded as the generic boosting algorithm. The basic idea of AdaBoost is to iteratively re-weight the training samples based on the outputs of some weak learners. Misclassified samples will receive higher weights in the next iteration. This forces the classifier to focus more on the incorrectly classified examples.

However, during this procedure, only weights of samples are updated. It doesn't update any feature element weight, which is important and very useful especially for image databases using high dimensional image features [5]. In this paper, we incorporate feature re-weighting into boosting and propose a new feature re-weighting approach for Adaptive Discriminant Analysis (ADA) [3]. In addition, considering feature re-weighting on small training data set tends to bias to the training set and causes overfitting, we integrate user feedback into boosting scheme and propose a novel interactive boosting framework (i.Boosting). i.Boosting not only weights the samples, but also weights the feature elements iteratively. Besides, in i.Boosting, relevance feedback provides boosting with more misclassification information. And better than simple relevance feedback, Adaboost forces classifiers to pay more attention to wrongfully predicted samples in user feedback.

In this paper, i.Boosting is implemented using ADA as base classifiers. It not only enhances but also combines multiple ADA classifiers into a more powerful one. Extensive experiments on the UCI benchmark data sets, three facial image data sets and COREL color image data sets show that the superior performance of i.Boosting.

In section 2, feature re-weighting and relevance feedback techniques will be briefly described. Interactive boosting scheme and i.Boosting with ADA are illustrated in section 3. Experimental results and conclusions will be given in section 4 and 5 respectively.

2 Feature Re-weighting and Relevance Feedback

2.1 Classic Feature Re-weighting

In image database, each image $i \in I$ is represented by its M features $\boldsymbol{f}_i = [f_{i1}, f_{i2}, \cdots f_{iM}]^T$. Let the feature of query image q be $\boldsymbol{f}_q = [f_{q1}, f_{q2}, .. f_{qm}]^T$, the Euclidean distance between query image and the image in the database is

$$d = (\boldsymbol{f}_i - \boldsymbol{f}_q)^T W (\boldsymbol{f}_i - \boldsymbol{f}_q)$$

W is the feature weighting matrix indicating the importance of each component of features. After relevance feedback, the user provides the relevance of each image to the query and the feature weights can be updated to make similar images close to each other and dissimilar images far away from each other. Traditional feature re-weighting methods are based on distance metric, *e.g.*, generalized Euclidean distance

[5]. In this paper, we apply a dynamic feature re-weighting method before dimension reduction in order to obtain a better projection after each iteration in relevance feedback.

2.2 Adaptive Discriminant Analysis

In our paper, we use Adaptive Discriminant Analysis (ADA) [3] as our dimension reduction method. It can provide an accurate model of the complex distribution for positive and negative samples by finding an optimal projection in the following way:

$$W_{ADA} = \arg\max_{W} \frac{|W^T[\lambda S_{P->N} + (1-\lambda)S_{N->P}]W|}{|W^T[\eta S_P + (1-\eta)S_N]W|} \tag{1}$$

in which

$$S_{N->P} = (\boldsymbol{f}_N \boldsymbol{f}_N{}^T) .* \sum_{i \in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{2}$$

$$S_{P->N} = (\boldsymbol{f}_P \boldsymbol{f}_P{}^T) .* \sum_{j \in Positive} (\boldsymbol{x}_j - \boldsymbol{m}_N)(\boldsymbol{x}_j - \boldsymbol{m}_N)^T \tag{3}$$

$$S_P = (\boldsymbol{f}_P \boldsymbol{f}_P{}^T) .* \sum_{j \in Positive} (\boldsymbol{x}_j - \boldsymbol{m}_P)(\boldsymbol{x}_j - \boldsymbol{m}_P)^T \tag{4}$$

$$S_N = (\boldsymbol{f}_N \boldsymbol{f}_N{}^T) .* \sum_{i \in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{5}$$

The $\boldsymbol{m}_P$ and $\boldsymbol{m}_N$ are the means of positive and negative samples, respectively. $\boldsymbol{f}_P$ and $\boldsymbol{f}_N$ are feature element weights of positive and negative samples, respectively. $.*$ stands for Hadamard product operation. S_P (or S_N) is the within-class scatter matrix for the positive (or negative) examples. $S_{N->P}$ (or $S_{P->N}$) is the between-class scatter matrix from the negative (or positive) examples to the centroid of the positive (or negative) examples. The two parameters $\lambda \in (0,1), \eta \in (0,1)$ control the bias between positive and negative samples. Proper setting of parameters may fit the real distribution of data better than LDA and PCA [3].

However, to find an optimal setting, exhaustive searching in 2D parameter (λ, η) space is needed and computationally expensive. Boosting can alleviate this problem by enhancing and combing a set of weak ADA classifiers into a more powerful one. (An ADA classifier is denoted as ADA projection and a base classifier in the projected space)

For each weak ADA classifier, to find a better projection, the ratio of $\frac{trace(\lambda S_{P->N} + (1-\lambda)S_{N->P})}{trace(\eta S_P + (1-\eta)S_N)}$ need to be maximized. Intuitively, it is to minimize the "within-class scatter" and maximize the "between-class scatter". Therefore, the criterion can be redefined to maximize

$$\begin{aligned} &trace(\lambda S_{P->N} + (1-\lambda)S_{N->P}) - trace(\eta S_P + (1-\eta)S_N) \\ &= trace(\lambda S_{P->N} - \eta S_P) + trace((1-\lambda)S_{N->P} - (1-\eta)S_N) \end{aligned} \tag{6}$$

Hence, re-weighting scheme for ADA is to update $\boldsymbol{f}_P$ by maximizing $trace(\lambda S_{P->N} - \eta S_P)$ and update $\boldsymbol{f}_N$ by maximizing $trace((1-\lambda)S_{N->P} - (1-\eta)S_N)$ based on (2), (3), (4), and (5).

2.3 Relevance Feedback

To efficiently incorporate user feedback and enhance the retrieval accuracy, relevance feedback can also be integrated in the boosting.

Relevance Feedback was initially developed in document retrieval [6] and widely applied in content-based image retrieval (CBIR) [7, 8]. A challenge in CBIR is the *semantic gap* between the high-level semantics in human mind and the low-level computed features (such as color, texture, and shape). In order to bridge the gap between low-level features and high-level semantics, relevance feedback is introduced.

The basic idea of relevance feedback is to get human in the loop. At first, computer processing provides initial retrieval results. Users are then asked to evaluate the current retrieval results according to degrees that are relevant or irrelevant to his/her request. The system then applies the user's feedback to update the training examples to improve performance for the next round. This learning process can be applied iteratively if the user desires. Relevance Feedback algorithms have been shown to provide dramatic performance improvement in image retrieval systems [8].

3 Interactive Boosting

3.1 Methodology

Motivated by the strength and success of adaboost, dynamic feature re-weighting and Relevance Feedback, we propose an interactive boosting framework called

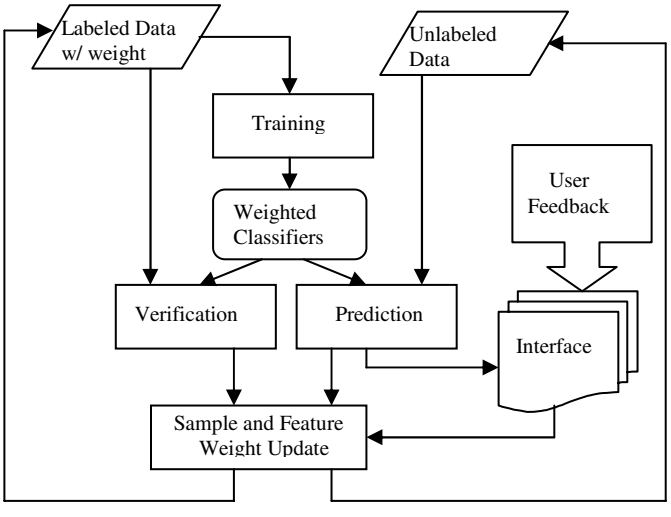


Fig. 1. Interactive Boosting framework

i.Boosting. It can integrate user relevance feedback, adaboost (sample re-weighting) and feature re-weighting in the loop of boosting and better bridge the gap between semantic concept and image features. Figure 1 gives an illustration of the basic idea of the interactive boosting framework.

3.2 Interactive Boosting in ADA

Based on the framework described in section 3.1, the brief algorithm below shows how the i.Boosting is implemented with multiple ADA classifiers.

Algorithm: i.Boosting with ADA as weak classifiers

Input: Labeled Sample Set $\boldsymbol{X}$ and label Y, Unlabeled Sample Set $\boldsymbol{U}$
Feature vector D and Feature element d
K ADA classifiers with different (λ, η)
M: The dimension of feature (feature size), p: positive samples, n: negative samples
T: The total number of runs

Initialization: sample weight $w_{k,t=1}(x) = 1/|X|$ and feature weight $f_{p,t=1}(d) = 1, f_{n,t=1}(d) = 1$

Boosting

For each classifier $k = 1, \ldots, K$ do

For $t = 1, \ldots, T$

- Train each ADA classifier on labeled samples with weights. Note that $\sum_{x \in X} w_{k,t}(x) = 1$,

  $$\sum_{d \in D} f_{p,t}(d) = \sum_{d \in D} f_{n,t}(d) = M$$

- Get the probability-rated prediction on labeled and unlabeled sample.
- Compute the weights of classifiers based on its classification error rate $\varepsilon_{k,t}$ on labeled samples $\alpha_{k,t} = \frac{1}{2} \ln(\frac{1-\varepsilon_{k,t}}{\varepsilon_{k,t}})$
- Present samples from the unlabeled data set with their predicted labels to user
- Obtain user feedback on the ground truth labels
- Construct new labeled training set by adding data and corresponding labels obtained from user feedback
- Update the weight of all training samples $w_{k,t+1}(x) = w_{k,t}(x) \exp(-\alpha_{k,t} \cdot h_{k,t}(x) \cdot y)$
- Compute the new $|\lambda S_{P \to N} - \eta S_P|$ and $|(1-\lambda) S_{N \to P} - (1-\eta) S_N|$ in eq. (6)
- Update the weight of features $f_{p,t}(d), f_{n,t}(d)$ accordingly

End for t

End for each classifier

The final prediction $H(x) = sign(\sum_{k,t} \alpha_{k,t} h_{k,t}(x))$, using sum rule to combine multiple classifiers

4 Experiments and Analysis

In this section, we experimentally evaluate the performance of the interactive boosting on benchmark dataset, image and face classification. The test data sets include UCI benchmark data sets, COREL image data set and three popular face image data sets, which cover a wide range of data in computer vision applications. In

order to comprehensively evaluate the performance of our proposed method, we compare it with AdaBoost, ADA with Relevance Feedback and other state-of-the-art projection techniques. In all experiments, our boosted ADA is trained on 36 ADA classifiers with (λ, η) evenly sampled from 0 to 1 with step size of 0.2. Bayesian classifier is used on the projected data for all projection-based methods. In order to have statistical analysis of our scheme, we perform a pseudo relevance feedback. At each relevance feedback, 5 images are fed to the system automatically based on their ground truth labels. In all experiments, average prediction error rate of 50 runs is reported.

4.1 Interactive Boosting on UCI Data Set

First, we tested the effectiveness of the proposed i.Boosting on benchmark data sets from UCI repository. For comparison purpose, four independent experiments are designed and implemented to compare i.Boosting with other related variants.

Due to the limited space, we only show the results on SPECTF heart databases, which describe diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Similar results are obtained on other data sets. This SPECTF data set contains 267 instances (patients) and totally 43 attributes. Each of the patients is classified into two categories: normal and abnormal. The sizes of the training set and testing set are 80 and 187, respectively. The average error rate across five iterations is plotted in Fig. 2, where the *x*-axis denotes the iteration number (between 0 to 5). 0 stands for the starting status before iterations begin.

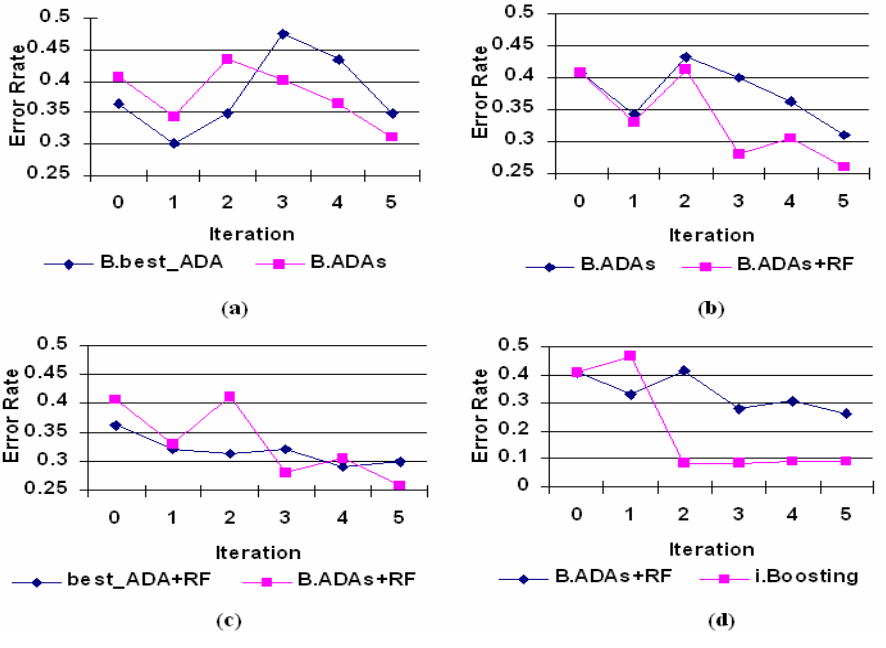


Fig. 2. Comparison between i.Boosting and other related variants

a) *boosting single ADA classifier vs boosting multiple ADA classifiers*
In this experiment, we compare a boosted single best ADA classifier (B.best_ADA) with boosted multiple ADA classifiers (B.ADAs). The boosted multiple ADA classifiers are trained on 36 ADA classifiers with (λ, η) evenly sampled from 0 to 1 with step size of 0.2. The single best ADA is the best one chosen from these 36 classifiers. Showed in Fig. 2 (a), as iteration goes on, the error rates of the B.best_ADA classifier and B.ADAs decrease. Although B.ADAs starts with a set of weak classifiers, after three iterations (*T=3*), it outperforms the B.best_ADA. It verifies that AdaBoost provides a general way of combining and enhancing a set of ADA classifiers in the parametric space.

b) *boosting multiple ADA classifiers with and without relevance feedback*
Secondly, we evaluate the effect of integrating user feedback into boosting scheme. From Fig.2 (b), we can find the performance improvement of using AdaBoost alone (B. ADAs) is less than that of boosted ADAs with relevance feedback (B.ADAs+RF). The performance of B.ADAs+RF is consistently better than that of B.ADAs (w/o relevance feedback) by up to 30.4% on SPECTF heart set. Obviously, the reason is that user feedback and human judgement could be accumulated iteratively to facilitate learning process.

c) *single ADA classifier+RF (w/o boosting) vs boosting multiple ADA classifiers+RF*
The third experiment is designed to verify if the performance improvement of B.ADAs+RF is introduced by relevance feedback only. Hence, we compare the single best ADA classifier with only relevance feedback (best_ADA+RF) and boosted multiple ADA classifiers with relevance feedback (B.ADAs+RF). From the experimental result in Fig. 2 (c), we can conclude that: 1) B.ADAs+RF and Relevance Feedback only starts with similar performance in iteration 1; 2) After several iterations, simple relevance feedback gain less performance improvement than B.ADAs+RF. In conclusion, B.ADAs+RF has obvious advantage over the simple relevance feedback method in that the classifiers are trained to pay more attention to wrongfully predicted samples in user feedback through a reinforcement training process.

d) *boosting multiple ADA classifiers+RF(w/o feature re-weighting) vs i.Boosting*
The last experiment is to evaluate the performance of feature re-weighting in interactive boosting. In Fig. 2 (d), we can find after two iterations, the i.Boosting performed much better than B.ADAs+RF (without feature re-weighting). Besides, i.Boosting becomes much steadier after several iterations. It is clear that our method i.Boosting boosts not only a set of weak classifiers but also the individual features.

4.2 Interactive Boosting for Image Classification

In order to evaluate interactive boosting for image classification, we test i.Boosting on COREL image databases. This database contains 1386 color images, which are categorized into 14 classes. Each class contains 99 images. Each image is represented by 37 feature components including color moments [9], wavelet-based texture [10] and water-filling edge-based structure features [11]. For simplicity, we randomly pick

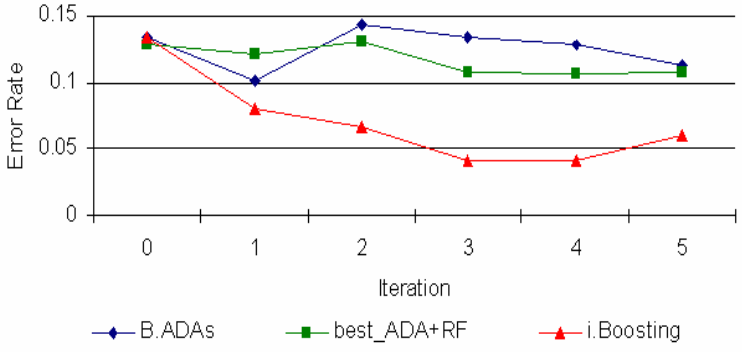


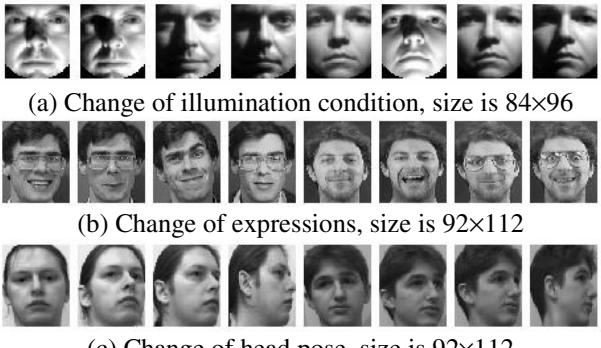
Fig. 3. i.Boosting on COREL data set

up two classes of images for classification. One-third of the images are used for training while the rest are used for testing.

The experimental result showed in Fig. 3 is consistent with the results on the UCI data set. i.Boosting, boosted multiple ADA classifiers (without relevance feedback) and the best ADA classifier with relevance feedback start with similar performance in iteration 1. But as iteration goes on, i.Boosting gains much better performance improvement than the other two. It demonstrates that interactive boosting exploits the favorable attributes of adaboost, feature re-weighting and relevance feedback well.

4.3 Interactive Boosting for Face Classification

To evaluate interactive boosting for face classification, we tested i.Boosting on three well-known face image databases with change of illumination, expression and head pose, respectively. The Harvard Face image database contains images from 10 individuals, each providing total 66 images, which are classified into 10 sets based on increasingly changed illumination condition [12]. The AT&T Face Image database [13] consists of grayscale images of 40 persons. Each person has 10 images with



(a) Change of illumination condition, size is 84×96

(b) Change of expressions, size is 92×112

(c) Change of head pose, size is 92×112

Fig. 4. Example Face images from three facial databases

different expressions, open or closed eyes, smiling or non-smiling and wearing glasses or no glasses. The UMIST Face Database [14] consists of 564 images of 20 people, which covers a range of poses from profile to frontal views. Figure 4 gives some example images from the databases. Sixty image features are extracted to represent these images including histogram, wavelet-based texture and water-filling edge-based structure features.

For each database, we randomly chose one person's face images as positive and the rest face images of others are considered as negative. For comparison purpose, 6 state-of-the-art projection-based techniques: Eigenface [12], LDA, BDA [15], DEM [16], KDEM [16], ADA [3] are tested on the same databases. To be consistent, the results for these techniques are obtained after 5 iterations of relevance feedback.

The results are listed Table 1 with the smallest error rate in bold. It is clear that i.Boosting performs best in 4 out of 5 tests and second to the best ADA in one test. It is clear that i.Boosting provides more robustness to the changes of illumination, expression and pose than other techniques.

Table 1. Comparison of i.Boosting with state-of-the-art techniques on three different face databases

Error Rate (%)		Harvard Dataset			ATT Dataset	UMIST Dataset
		Subset 1	Subset 2	Subset 3		
Methods	Eigenface	6.33	9.1	4.16	0.31	3.81
	LDA	15.06	15.17	15.33	2.07	0.51
	BDA	1.42	4.0	1.43	0.83	1.36
	DEM	14.96	15.18	15.26	3.35	1.28
	KDEM	11.21	13.33	11.18	1.67	2.64
	Best single ADA	0.33	**2.7**	0.84	0.04	0.17
	i.Boosting	**0.16**	3.0	**0.58**	**0.02**	**0.11**

5 Conclusion and Future Work

In this paper, we propose a novel interactive boosting framework to integrate feature re-weighting and relevance feedback into standard boosting scheme. Compared to the traditional boosting scheme, the proposed method updates both sample weights and feature weights iteratively. It obtains more performance improvement from the relevance feedback by putting human in the loop to facilitate learning process. It has obvious advantage over the simple relevance feedback method in that the classifiers are trained to pay more attention to wrongfully predicted samples in user feedback through a reinforcement training process. All experimental results show the superior performance of the proposed interactive boosting framework.

Our future work includes testing different user feedback schemes such as active learning techniques [17] in the interactive boosting. Different base classifiers and

their corresponding feature re-weighting schemes will be implemented. We will also evaluate the performance difference among different boosting and fusion schemes.

Acknowledgement

This work was supported in part by Army Research Office (ARO) grant under W911NF-05-1-0404, by Department of Homeland Security (DHS) and by the San Antonio Life Science Institute (SALSI).

References

1. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer-Verlag, New-York (2002)
2. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. John Wiley & Sons, Inc. (2001)
3. Yu, J., Tian, Q.: Adaptive Discriminant Projection for Content-based Image Retrieval. In: Proc. of Intl. Conf. on Pattern Recognition August 2006, Hong Kong (2006)
4. Freund, Y., Schapire, R.: A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
5. Wu, Y., Zhang, A.: A Feature Re-weighting Approach For Relevance Feedback. In: Image Retrieval. IEEE International Conference of Image Processing (2002)
6. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill Book Company, New York (1992)
7. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool in interactive content-based image retrieval. IEEE Trans Circuits and Systems for Video Tech. 8(5), 644–655 (1998)
8. Zhou, X., Huang, T.S.: Relevance feedback in image retrieval: a comprehensive review. ACM Multimedia Systems Journal, special issue on CBIR 8(6), 536–544 (2003)
9. Stricker, M., Orengo, M.: Similarity ofcolor images. In: Proceedings of SPIE Storage and Retrieval for Image and Video Databases, San Diego, CA (1995)
10. Smith, J.R., Chang, S.F.: Transform features for texture classi.cation and discrimination in large image database. In: Proceedings of IEEE International Conference on Image Processing, Austin, TX (1994)
11. Zhou, X.S., Rui, Y., Huang, T.S.: Water-.lling algorithm: a novel way for image feature extraction based on edge maps. In: Proceedings of IEEE International Conference on Image Processing, Kobe, Japan (1999)
12. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. PAMI 19(7) (July 1997)
13. Rowley, H. A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. PAMI, 20 (1998)
14. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: IEEE Workshop on Applications of Computer Vision December 1994, Sarasota FL (1994)
15. Zhou, X., Huang, T.S.: Small sample learning during multimedia retrieval using biasMap IEEE CVPR (2001)
16. Tian, Q., Wu, Y., Yu, J., Huang, T.S.: Self-supervised learning based on discriminative nonlinear features for image classification. Pattern Recognition, Special Issue on Image Understanding for Digital Photographs, 38 (2005)
17. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: Proc. of ACM Int'l. Conf. Multimedia, pp. 107–118 (2001)

FBSA: A Self-adjustable Multi-source Data Scheduling Algorithm for P2P Media Streaming

Shuangjia Chen[1,3], Longshe Huo[2], Qiang Fu[2], Rui Guo[1,3], and Wen Gao[2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, 100080 Beijing, China
[2] Institute of Digital Media, Peking University, 100871 Beijing, China
[3] Graduate School, Chinese Academy of Sciences, 100039 Beijing, China
{sjchen, lshuo, qfu, rguo}@jdl.ac.cn, wgao@pku.edu.cn
http://idm.pku.edu.cn

Abstract. Recently, there is great interest of using P2P network in multi-source media streaming. In this paper, we present a self-adjustable scheduling algorithm for multi-source P2P media streaming, called feedback based scheduling algorithm (FBSA). FBSA is simple to implement, fully distributed, asynchronous, and can afford better quality of service. Furthermore, it can adjust itself automatically to adapt network heterogeneity, load variety, bandwidth fluctuation, and overlay dynamics. We have implemented FBSA, and tested it in a P2P live media streaming system: AVStreamer. Experimental results show that FBSA outperforms other conventional P2P data scheduling algorithms, especially in heterogeneous networks.

Keywords: peer-to-peer, media streaming, data scheduling.

1 Introduction

During recent years, Peer-to-Peer (P2P) network has attracted tremendous attention from both academy and industry. In a P2P system, nodes communicate directly with each other for the sharing and exchanging of data as well as other resources such as storage and CPU capacity, each peer acts both as a client who consumes resources from other peers, and also as a server who provides service for others.

Nowadays, various P2P systems, especially live media streaming systems, have been proposed and developed. Each system constructs an overlay network, such as data-driven overlay network (DONet) [1-3], source-driven overlay network [4-5], and receiver-driven overlay network [6]. Among these systems, DONet has achieved great success due to its robustness to high churn of nodes. DONet forms an unstructured network, and its robustness has been verified in a systematical study [7]. Some practical operational DONet based systems, such as PPlive [8], PPStream [9], and Gridmedia [10], have reported that they can enable several 100,000 users to watch a hot live media simultaneously with bit rate of 300-500 Kbps through only one streaming server, whose bandwidth is merely 5-10 Mbps.

In DONet based systems, the streaming media is partitioned into a series of data blocks, and the exchanging of blocks between different nodes relies mainly on two

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 325–333, 2007.
© Springer-Verlag Berlin Heidelberg 2007

important steps: overlay construction and data scheduling. In the first step, each node randomly selects some neighbors as its partners so as to form an unstructured overlay network. When a node receives a new data block, it first updates its own buffer, and then exchanges the buffer occupation information with all its partners. In the second step, each node periodically requests its absent blocks from some of its partners, and sends the blocks already cached in its buffer to other partners who request it. The performance of the system relies on the algorithms adopted in the two steps. Many existing works have proposed different gossip-style algorithms to construct unstructured overlay to enhance its randomicity [11-12], however fewer attentions were paid to the data scheduling algorithm which is also very important for the overall performance.

Based on previous works, recently we also developed a system named AVStreamer [13], which is the first P2P live media streaming system supporting the Chinese AVS standard (Audio Video coding Standard). AVStreamer adopted the same data-driven overlay like DONet, however, lots of new contributions have been put forward. In AVStreamer, we regard a block as a GOP (Group of Pictures) of a video stream, which starts with an I-frame, and follows by a series of P- and B-frames. In each node, a fix-sized (default 120) buffer window is designed to cache the GOPs received from other nodes. With the incoming of new GOPs, the buffer window slides forward continuously. The availability of each GOP cached in the buffer window of a node is represented by a 120-bit vector called Buffer Map (BM). In this paper, we put our emphasis mainly on the P2P data scheduling algorithm used in AVStreamer, and present a new feedback based scheduling algorithm (FBSA) which is simple, distributed and asynchronous. FBSA can adjust itself automatically according to network heterogeneity, variety of the load, bandwidth fluctuation, and overlay dynamics (i.e., nodes frequently subscribe or unsubscribe).

The rest of this paper is organized as follows. In Section 2, we briefly present the related work. In Section 3, we describe the FBSA in detail and also provide the pseudo code with some analysis. In Section 4, we evaluate the performance of FBSA through experimental results. Finally, we conclude this paper and present future plans in Section 5.

2 Related Work

There have been some research efforts toward P2P data scheduling. Until now, some strategies, such as random, rarest-first (RF), and round-robin (RR), have been proposed and adopted in many existing media streaming systems. In this section, we give a brief description about these strategies.

Random is a very simple P2P data scheduling strategy adopted by Chainsaw [3]. According to this strategy, for every absent media block, each node randomly selects a partner from multiple partners who hold the block, and then request it from the selected partner. This strategy can be used in DONet based systems, but its performance is not always steady especially in heterogeneous network environment.

Round-robin (RR) is used by PALS [2] for layered media streaming. According to this strategy, all the requested blocks are assigned to one partner proportionally

in a determinate order (namely round-robin). If only one partner holds the block, then it is requested from that partner; otherwise, it is requested from a partner who has the maximum available bandwidth. This strategy can achieve good balance of load.

Rarest-first (RF) is a heuristic algorithm with fast response time, and was proposed in CoolStreaming [1]. The main ideas of this strategy include: requesting blocks with less potential suppliers preemptively; among multiple potential suppliers, selecting the one with highest superfluous bandwidth and enough available time. However, this strategy has some inherent disadvantages that we will analyze in detail in the next section.

3 FBSA: A Self-adjustable Scheduling Algorithm

In this section, we propose a self-adjustable distributed P2P data scheduling algorithm: feedback based scheduling algorithm (FBSA). In AVStreamer, we index each media block (GOP) with a serial integer, which is called GOP sequence number. The GOP with smaller sequence number should be rendered at the client side before the one with greater sequence number. To afford better QoS (Quality of Service), we assume that the GOP with smaller sequence number has higher priority than the one with greater sequence number. Our goal is to minimize the start delay time and achieve more fluent rendering experiences under certain constraints, such as heterogeneous network topology, fluctuant bandwidth, and dynamic overlay, etc.

3.1 Motivation

In a DONet based P2P media streaming system, it is demanded to have an intelligent scheduling algorithm. The reasons can be summarized as follows. First, since each node has multiple direct data providers (named partners), there are multiple choices when requesting a GOP from these partners. Second, since the network is dynamic and heterogeneous, even for the same peer, its bandwidth cannot always be determined with the time elapses. Third, in real situation, each GOP should be obtained before its playback deadline, however, this target cannot always be achieved; in the condition of bandwidth shortage, scheduling is needed to keep the number of GOPs missing deadline as minimal as possible.

As mentioned in CoolStreaming [1], the scheduling problem is a variation of the parallel machine scheduling, which is known as NP-hard [14]. Thus in a system of real-time media streaming, heuristic algorithms which can achieve approximate optimal scheduling are helpful. Recently, researchers have proposed several algorithms. Chainsaw [3] adopted the random strategy, which is simple yet not steady. Especially in heterogeneous networks, a poor choice may cause a block never arriving its destination even the available bandwidth is abundant. The round-robin strategy used by PALS [2] can achieve better load balance; however, neither it acts so well in heterogeneous networks, since the out-bound bandwidth of each partner is not equivalent.

CoolStreaming [1], in which the DONet was first introduced, proposes the rarest-first strategy. However it performs not so well for streaming applications with

real-time constraint. In AVStreamer, there are two situations that a GOP has few potential suppliers. First, the GOP has a smaller sequence number and it has slid out of the buffer windows in some partners. Second, it is a very new GOP with almost greatest sequence number and has not been spread widely, thus only the media source node and its direct partners hold this GOP temporarily. For the first situation, it is suitable to request GOP with fewer potential suppliers first using the rarest-first algorithm. However for the second situation, particularly in the condition that the bandwidth is not abundant, requesting GOP with fewer potential suppliers first may result that the GOPs with smaller sequence number will never be requested before their playback deadlines. Besides, the rarest-first strategy needs to know the exact bandwidth of each partner in advance, which is hard to measure in real-time.

Facing these problems, we propose a new self-adjustable scheduling algorithm: FBSA, and implement it into AVStreamer.

3.2 Feedback Based Scheduling Algorithm

To minimize the start delay time and get more fluent rendering experience, the scheduling algorithm should try to obtain the GOP with smaller sequence number first. At the same time, the algorithm must be able to adapt highly dynamic network conditions and can achieve balance load. To meet these requirements, we present a feedback based scheduling algorithm (FBSA) with the characteristic of self-adjustable.

In FBSA, we assume that the GOP with smaller sequence number has higher priority. Based on this assumption, we request a GOP only if its sequence number is smaller than anyone else and has at least one supplier. Otherwise, if a GOP has the smallest sequence number but has no suppliers, it will be inserted into a set called *prepare_set* at the moment. During each loop of the scheduler, we first consider the GOPs in *prepare_set* and then the other GOPs. If there are multiple potential suppliers for a GOP, then the supplier node with minimum *reqNum* is selected. If there are still two or more suppliers with the same minimum *reqNum*, then the one whose IP address matches best to the IP of the current node is selected. In the above statements, *reqNum* is a predefined variable for each partner and is initialized to zero. For each partner, its *reqNum* increases by one when a GOP is requested from it and decreases by one when it has sent a GOP out successfully. By this way, the scheduling strategy is adjusted by itself based on the actual status of media data transmission. This is also the reason why the algorithm is named Feedback Based Scheduling Algorithm. According to FBSA, each node can always request data from a partner who is capable of providing data more quickly. In other words, the partners who have abundant resources (for example bandwidth) can always provide more data than others.

The pseudo code of FBSA running at each peer is shown in Fig.1. Its compute complexity is bounded by $O(W \cdot N)$, where W is the maximum number of GOPs cached in the buffer widow, and N is the number of partners of this node.

Since a partner may unsubscribe at any moment and its access bandwidth may fluctuate with time, we need a re-try mechanism to re-request the GOP which has

```
Input:
 bm[s]:        buffer map of partner s;
 reqNum[s]:    number of GOPs this node has requested from s but has not received yet;
 maxReq:       max number of GOPs this node can request from a partner at the same time;
 currGop:      the sequence number of the current absent GOP to be requested;
 partner_set:  set of partners of this node;
 prepare_set:  set of GOPs whose sequence numbers are smaller than currGop, but has not
               been requested due to that none of its partners holds it then.
Scheuler:
  while true do
      { erase the GOPs missing deadlines in prepare_set }

      //in this loop, i is fetched in ascendant order by sequence number
      for each i ∈ prepare_set do
            supplier ← arg_s{ reqNum [s] <= reqNum [s'] | bm[s,i] = 1, bm[s',i] = 1,
                        reqNum [s] < maxReq, reqNum [s'] < maxReq, s, s' ∈
                        partner_set};
            if supplier <> null then
               prepare_set ← prepare_set – {i};
               requestGop ← i;
               break for i;
            end if;
      end for i

      if supplier = null then
            if currGop > { the maximum sequence number of GOP in local buffer
                        window } then
               break while;
            else
               requestGop ← currGop;
               supplier ← arg_s{ reqNum [s] <= reqNum [s'] | bm[s,currGop] = 1,
                           bm[s',currGop] = 1, reqNum [s] < maxReq, reqNum [s'] <
                           maxReq, s, s' ∈ partner_set};
               if supplier = null then
                        prepare_set ← prepare_set ∪ {currGop};
               end if;
               currGop ← currGop + 1;
            end if;
      end if;

      if supplier <> null then
            { request GOP of (requestGop) from (supplier) };
            reqNum [supplier] ← reqNum [supplier] + 1;
      end if;
  end while
```

Fig. 1. Pseudo code of FBSA

been requested but has not received for a long time. However, too much re-requests may cause network congestion and result in more GOPs missing their playback deadlines. To avoid this phenomenon, we define a variable *maxReq*, which indicates the max number of GOPs that this node can request from a partner at the same time, as shown in the pseudo code. This variable can be adjusted according to the average size of GOPs. If the size is great, then *maxReq* should be decreased to avoid network congestion due to redundant re-requests; otherwise if the size is small, then *maxReq* should be increased to took full advantage of the network bandwidth and avoid frequently pauses of the client side media player.

3.3 Analysis

As mentioned above, in each node, FBSA selects appropriate data suppliers based on the actual status of data transmission, thus it is especially efficient for heterogeneous networks. In this section, we provide an example to show how FBSA works intuitively.

An actual illustration of scheduling problem is shown in Fig.2. The node in the center has three partners, and the numbers beside the arrows represent the current available bandwidth from that partner to this node. The numbers in the rectangles beside each partner represent the GOPs it holds now. The IP address of partner 2 matches best to the central node, next partner 1, and the last partner 3.

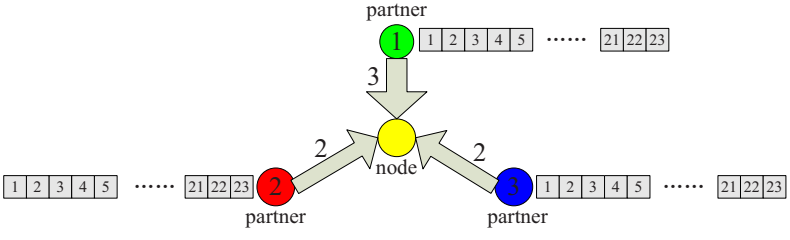


Fig. 2. Illustration of scheduling problem

Fig.3 shows the scheduling timing diagram of this node which adopts FBSA. To make the diagram more concise, in this example we set *maxReq* as 2. The number on dashed arrow stands for the sequence number of the GOP requested from the partner. As shown in Fig.3, this node requests at most *maxReq* (namely 2) GOPs from each partner at the same time, then it will request a new GOP from a partner whenever this partner has sent a pre-requested GOP out successfully. The partners with higher available bandwidth spend less time sending a GOP, so we will request more GOPs from these partners. During this period, this node requested 9, 6, 6 blocks from partner 1, 2 and 3 respectively, and the ratio is just consistent with the ratio of their current available bandwidth. From this figure we can see that FBSA results in an approximate optimal scheduling.

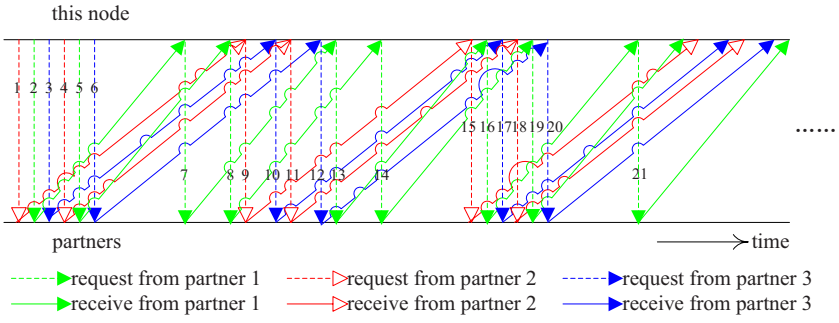


Fig. 3. Scheduling timing diagram

4 Performance Evaluation

In this section, we evaluate FBSA and compare it with other three P2P data scheduling algorithms: random, round-robin, and rarest-first.

As aforementioned, our goal is to minimize the start delay time, thus we consider the time used to buffer the same amount of media data as a main performance criterion for each algorithm. For a fair comparison, we implement all the algorithms in our AVStreamer system, and they rely on overlay networks constructed by the same receiver-driven gossip algorithm. Furthermore, each experimental result is an average value of no less than 10 similar runs. In our experiment, the overlay has 50 nodes and each node has 3 to 6 partners. The scheduling algorithm is executed periodically, and we set the period as 1 second. The average bit rate of the media stream used in our experiment is 420 Kbps, and the size of each GOP is about 180K bytes. The value of *maxReq* is set as 5.

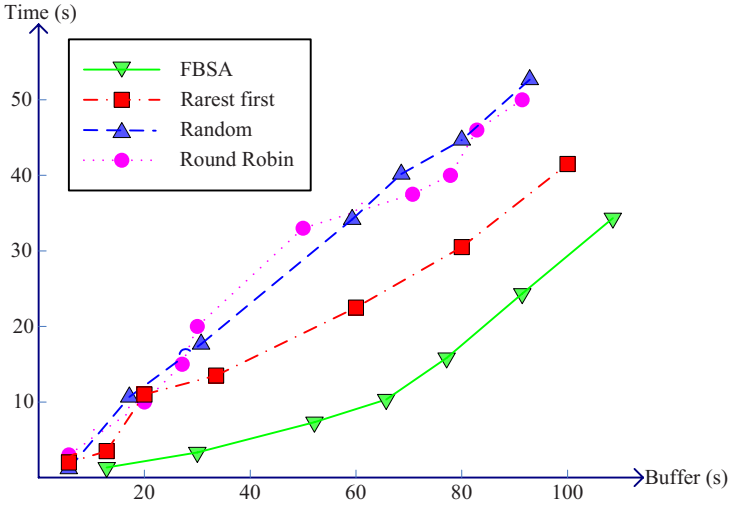


Fig. 4. Time used to buffer media data

As shown in Fig.4, we plot the curves of FBSA and other three algorithms. The horizontal axis denotes the amount of buffered data (in second that the data can be played), while the vertical axis denotes the actually buffering time (in second). From this figure we can see that FBSA outperforms other algorithm obviously. For example, to buffer 60-second data, FBSA spends about 10 seconds, RF spends 20 seconds, RR and Random spends 30 seconds. The curve of RF fluctuates at the beginning since it requests some relative new blocks as analyzed in section 3. The Random strategy and the RR strategy have the poorest performances because they can not adapt themselves to heterogeneous networks. There is a variation of RR called Weighted Round-Robin (WRR), which may perform even better than RF. However, WRR relies on the exact bandwidth of each partner, thus we do not simulate it here.

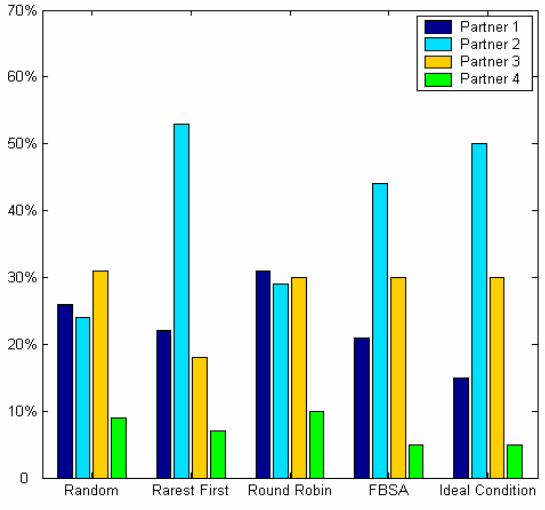


Fig. 5. Scheduling from heterogeneous partners

Fig.5 shows a scheduling histogram of a node with 4 partners. The height of the column indicates the percentage of the number of blocks obtained from a partner to the total number of blocks this node has received. The "Ideal Condition" columns show the relative bandwidth of each partner, and the other four groups of columns show the actual scheduling histogram. We note that the FBSA columns are the most similar to the "Ideal condition" columns, then the RF, RR and Random. This also verifies that FBSA performs better than other strategies, especially in heterogeneous networks.

The above experimental results show that FBSA can shorten the start delay time and make the player play more fluent under any constraints. Particularly, under the condition that the bandwidth of each partner is not abundant but their aggregate bandwidth is sufficient for full-quality streaming, FBSA can guarantee better QoS than others. At the same time, FBSA is fully distributed, and it can achieve better performance than conventional strategies without any knowledge of the bandwidth of its partners or any special mechanism for the load balance.

5 Conclusion

In this paper, we presented FBSA: a self-adjustable, distributed, and asynchronous scheduling algorithm for multi-source media streaming in P2P networks. Experimental results show that FBSA can afford shorter start delay time than existing algorithms, and is more adaptive to network heterogeneity, bandwidth fluctuation and overlay dynamics. For future work, we will study the layered scheduling, and try to exchange media data with a video frame as the smallest scheduling data unit. We will also do some further experiments in Planet-Lab [15] and practical operational networks.

Acknowledgment

This work was performed in Peking University, and was partially supported by the Key Technologies R&D Programme of China under grant No. 2006BAH02A10, the China Postdoctoral Science Foundation under grant No. 20060400001, and the“985” construction project of Peking University.

References

1. Zhang, X., Liu, J., Li, B., Yum, T.S.P.: Coolstreaming/DONet: A data-driven overlay network for peer-to-peer live media streaming. In: IEEE INFOCOM, Miami, FL, USA (2005)
2. Agarwal, V., Rejaie, R.: Adaptive multi-source streaming in heterogeneous peer-to-peer networks. In: SPIE/ACM MMCN, San Jose, CA, USA (2005)
3. Pai, V., Kumar, K., et al.: Chainsaw: Eliminating trees from overlay multicast. In: Castro, M., van Renesse, R. (eds.) IPTPS 2005. LNCS, vol. 3640, Springer, Heidelberg (2005)
4. Bawa, M., Deshpande, H., Garcia-Molina, H.: Transience of peers & streaming media. ACM SIGCOMM Computer Communication Review 33(1), 107–112 (2003)
5. Chu, Y.-H., Rao, S.G., Zhang, H.: A case for end system multicast. ACM SIGMETRICS Performance Evaluation Review 28(1), 1–12 (2000)
6. Hefeeda, M., Habib, A., Boyan, B., Xu, D., Bhargava, B.: PROMISE,: peer-to-peer media streaming using CollectCast. In: ACM MM’03, Berkeley, CA, USA (2003)
7. Silverston, T., Fourmaux, O.: Source vs data-driven approach for live p2p streaming. In: IEEE International Conference on Networking, Mauritius (2006)
8. PPLive: http://www.pplive.com/
9. PPStream: http://www.ppstream.com/
10. Gridmedia: http://www.gridmedia.com.cn/
11. Liang, J., Nahrstedt, K.: Randpeer. Membership management for QoS sensitive peer-to-peer applications. In: IEEE INFOCOM, Barcelona, Spain (2006)
12. Vishnumurthy, V., Francis, P.: On heterogeneous overlay construction and random node selection in unstructured p2p networks. In: IEEE INFOCOM, Barcelona, Spain (2006)
13. Huo, L.: Study on key techniques of media streaming over the internet, Ph.D. dissertation, Graduate University of Chinese Academy of Sciences (2006)
14. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge, MA (2001)
15. PlanetLab: http://www.planet-lab.org/

An Optimized Topology Maintenance Framework for P2P Media Streaming

Rui Guo[1,3], Longshe Huo[2], Qiang Fu[2], Shuangjia Chen[1,3], and Wen Gao[1,2]

[1] Institute of Computing Technology, Chinese Academy of Sciences, 100080 Beijing, China
[2] Institute of Digital Media, Peking University, 100871 Beijing, China
[3] Graduate School, Chinese Academy of Sciences, 100039 Beijing, China
{rguo, lshuo, qfu, sjchen}@jdl.ac.cn, wgao@pku.edu.cn
http://idm.pku.edu.cn

Abstract. In this paper, we present an optimized topology maintenance framework used in AVStreamer P2P media streaming system. To help new incoming clients get an initial view of the whole network, a new weak hierarchy model is designed to share burdens among different servers. Based on this new model, several key techniques are adopted: node evaluation criteria which can help optimize the node-list cache, receiver-driven gossip which is used to update node-list cache from others more effectively, random walk algorithm which can make the network more random, and status polling which is used to ensure the aliveness of each node cached. Experimental results show that the proposed techniques are effective.

Keywords: Peer-to-Peer, media streaming, topology maintenance, weak hierarchy model, receiver-driven gossip.

1 Introduction

Recently, P2P (Peer-to-Peer) technique has been widely used in the realm of media streaming. As a distributed system, it can support a very large amount of users in the condition that few equipments and network bandwidth are needed at the server's side. However, there still exist some challenging problems. On the one hand, since the experiences of P2P related technique are almost originated from the researches of P2P file sharing, and there are lots of differences between media streaming and file sharing, therefore the experiences should be modified to fit this new realm; on the other hand, because of distribution, dynamic peer failures may lead to certain damages to the living peers, which further result in poor quality of service at the client side.

To address these issues, several P2P structures have been put forward, which are more or less different from structures in file sharing systems. Tree-based structure, such as PeerCast [1] etc., is one of them. The idea is from IP multicast, except that the users in IP multicast are more likely to be steady and similar. Such a structure is not suitable for scenarios where there is a high probability of node failures.

Next to the tree-based structure, some improvements are proposed, among which multi-connected structure is mentioned, such as P2PCast [2] etc. Though these

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 334–342, 2007.
© Springer-Verlag Berlin Heidelberg 2007

structures still cannot solve all of the above problems, they make important progress in the field of P2P media streaming.

In [3], Zhang et al. proposed to use unstructured overlay network. From the view of topology maintenance, this paper gives some important points: (1) There is no need to maintain a certain structure; (2) multi-connection should be used to get data from other peers; (3) backup peers should be maintained to recover from failures of connected nodes. Such points have been accepted by most people, and have been implemented in lots of present live P2P media streaming systems [4]. But new problems appeared soon. Where to get the backup peers? How to decide which one should be preferred as a backup peer and which one should be preferred as a connected peer? What should the whole network look like? How to ensure the data transition is on the rails?

Based on previous works, recently we developed a system named AVStreamer [5], which is the first P2P media streaming system supporting the Chinese AVS standard (Audio Video coding Standard). This paper focuses mainly on the topology maintenance aspect of AVStreamer, and tries to answer the above questions encountered in the development of AVStreamer.

2 Framework of AVStreamer

As a distributed system, the framework design of a P2P streaming system from global view is very important. It is desirable to make new incoming nodes get robust service quickly and be scalable to provide service in a large extension. To meet these requirements, in our AVStreamer system four kinds of nodes are designed: Node-List Server (NLS), Normal-Client Node (NCN), Leased-Client Node (LCN), and Media Fetching Node (MFN). In this section, we first describe these nodes and their mechanisms respectively, and then propose a P2P topology maintenance model named weak hierarchy model based on them.

2.1 Node-List Server

The aim of designing Node-List Server (NLS) is to give the new incoming client node an initial list of nodes, so as to help them build their own view of the whole topology. To ensure that the initial node-list can serve various nodes and to avoid overload on some certain nodes, the capacity of the node-list cache on this server should be sufficient enough, and the contents of the cache should be updated periodically. To ensure the activeness of the cached nodes, we restrict them in two categories: (1) nodes which announce their activeness periodically to this server; (2) nodes which have just joined in the network. The criterion to determine which node should be stayed in cache lays in the capability of each node. When a node's capability has been exhausted, the Lease will be released. By using this criterion, it can be guaranteed that there are enough unsaturated nodes for new incoming nodes.

However, there still exists a serious problem: since new nodes are likely to be more unsaturated [6], the topology constructed using this method may become long and narrow, and the probability of division is higher than that of a totally random topology. To address this problem, we introduce two mechanisms to restrain the new incoming nodes. First, new nodes are refused to be used as Leased node according to a certain

probability, so as to let more old nodes be used in the node-list. Second, random-walk based technique is introduced, which will be explained in detail in Section 3.3.

2.2 Normal-Client Node

Each client-side user who enjoys P2P media streaming service is classified as Normal-Client Node (NCN). In each NCN, there also exists a node-list cache. Connected-peers (partners) are selected from the cached node list. In the next section, we will introduce several mechanisms, such as receiver-driven gossip, random walk and lease, to maintain the cached node list. Besides, an evaluation criterion should be used to select partners from the cached node list.

2.3 Leased-Client Node

Leased-Client Nodes (LCNs) are the nodes who send lease messages to NLS periodically to announce their aliveness. From the view of topology, these nodes are connected directly to the NLS, and are a part of the backup nodes which might be given to new incoming nodes as contents of their initial cached-node list. Each LCN sends a message to keep its lease on NLS periodically. If leased time has passed and no re-lease message has arrived, then this node will be removed from the NLS's node-list cache. NLS can confirm the aliveness of LCNs or gather topology information from them. Besides, NLS can redirect new incoming nodes to a LCN to lighten its own burden.

LCN is developed from NCN. There are two chances for a NCN to become a LCN: (1) when a node joins the overlay network; (2) when a node's node-list cache is empty and turns to NLS for help.

2.4 Media-Fetching Node

Media-Fetching Node (MFN) is used to get media data from the streaming server, and then distribute them in the P2P network. The data receiving status of MFN should be sent to its partners, so as the partners can request media data from it.

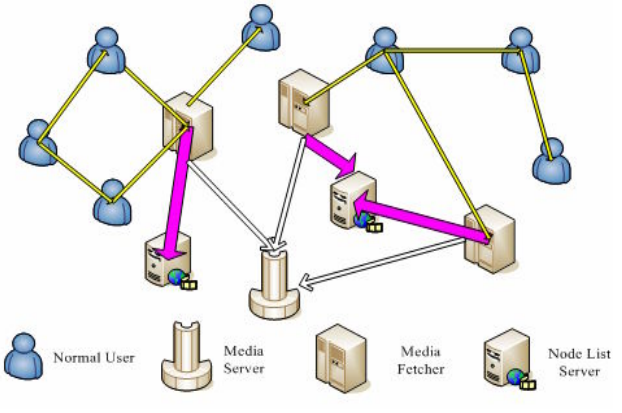


Fig. 1. How NLS, MFN, LCN and NCN are connected

In a practical system, MFN and NLS can be installed separately. Such a design has a salient advantage, i.e., it provides more selections for implementing the system. For example, we can deploy a MFN for each ISP, thus clients from every ISP can get better service, while only one NLS is installed in the whole network. As shown in Fig. 1, one NLS can provide services for different MFNs.

2.5 Weak Hierarchy Model

In AVStreamer, when a new client node joins the network, it should be given a node list to help it to get the initial view of the whole network. Therefore NLS is designed to deal with this problem. If we want to give an optimized node list, the NLS must track the records of all the nodes in the system, and extensive computations are needed to decide which ones should be the best. It may become the bottleneck both for performance and scalability, and is vulnerable to be attacked.

To address the above problem, we propose to use Leased Client Node (LCN). The advantages of using LCNs can be summarized as follows: (1) the cost of maintaining the node list to be sent and the computation of deciding which nodes should be sent can be shared with the NLS. (2) LCNs are selected from NCNs by capability, and are given greater responsibilities. This makes the whole network more efficient. (3) LCNs can also be used to gather the information about the whole overlay network. (4) For an existing NCN, what a LCN is responsible for is not critical. If a LCN becomes fail, no existing NCNs will be affected, thus there is enough time to recover a new LCN.

This model is similar as but also has notable difference from traditional Hierarchy Model [7], since LCNs are only used when new incoming users join the overlay network, and from NCN's view there are no differences between LCNs and themselves, therefore we call this model weak hierarchy model.

3 Proposed Key Techniques

For P2P media streaming, to make the whole system more robust and effective, partners and node-list cache play a very important role. Partners of a node are the neighbor nodes which directly connect to and exchange data with it. Node-list cache contains backup partners preparing for the fails or updates of current partners. Whenever partners and cached nodes are optimized, nodes can get better service from the overlay and the whole system can be more robust and steady. Aiming at this requirement, first we need a mechanism to get an overview of the status of the new nodes, and then use a criterion to select from them. In AVStreamer, we use receiver-driven gossip to exchange node-lists between nodes to help nodes get more information about new nodes. We use our node evaluation criteria to pick up partners and merge the exchanged node-list into nodes' own node-list cache and make the cache optimized. We also use random walk to make the overlay more random and steady, and status-polling to ensure the aliveness of the nodes cached. As a whole, these four techniques can be combined together and make the topology in AVStreamer system effective and robust.

3.1 Node Evaluation

In order to optimize the topology of a P2P media streaming system, some performance standards should be built first. While developing AVStreamer, we summarized four criteria to evaluate its topology maintenance system. First, the topology should be random enough, so that organization in which no peer is directly or indirectly connected with data provider has little chance to appear. Random topology also makes the network data transmission steady and robust in case of topology changes. Second, the importance of a peer should be affected strongly by its capability. If a peer has strong capability (such as abundant bandwidth, etc.), then it should take more responsibility. If a peer has poor capability, then it must not play an important role, otherwise it may become a bottleneck of the whole overlay network. Third, application-layer topology should be strongly affected by network-layer topology. If the application-layer topology is consistent with the network-layer topology, throughput between any two nodes can be maximized; as a result, the resource of the whole network can also be utilized more efficiently. Finally, the cost for topology maintenance should be as small as possible.

Some of the above criteria are conflict with each other. For example, if the topology is totally random, then the second and the third criterion may not be fit; if the application-layer topology is determined by peers' capacity and the network-layer topology, then the first criterion cannot be fit. Therefore a balance should be achieved among them.

To reach the balance, we need a standard to evaluate different nodes to which a node may connect. In AVStreamer, the following strategy is used: when comparing two nodes in cache, the popularities of them are considered first; if the difference is not big enough, then their capabilities are taken into account; if the difference of capability is rare, then the freshness of the nodes is considered.

There are two reasons for choosing the above strategy. On the one hand, they reflect a node's quality in some degree. In AVStreamer, popularity is decided by the times a node is recommended to others. If a node has been recommended more times, it is less tended to be recommended again, unless this node has more capability. Freshness is determined by a node's latest message time-stamp. Since a node's aliveness cannot be ensured all the time, thus the freshness is an important feature to guess whether the node is still alive. Capability is determined by the announced capability in periodically exchanged messages. Nodes exchange each others' capability at a certain short interval. Because recent node's capability is close to its current capability, it can be used to play the same role. On the other hand, they are easy to implement. If we want to confirm the exact aliveness, popularity, or current capability of a node, then messages will be transmitted more frequently and very high bandwidth will be cost. To avoid these costs, only freshness and capability at a certain time of a node should be used.

3.2 Receiver-Driven Gossip

In order to maintain and optimize a node' local cached node-list, it needs to get new nodes periodically and use the ones which are fit for it to replace the old unfit ones in the cache. Although each node has its own node-list cache, a node can get new node information from other nodes.

Gossip based dissemination protocol seems to be helpful to solve this problem and can provide good scalability and reliability properties [8][9][3]. However, traditional gossip protocol is not fit for our scene, since traditional gossip is sender-driven. In AVStreamer, a node needn't to transfer its node list to every one in the network, because not all of the receivers need the information. Besides, if all the nodes send its message to every one in the network periodically, there will be a very high traffic load over the network.

Based on this, we make an important improvement in AVStreamer: a receiver is not passively waiting for node-lists, but actively asking for them when it thinks it is necessary. In this way, nodes are no longer responsible for sending its own node-list to others actively, but only sending on demand. We call this mechanism receiver-driven gossip.

There are several advantages for receiver-driven gossip. (1) No broadcasting is involved. If gossip is driven by sender, every time the status of the sender is changed, a message should be send through the network. But for receiver-driven gossip, one message may contain several pieces of information which show all the changes during query interval. (2) It is more efficient. Once a message is only transmitted between a pair of peers, lower bandwidth is required. Also, the receiver asks for new peers only when it thinks its own node-list cache needs to be improved. It never receives messages which it is not interested in. (3) Node information can be transmitted from peers to peers, spread to certain scope, just like the traditional gossip protocols. For example, if *A* is a part of the node-list cache on *host_a*, when *host_b* queries *host_a* for node-lists, *A* might be sent to *host_b*. Next time when *host_c* query *host_b*'s node-lists, *A* also might be sent to *host_c*. (4) Newest nodes can always be used to query new nodes to make the diffusion of cached nodes fast enough.

By using receiver-driven gossip, a node can get to know other nodes which are currently far away from it. Its local node-list can be more and more random and fit for itself.

3.3 Random Walk

To make the topology more random, we introduced random walk [6] into the implementation of AVStreamer. In our design, we use it to help a peer to select its partners following some certain rules. When a node has cost more than half of its capability, it begins to use a probability which is inverse proportional to its current capability to decide whether to accept this new incoming node. If the new node is not accepted, it is redirected to another node, and this node makes the same decision repeatedly. Like this process, the joining message will also be redirected until one node decides to accept it or become out of time.

In our system, random walk plays an important role in several aspects. (1) The method can reduce the burden of LCNs and avoid them to be saturated too soon, because LCNs can redirect requests to other nodes. 2) If one partner is get by redirecting, this partner is tend to be far away from original node topologically. When original nodes exchange node-list with this partner, it can get more new nodes far away. The whole overlay network will be more random, too.

3.4 Status Polling

In AVStreamer, we adopted status polling to ensure the aliveness of nodes in local node-list cache. When a node is inserting into one's cached node-list, the inserting time is recorded. Whenever a data transmission occurs, the inserting time will be updated to the current time. If the interval between current time and the inserting time is longer than a certain value, a status-query message will be sent to make sure the node is still alive. Some other information about the node's current status is wanted, too.

By using status polling, the aliveness of cached nodes is ensured in a certain extent. If a node is selected as partner, the probability of success is increased.

4 Experimental Results

To evaluate the performance of the proposed techniques, we test AVStreamer in our laboratory. In our experiment, 50 PCs are involved, each PC runs 2~8 instances of AVStreamer during the test. For each instance, limitation of the size of node-list cache is set to 8, limitation of the size of partners is set to 5, limitation of the size of active connected partners is set to 2. We organized 4 groups of experiments. Whole sizes of the overlay network are different between each experimental group.

Table 1. Average hops to media-fetching node in different size of network

Total size	100	200	300	400
Average Hop	5.8	6.7	7.3	7.6

4.1 Average Hops to Media-Fetching Nodes

If a overlay network is totally random, its diameter should be proportioned to log(N), where N is the size of the network. A random network can be more robust to topology changes. Besides, the average data transmission delay can be reduced. In our experiments, we measure average hop to MFNs instead, because the average diameter is not easy to get. MFNs are the source of data transmission. The distance between MFNs and NCNs show the semi-diameter of the network to some extents.

As shown in Table 1, we can see that the average hop to media fetching nodes is close to being logarithmic to the size of the network. It should be because of random walk mechanism, and our node evaluation criteria.

4.2 Node Information Diffusing

Table 2 shows the average amount of different nodes that a client can get by using receiving-driven gossip. Member exchanging is done at interval of 40s, each exchange queries 1/3 of the size of node-list cache nodes from other peers.

Table 2. Average amount of different nodes a client has got after 20 minutes

Total size	100	200	300	400
Average nodes get	72	77	81	83

From this table, we can see that, by using receiver-driven gossip, each node has the chance to know lots of other nodes in the network. That is helpful to recovering from partner failures.

4.3 Node Degree Distribution

In the whole network, there should be enough unsaturated nodes for the new incoming client, but for each individual, a node should have enough partners to make the data transmitting steadier, therefore we give the node degree distribution of each experiment.

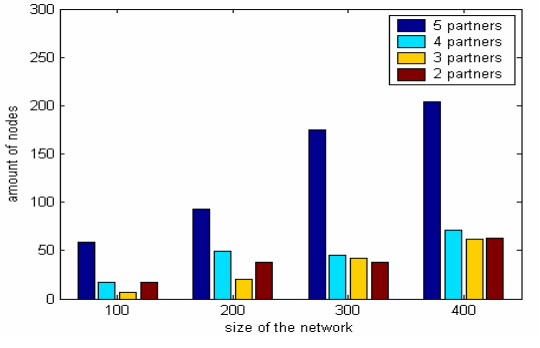


Fig. 2. Node degree distribution

As shown in Fig. 2, we can see almost half of the nodes are unsaturated, which means when new node joins the network, there are nearly half nodes available to be connected to them. The reason results from the difference between the limitation of the active connected partners and the limitation of the total partners.

4.4 Bandwidth Cost

We test the bandwidth costs of four experiments. The results shown in Table 3 tell us the bandwidth cost is rare, and it increases very slowly when the size of network become larger. This is because all the messages are transmitted locally.

Table 3. Average bandwidth cost in 20 minutes

Network Size	100	200	300	400
Average download bitrate(kbps)	1.1	1.2	1.3	1.3

5 Conclusions

In this paper, we presented the design of an optimized topology maintenance system in AVStreamer, and discussed the new weak hierarchy model. This model is more effective, it shares burdens from Server to help new incoming client to get an initial view of the network. We also discuss some key techniques used in AVStreamer, such as how to evaluate nodes to help optimizing the node-list cache, using receiver-driven gossip to help fetching new nodes from others more effectively, using random walk to redirect nodes' burden and make the network more random, and using status polling to ensure the aliveness of each node. Experimental results show that the proposed techniques are effective.

Acknowledgments. This work was performed in Peking University, and was partially supported by the Key Technologies R&D Programme of China under grant No. 2006BAH02A10, the China Postdoctoral Science Foundation under grant No. 20060400001, and the"985" construction project of Peking University.

References

[1] Deshpande, H., Bawa, M., Garcia-Molina, H.: Streaming Live Media over a Peer-to-Peer Network. Stanford database group technical report (2001-20) (August 2001)
[2] Nicolosi, A., Annapureddy, S.: P2PCAST: A Peer-to-Peer Multicast Scheme for Streaming Data. In: Proceedings of First IRIS Student Workshop August 2003. Cambridge, Massachusetts (2003)
[3] Zhang, X., Liu, J., Li, B., Yum, T.S.P.: Coolstreaming/DONet: A data-driven overlay network for peer-to-peer live media streaming. In: IEEE INFOCOM, Miami, FL, USA (2005)
[4] Gao, W., Huo, L., Fu, Q.: Recent Advances in Peer-to-Peer Media Streaming Systems. China Comminications.5, 52–57 (2006)
[5] Huo, L.: Study on key techniques of media streaming over the internet, Ph.D. dissertation, Graduate University of Chinese Academy of Sciences (2006)
[6] Gkantsidis, C., Mihail, M., Saberi, A.: Random Walks in Peer-to-Peer Networks. In: IEEE INFOCOM. HongKong, China (2004)
[7] Yang, B., Garcia-Molina, H.: Designing a Super-Peer Network. In: Int'l Conf. on Data Engineering, Washington, US (2003)
[8] Ganesh, A.J., Kermarrec, A.M., Massoulie, L.: Scamp: Peer-to-peer lightweight membership service for large-scale group communication. In: Workshop on Networked Group Communications, London, UK (2001)
[9] Jelasity, M., Babaoglu, O.: T-Man: Gossip-based overlay topology management. Engineering Self-Organising Applications. Utrecht, The Netherlands (2005)

Building Large Scale 3D Face Database for Face Analysis*

Yuxiao Hu, Zhenqiu Zhang, Xun Xu, Yun Fu, and Thomas S. Huang

Beckman Institute, University of Illinois at Urbana-Champaign
405 North Mathews Avenue, Urbana, IL 61801 USA
{hu3, zzhang6, xunxu, yunfu2, huang}@ifp.uiuc.edu

Abstract. We propose to build a large scale 3D face database with dense correspondence for variant face analysis research purposes. Large scale means that the number of subjects in the database is more than 400, which is, to our best knowledge, the biggest[1] one at this time. 3D face means that we provide both the texture and shape of human faces, which is also balanced in gender and race. Dense correspondence means that the key facials points with semantic meanings are carefully labeled and aligned among different faces, which can be used for a broad range of face analysis tasks. We provide the data description, data collection schema and the post-processing methods to help the usage of the data and future extension. More and more data is still being collected and processed to enlarge the extensive 3D face database. The proposed face database provides solid ground truth for human face related tasks such as alignment, tracking, recognition and animation, etc.

1 Introduction

Human faces are the most popular and important objects in the real world. Computer graphics researchers are interested in face modeling, synthesis, and animation, while computer vision researchers are investigating the methods for face analysis, including face detection, alignment, tracking and recognition. Though research on frontal detection and recognition has achieved a lot, non-frontal face analysis, especially face alignment and recognition still remains open problems. In order to address these issues, more and more attentions are attracted to 3D face analysis. As the foundation to all these tasks, building a comprehensive 3D face databases is necessary.

There are many existing databases containing 3D face information. According to their construction methods, these databases can be categorized to three types: built from multi-view geometry, built from structure lights and captured by 3D scanners. The CMU Face In Action (FIA) Database [1] consists of 20-second videos of face data from 180 participants mimicking a passport checking scenario. The data is captured by six synchronized cameras from three different angles. Since it only includes

* This work is supported by NSF Grant CCF 04-26627.

[1] Since the data format is also compatible to some other database of the same type. The total number of the available subject is almost 700 now.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 343–350, 2007.
© Springer-Verlag Berlin Heidelberg 2007

unmarked face video sequences, sophisticated face tracking algorithms need to be applied to recover the head poses and face shapes. Some other 3D face databases are collected by structure lights. For example, in 3D_RMA [2], it contains 120 individuals. The 3D face shape was constructed using a camera and a projector. The 3D coordinates of the surface points were represented with high precision. The limitation of this method is that glasses and dark parts of faces could not be captured, resulting the shapes of the face are not complete. 3D scanners are widely used recently, which are able to provide precise shape and texture information of an object. In GavabDB [3], 61 individuals are scanned from different views with variant expressions. Since the Minolta VI-700 digitizer in [3] only captured one view of the face, there are some invisible parts in a single view due to self occlusion. Hence extra post-processing must be carried out to concatenate the shape and texture map from different views to get a complete face surface. Compared with the aforementioned scanners, a Cyberware Scanner [4] **Error! Reference source not found.**is able to scan the complete view of human heads in one scan. It is convenient both for the data collection and post processing. Vetters et al. provided a 3D morphable face database based on USF 3D face database, which included 200 subjects [5][6]. However, comparing to the high dimension of the 3D shape and texture data, the number of the samples is still far from enough for future face analysis tasks.

In this paper, we propose to build a larger 3D face database with Cyberware 3D scanner which includes more than 400 individuals. Different to the registration method used by Vetters, the raw 3D face shapes are aligned manually instead of by optical flow, which provide more accurate results. Currently, there are already 475 subjects in our 3D face database with both neutral and smile expressions available for most of them. The ages of the subjects range from 19 to 55 and balanced in ethnic group. The database can be used for a broad range of face analysis tasks including alignment, tracking, recognition and animation, etc. The data description, data collection schema and the post-processing methods are also provided to help the usage and future extension of the database.

2 The Data Collection Schema

In this section, we first brief our data collection schema. About 500 subjects were invited for the 3D face scanning during February to May in 2006, most of them are students. The subjects were carefully selected to balance gender. About half of the subjects were western people including Caucasia and African American ethnic groups and the other half was Asian. The ages of the subjects were also recorded for future reference, which ranged from 19 to 55. The subjects were asked to show neutral expression during the scanning and for most of them, another scan for smile expression was collected for possible future recognition purposes.

A 3D human face was first scanned by Cyberware scanner, which output the raw data including the head shape and corresponded texture map. Then 65 pre-defined key facial points, such as eye/mouth corners and nose tip, were labeled on the texture map manually. According to these key points and the reference mean face model, the face region was cropped out and the dense triangular mesh shape model was created by

interpolation so that the vertices in different face model corresponded with semantic meaning. Finally, the poses of the 3D faces were normalized so that their 3D coordinates were in the same reference frame.

The steps of the 3D face capture and post-processing are detailed in section 3.

3 Three Dimension Face Capture

3.1 Cyberware Scanner

To capture the 3D human face data, we used the Cyberware Head & Face Color Scanner bundle, including a Cyberware 3030 scanhead mounted on a Cyberware PS cylindrical motion platform. The 3030 shines an off-axis low-intensity laser illuminating a profile captured from two video sensors to triangulate its spatial location. Thousands of these profiles are combined to construct a single range scan dataset. The scanner additionally captures a 24-bit RGB texture map using a cylindrical surface parameterization.

3.2 Data Capture Environment

The CyberWare3030 scanner provides an ambient illumination system. We shielded all the other lights from the window and in the room to get consistent and uniform ambient rays so that the illumination condition is reasonable and consistent at each time we scan a new subject.

3.3 Accessories

The result of the scanner is not good when it scans complex structure objects such as hair and black objects. Since only face region is concerned, every subject was asked to wear a wig cap and take off glasses. The cap is white, blue or black which is different from the skin color of the subject. The cap shields the hair above the forehead and behind the ears, and keeps the face region at the most.

Some examples about the raw 3D data are shown in Figure 1.

4 Data Labeling and Post-Processing

The raw 3D data from the Cyberware scanner is in resolution of 734*457 for both the shape and texture map. The raw shape contains 734*457 vertices and each vertices is represented by x, y and z coordinates. The raw texture is represented with 734*457 .tiff image. Because the absolute position and orientation of the head are different for each subject, the raw data need to be labeled and post-processed to extract the useful human face information. In this section, we detailed the steps of labeling the key points, cropping the face region, building sparse/dense correspondence and aligning the face model.

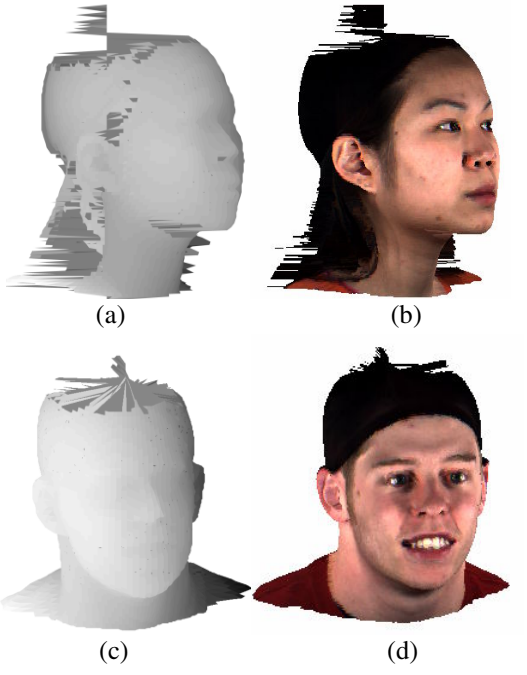


Fig. 1. Samples of the raw 3D data from Cyberware scanner. (a)(b) Shape and Texture of Sample#304, Female, in neutral expression (c)(d) Shape and Texture of Sample#310, Male, in smile expression.

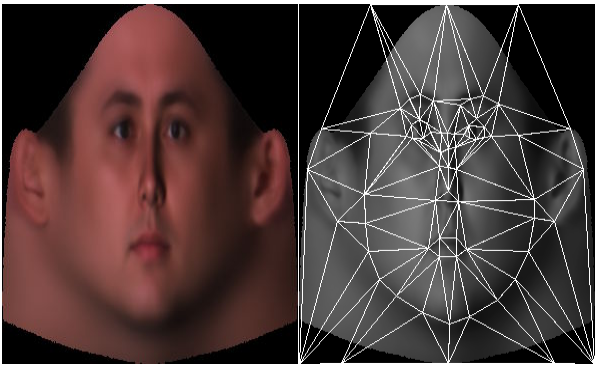


Fig. 2. The Reference Face and the Key Points

4.1 Data Labeling and Cropping

Typically the face region occupies an area about 500*400 within the whole scanning cylinder. In order to specify the face regions and build the correspondence between

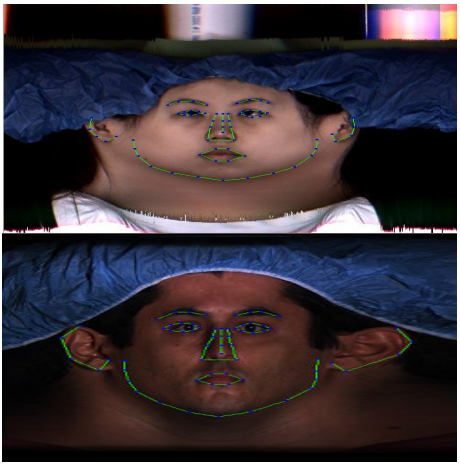


Fig. 3. The texture map with the key points labeled with blue dots. The key points are connected with green lines in different components.

different faces, a reference 3D face model is used. 75 key facial points on the reference face model is defined as anchor points, including eye corners, mouth corners, nose tip and face contour, etc. The sample data are also labeled on these anchor points manually. The reference face and the pre-defined key points are shown in Figure2. The triangular configuration is also defined for future warping as stated in Section 4.2. Again, since only face regions are concerned, the vertices and texture corresponding to the region outside the reference face are cropped. Some labeled face texture map are shown in Figure 3. We notice that some of the key points are not labeled compared to the reference face. These points will be calculated in the post-processing steps according to other labeled points and will be used for face warping in the next step.

4.2 Building the Correspondence

Because different subjects' heads are in different position and orientation when they are scanned, in the raw 3D data, the vertices and texture information of different subjects are unrelated. For example, the *i-th* vertex in one 3D face is on the nose tip, but the *i-th* vertex in another 3D face is on the cheek, i.e., these two vertices have different semantic meanings. Since vertices correspondence are preferred in typical face research tasks, especially for facial animation and subspace analysis, re-sampling the vertices and texture map is required to make the vertices "*corresponded*", i.e., the *i-th* vertex in different 3D faces are all nose tips and the *j-th* vertex in different 3D faces are all left corner of the mouth.

Since the 75 key facial points with semantic meanings have been located manually in the texture map, a sparse correspondence is already established. Other points within the face region can be interpolated and warped to the reference face. The points which can be warped to the same points on the reference face are corresponded. Because

Fig. 4. Verification the result of texture correspondence: (a) ref. face (b) input sample face (c) warped sample face

Fig. 5. Verification of result of shape correspondence: The first row is the original raw face shape, the second row is the reconstructed face shape with dense correspondence.

each point in the texture map corresponds to a vertex in the shape model, the correspondence between two vertices from different 3D faces are also established under the help of reference face.

There are two aspects to be verified for the result of our data labeling and the dense correspondence. On texture maps, we warp a sample face to the reference face according to the established correspondence. If the facial points are aligned perfect according to their semantic meanings, the result warped face should looked in the same shape to the reference face, but the texture is the same to the original sample face. On the contrary, if the facial points are not aligned well, there will be some distortions. A sample of the face warping is shown in Figure 4. For shape vertex, the vertices actually are re-arranged to represent the 3D face shape. If the shape correspondence is perfect, the 3D shape constructed with the re-arranged 3D vertices should be the same shape to the original raw 3D face shape. On the contrary, the reconstructed 3D shape will be different to the original data. As shown in Figure 5, the 3D face shape remains the same before and after the processes.

Fig. 6. Key facial points plots before/after the Procrustes Analysis. Left: before, Right: after.

Fig. 7. Samples of the 3D face after post-processing

From above verification, we may claim that building the dense correspondence did not affect the original 3D face shape and texture.

4.3 Vertex Coordination Normalization

Although the vertices are already in the same semantic meaning for different 3D faces after the dense correspondence has been established, their coordinate systems are still not consistent, i.e., the faces are not in the same poses if we put them in a 3D space. In order to normalize the 3D pose (location, rotation and scale) of the collected shape data so that they achieve the best alignment with each other, 3D Procrustes analysis [7] is applied. After the 3D Procrustes analysis, all the 3D face shapes vertices are transformed to the same coordinate system, in which all the corresponded face vertices are well aligned. The 3D face shape vertices before and after the 3D Procrustes analysis are compared in Figure 6. In the left plot, we plot the key facial points of several faces together, the 3D points scattered around because are not well alignment. The same set of points is plotted again in the right plot after Procrustes analysis, which is more consistent. (Because of the individual variance, they will not completely overlap.)

Samples of the 3D face data after post-processing are shown in Figure 7.

5 Data Organization

Both the original raw data and the normalized 3D face data with dense correspondence are available with the information of the subjects, including age, gender and ethic group. The shape and texture map of the raw data are stored separately, which occupied about 3GB disk space. Each file pair is named by the subject number with suffix "-N" or "-S" for different expression. After post-processing, there are 33420 vertices/points for each face, which is about 400MB in file size.

6 Conclusion

A large scale 3D face database with dense correspondence was described in this paper. The database contains more than 400 subject which balances in gender and race. Dense correspondence and alignment among different faces are also established, which makes the proposed database is useful for a broad range of face analysis tasks. The data description, collection schema and the post-processing methods are detailed to help its usage and for future extension. Since the data format is compatible to some existing databases [8], more subjects can be added and processed to further enlarge it. In Conclusion, the proposed face database provides solid ground truth for human face related tasks such as alignment, tracking, recognition and animation, etc.

References

1. Goh, R., Liu, L., Liu, X., Chen, T.: The CMU Face. In Action (FIA) Database, In: the Proceedings of the 2005 IEEE International Workshop on Analysis and Modeling of Face and Gestures, in conjunction with the 2005 International Conference of Computer Vision (ICCV '05), Beijing, China, October (2005)
2. http://www.sic.rma.ac.be/~beumier/DB/3d_rma.html#1
3. Moreno, A.B., Sanchez, A.: GavabDB: A 3D Face Database. In: Proc. 2nd COST275 Workshop on Biometrics on the Internet, Vigo (Spain) (2004)
4. http://www.cyberware.com/products/scanners/ps.html
5. Blanz, V., Vetter, T.: A Morphable Model for the Synthesis of 3D Faces, In: SIGGRAPH'99 Conference Proceedings, pp. 187-194 (1999)
6. USF DARPA HumanID 3D Face Database, Courtesy of Prof. Sudeep Sarkar, University of South Florida
7. http://www.dcs.gla.ac.uk/ mc/1stYearReport/2.5_GPA.htm
8. The BJUT-3D Large-Scale Chinese Face Database, Technical Report of The Multimedia and Intelligent Software Technology Beijing Municipal Key Laboratory in Beijing University of Technology (2005)

Senary Huffman Compression – A Reversible Data Hiding Scheme for Binary Images

Chung-Chuan Wang[1], Chin-Chen Chang[2], Xinpeng Zhang[3], and Jinn-Ke Jan[1]

[1] Department of Computer Science, National Chung Hsing University, Taichung 402, Taiwan, R.O.C.
ccwang@dragon.ccut.edu.tw, jajan@cs.nchu.edu.tw
[2] Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan, 40724, R.O.C.
ccc@cs.ccu.edu.tw
[3] School of Communication & Information Engineering, Shanghai University, Shanghai 200072, P.R.C.
xzhang@staff.shu.edu.cn

Abstract. Over the past few years many studies have proposed reversible data hiding schemes, but few have been applied to binary images. Some studies have utilized spread spectrum, compression and binary operations methods to achieve the data hiding goal, but most of them suffered from poor visual quality, capacity and inability to extract the hidden data during recovery. The performance of existing methods also is unsatisfactory. Therefore, this paper proposes a reversible data hiding scheme for binary images: **SHC** (***senary Huffman compression***). SHC adopts the half-white and half-black pixels of 4×1 or 2×2 blocks of six types to increase visual quality. Moreover, SHC is senary instead of binary and becomes double senary as a compression unit to increase compression rate and secret hiding capacity. Experimental results show that recovered images are well within human visual perception and have PSNR(s) greater than 33, high secret hiding capacity for 1.1 secret bits / 2 bits, and an effective compression rate of over 17% on average. All results demonstrate that the scheme has advantages in reversible data hiding for binary images.

Keywords: reversible data hiding, PSNR, visual perception.

1 Introduction

With the advancement in uses of the Internet, information security has become more and more important [6, 11]. Two of the most popular information security technologies are cryptography and steganography. While cryptography emphasizes meaning, steganography emphasizes existence [12]. However, in public networks, contents with cryptography are meaningless codes that readily attract a variety of attacks [3, 5, 14]. On the other hand, steganography is used to embed secret data into a cover image, which should have visual imperceptibility, security, high capacity and low distortion [1, 2, 9].

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 351 – 360, 2007.
© Springer-Verlag Berlin Heidelberg 2007

The best known steganography scheme involves embedding secret data in the least significant bit (LSB) of the stego image [10, 17]. The distortion from changing pixels on the last one to three bits is slight and the method is simple [9]. However, the LSB method cannot be applied directly to binary images; therefore, few research papers have discussed data hiding in binary images [6, 17]. A binary image contains only two colors, black and white, making any modification easily noticeable; thus, it is more difficult to distort the image for embedding data than with color or gray-scale images [15, 16]. To achieve reversible data hiding in binary images, the secret data is extracted from the stego image, and then the original image is recovered. Binary image processing is more convenient and more popular than color or gray images and developing such techniques is important for many applications. For instance, medical diagnosis and law enforcement use these techniques to authenticate critical legal documents [12].

Tsai etc. proposed the PWLC (pair-wise logical computation) [15] reversible data hiding scheme for binary images. However, their scheme fails to recover the original image without extra information because it cannot clearly distinguish valid blocks between with the embedded data and the original unaltered blocks [13].

This paper proposes a novel reversible data hiding scheme for binary images based on the senary Huffman compression (SHC) mechanism, which can achieve the goals of reversibility, losslessness and with a specific senary transformation to protect it from being divulged simultaneously. Moreover, the Huffman method is a widely applied and very effective algorithm for compressing data [4].

Any arbitrary alteration in the uniform blocks of a binary image is significantly perceptible [8, 10]. For imperceptible and reversible data hiding in a binary image with SHC, each embedded block must own six variances types with half-white and half-black pixels. Therefore, 4×1 and 2×2 2-white and 2-black pixel blocks are satisfied and adopted. For convenience, in this paper we utilize 4×1, 2×2 blocks to represent 4×1 2-white and 2-black, 2×2 2-white and 2-black pixel blocks, respectively. Both types will be compared in our experiments.

In essence, the SHC scheme retrieves all non-interleaving 4×1 or 2×2 blocks from the host image. Utilizing these blocks could compress not only the blocks themselves but could also encode and embed secret data for those saving spaces.

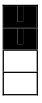
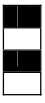
The rest of this paper is organized as follows: Section 2 gives some simple notations and definitions for the scheme. Section 3 presents an SHC reversible data hiding scheme for binary images, and Section 4 shows the experimental results. Conclusions are provided in Section 5.

2 Notations and Definitions

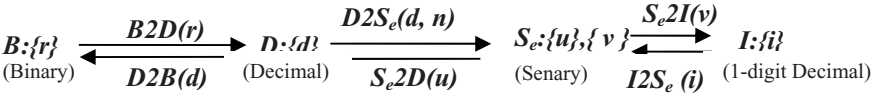
Initially, we employ senary to represent the six various types of 2×2 and 4×1 blocks listed in Table 1.

In this section, we present the notations for variables and the definitions of functions that are applied to the SHC scheme. In Fig. 1, functions (***B2D(r)***, ***D2B(d)***,

Table 1. Six types of 4×1 and 2×2 blocks with 2-white and 2-black pixels

Unit	Content					
4×1 block						
2×2 block						
r	001	010	011	100	101	110
Senary e	0	1	2	3	4	5

D2Se(d,n) and ***Se2D(u)***) among ***B***(Binary), ***D***(Decimal) and ***Se***(Senary) are general base transformations. The subsequent functions (***S_e2I(v)*** and ***I2S_e(i)***) between ***I***(1-digit Decimal) and ***Se***(Senary) are specific base transformations for employing 1- or 2-digit senary to represent a 1-digit decimal.



B: set of 3-bit *binary* numbers, $r \in B$ ***D***: set of *decimal* numbers, $d \in D$

I: set of 1-digit *decimal* numbers, $i \in I$ S_e : set of *senary* numbers, $u, v \in S_e$

Fig. 1. An outline of functions and notations between base transformations

Definition 1. Let i be a 1-digit decimal. Function $I2S_e(i)$ converts i to v, where $v = \begin{cases} i & \text{if } 1 \le i \le 5, \\ 0 \| (i \bmod 5) & \text{otherwise,} \end{cases}$ where "‖" denotes the notation of concatenation.

For example, if i=7, then $I2S_e(7)$=0‖(7 mod 5)=02.

Definition 2. Let v=e or $0e$, where e is a senary digit. The function $S_e2I(v)$ is to convert v to a decimal digit i, where $i = \begin{cases} v & \text{if } v = e, \\ \left\lceil \frac{e}{5} \right\rceil \times 5 + e & \text{if } v = 0e. \end{cases}$

For example, if v=02, then $S_e2I(02) = \left\lceil \frac{2}{5} \right\rceil \times 5 + 2 = 7$.

Definition 3. Let b be a base unit. The $(b)^n$ denotes $\underbrace{bb...b}_{n}$.

3 The Proposed Scheme

This section introduces an SHC reversible data hiding scheme for binary images. The scheme has two phases: embedding (listed in Fig. 2(a)) and extracting (listed in Fig. 2(b)).

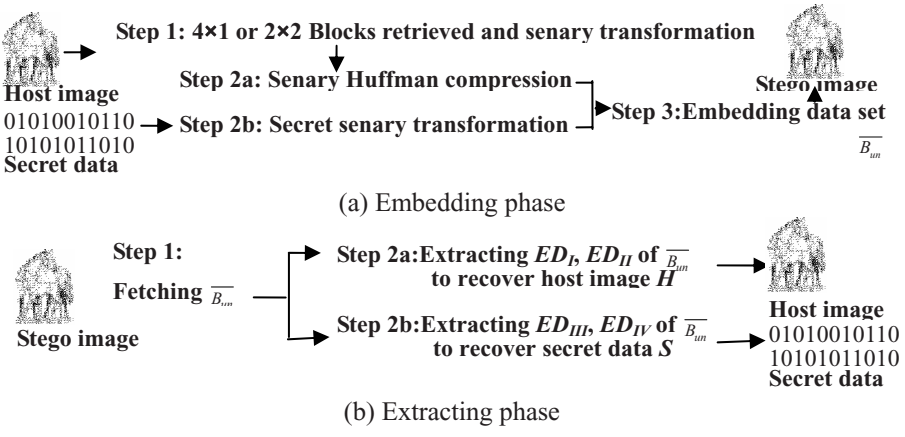


Fig. 2. The embedding and extracting phases of SHC scheme

3.1 Embedding Phase

In this section, a detailed description and clear examples illustrates how secret data is embedded into host image, broken into the two steps:

Step 1: 4×1 or 2×2 blocks retrieved and senary transformation
The first 3-bit binary (r) of 4×1 or 2×2 block could simply represent the block as a senary e by $B2D(r)$-1, as listed in Table 1. Table 2 shows the SHC pair-wise senary $u= e^2$. A clear example is given in Fig. 3.

Table 2. Types of specific senary $u(=e^2)$

Unit	Value										
$u(= e^2)$	00	01	02	03	04	05	10	11	. . .	54	55
$i(=S_e2D(u))$	0	1	2	3	4	5	6	7	. . .	34	35

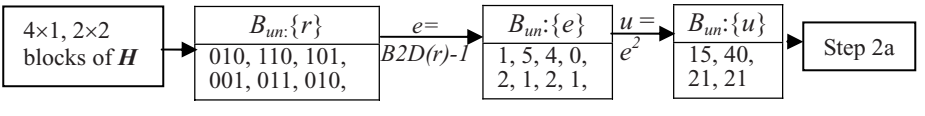


Fig. 3. An example of base transformation between

Step 2a: Senary Huffman compression

The SHC scheme includes four processes: (a) Statistics: All non-zero frequencies of u(s) transformed from an image are to be counted and sorted; (b) Building a senary tree: The least six frequencies of u(s) are retrieved and summed to a new frequency. Repeating the step until all the frequencies are retrieved; (c) Creating a Huffman table: Each leaf node depth and path of u in the senary tree is recorded in sub-tables T_d and T_p , respectively; (d) Compression: Initially, the Huffman table(T_d, T_p) is embedded. Then, according to the Huffman table, all the u(s) transformed from the image are replaced with the corresponding path of sub-table T_p. In Table 4 and Fig. 4, an example illustrates the SHC scheme. For simplicity, assume a small range of u(s) with frequencies from 00 to 15 as listed in Table 4 and the method shown in Fig. 4.

Table 3. An example of senary Huffman compression ('null' - no any value)

Unit		Content											
u		00	01	02	03	04	05	10	11	12	13	14	15
$i(=S_e2D(u))$		0	1	2	3	4	5	6	7	8	9	10	11
Frequencies		39	25	1	9	8	13	0	7	4	11	3	17
Huff_table	T_d	1	1	2	2	2	1	0	2	2	1	2	1
	T_p	5	3	40	45	44	1	(null)	43	42	0	41	2

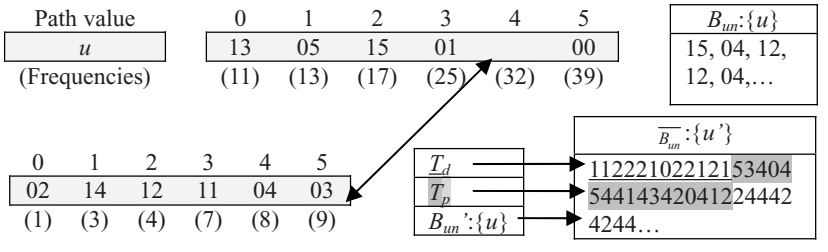


Fig. 4. The method of senary Huffman compression

Step 2b: Secret senary transformation

To increase capacity and security, the secret data are compressed into a specific senary in each 4×1 or 2×2 block. Table 4 shows that a total 24 secret bits are embedded into the corresponding 11 blocks and each block is sized 4 bits;on average, 1.1 secret bits are embedded on 2 bits. In practice, 10 blocks are recorded secret length. Consequently, every 3-bit binary s of secret data is converted into a specific embedding senary unit v through function $I2S_e(B2D(s))$, where $v \in$ {00, 1, 2, 3, 4, 5, 01, 02}.

Table 4. Secret lookup table

Unit	Value							
3-bit string s (decimal)	000 (0)	001 (1)	010 (2)	011 (3)	100 (4)	101 (5)	110 (6)	111 (7)
$v= I2S_e(B2D(s))$	00	1	2	3	4	5	01	02

If the length of the secret data is not tripled, 0 or 00 is appended to satisfy the triple. The secret lookup table is shown in Table 5.

An example of secret senary transformation is given in Fig. 5.

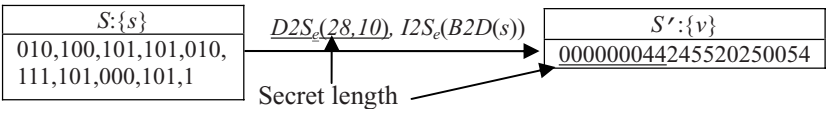


Fig. 5. An example of secret senary transformation and embedding

Step 3: $\overline{B_{un}}$: Embedding data set

Embedded data set $\overline{B_{un}}$ contains image recoverable information ED_I, ED_{II} and secret recoverable information ED_{III}, ED_{IV} , as shown in Fig. 6.

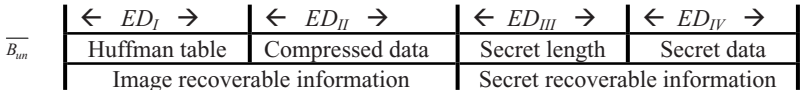


Fig. 6. The content of $\overline{B_{un}}$

Fig. 7 shows an example of the embedding process.

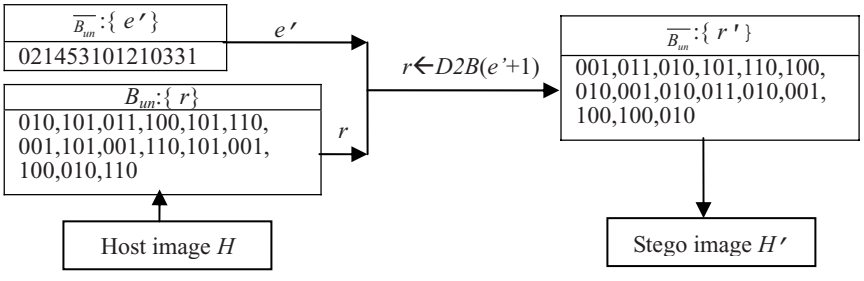


Fig. 7. An example of the embedding process

3.2 Extracting Phase

The extracting phase archives the stego image reversible and the secret data extractable. Therefore, the two steps of the extracting phase are:

Step 1: Fetching $\overline{B_{un}}$

All 4×1 or 2×2 blocks are retrieved as r's, then converted to e's, $\overline{B_{un}}$ is fetched out.

Step 2a: Extracting ED_I, ED_{II} of $\overline{B_{un}}$ to recover host image H

The two sub-tables T_d and T_p of the Huffman table are retrieved in ED_I. Subsequently, by retrieving the blocks of compressed data in ED_{II} then replacing each $T_p[i]$ with u

from the Huffman table until all u(s) are retrieved, the $ED_{II}(B_{un}:\{u\})$ is generated. An example of the extracting phase is given in Fig. 8.

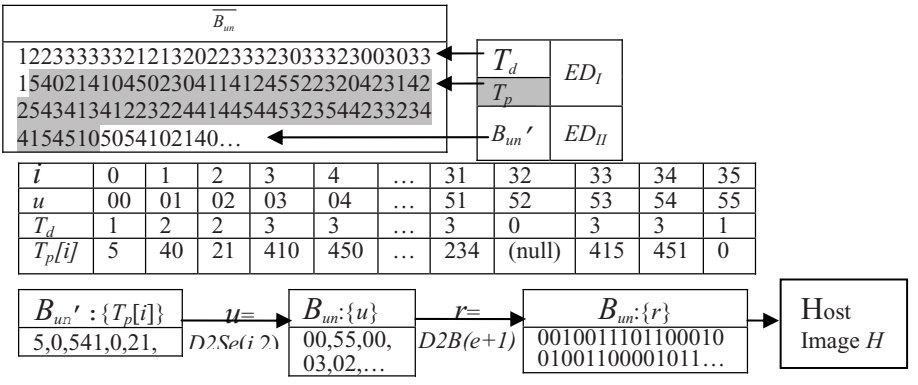


Fig. 8. An example of the embedding process

Step 2b: Extracting ED_{III}, ED_{IV} of $\overline{B_{un}}$ to recover secret data S

The secret length is transferred from ED_{III} :e^{10} to a decimal by $Se2D(e^{10})$. Consequently, the specific senary secret data ED_{IV} is transferred from each senary v to 3-bit r until the secret length is reached.

An example of extracting secret data from ED_{III}, ED_{IV} of $\overline{B_{un}}$ is given in Fig. 9.

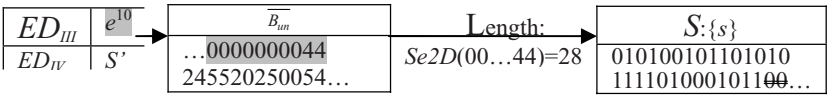


Fig. 9. An example of extracting ED_{III}, ED_{IV} of $\overline{B_{un}}$ and recovering the secret data

4 Experimental Results

Our experiments employed three binary images: (1) baboon - 512×512 pixels as shown in Fig. 10a, (2) scene - 700×470 pixels as shown in Fig. 10b and (3) horse - 570×619 pixels as shown in Fig. 10.

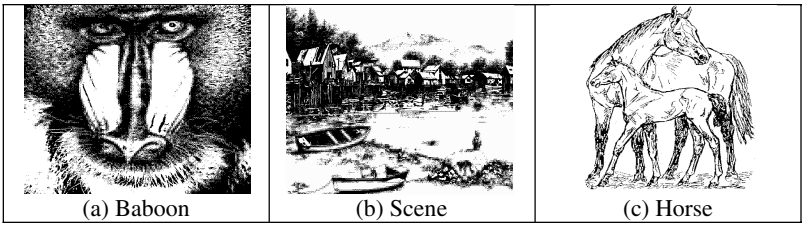


Fig. 10. Three test images

All the test images embedded their maximum capability of random secret data with 4×1 and 2×2 blocks are shown in Fig. 11. Peak signal to noise ratio (PSNR) is a general measure of visual quality. The PSNR values of 4×1 and 2×2 blocks in Table 6 show: 36.09, 36.44(baboon), 36.28, 35.66(sence) and 40.35, 38.71 (horse) are satisfied.

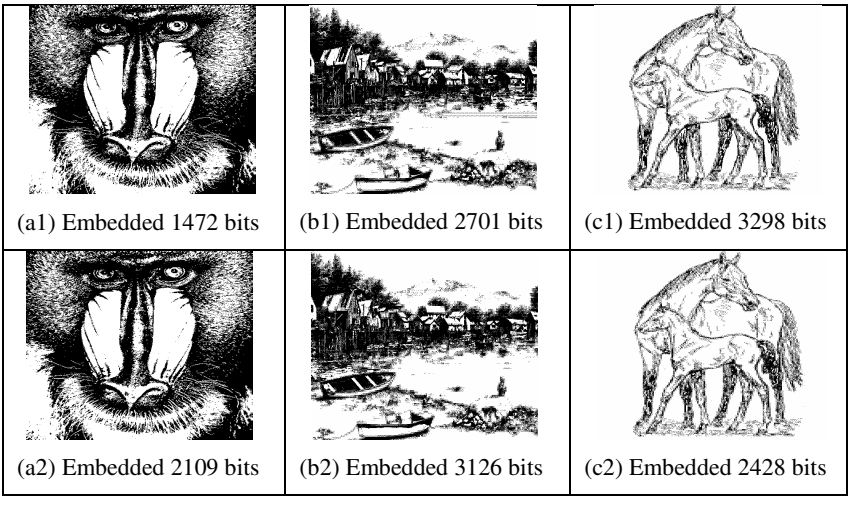


Fig. 11. Stego images – embedded the maximum capacity of secret bits with 4×1(a1, b1, c1) and 2×2(a2, b2,c2) blocks

The secret capacity is dependent on the number of valid blocks and compression rate. The data compression rate is 1 - (the number of compressed valid blocks / the number of valid blocks) (= 1 - ED_{II} / (ED_{I}+ED_{II}+ED_{III}+ED_{IV})). The results in Table 5 show 4×1 and 2×2 type of blocks for (a) secret capacity (1941<2572, 2701<3126 and 1084<2109), (b) the number of valid blocks (7926>7048, 7011<7317 and 4174<6008) and (c) compression rate (12.91%<18.69%, 19.54%<21.46% and 15.69%<18.48%).

Table 5. Test results for experiments

Image (pixels)	Block	PSNR	Valid/Total (blocks)	Secret capacity	Compression rate(%)	Entropy
Baboon (507×507)	4×1	33.82	7926 / 64262	1941	12.91	1.74
	2×2	36.69	7048 / 64262	2572	18.69	1.63
Scene (700×470)	4×1	37.73	7011 / 82250	2701	19.54	1.61
	2×2	37.57	7317 / 82250	3126	21.46	1.57
Horse (570×619)	4×1	40.35	4174 /88207	1084	15.69	1.62
	2×2	38.71	6008 / 88207	2109	18.48	1.59

Obviously, in the first image both numbers of valid blocks are closed (7926, 7048); the secret capacity (1941<2572) is dependent on the compression rate (12.91%<18.69%). The second and third images show the numbers of valid blocks (7011<7317, 4174<6008) and compression rates (19.54%<21.46%, 15.69%<18.48%) are proposed to secret capacity(2701<3126, 1084<2109).

Overall, the main time complexity is decided by Huffman's algorithm. Nevertheless, in general, the sorted time complexity is $O(nlogn)$, which is the key cost of Huffman's algorithm. Therefore, the time complexity for this scheme is $O(nlogn)$.

5 Conclusions

This paper presents a novel reversible data hiding scheme for binary images in which a stego image can be recovered to its original image. The proposed method has the following features. First, adopting 4×1 and 2×2 with non-interleaving and equivalent 2-white and 2-black pixel blocks increases visual quality and imperceptibility. Moreover, the SHC scheme with six variances blocks decreases the compression path and secret length. Applying a double senary ($u=e^2$) as the unit of the SHC method increases the compression rate. To increase secret capability and security, the secret data is compressed to a specific senary archived to approximately 1.1 secret bits / 2 bits. Overall, our scheme not only successfully increases the compression rate, secret capacity and confidence but also achieves an image that is well within human visual perception. Moreover, the host image can be reversible through lossless compression, making the secret data extractable. Furthermore, the well-known and effective Huffman compression method is an efficient method requiring only sub-linear computation time.

References

[1] Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for Data Hiding. IBM Systems Journal 35(3,4), 313–336 (1996)

[2] Brisbane, G., Safavi-Naini, R., Ogunbrna, P.: High-capacity Steganography Using a Shared Colour Palette. IEE Proceedings-Vision, Image and Signal Processing 152(6), 787–792 (2005)

[3] Chen, J., Chen, T.S., Cheng, M.W.: A New Data Hiding Method in Binary Image. In: Proceedings of the Fifth International Symposium on Multimedia Software Engineering (ISMSE'03) December 2003, Taichung, Taiwan, pp. 88–93 (2003)

[4] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)

[5] Johnson, N.F.: Information Hiding Steganography and Watermarking – Attacks Countermeasures. Kluwer Academic Publishers, Dordrecht (2001)

[6] Kamstra, L., Henk, J., Heijmans, A.M.: Reversible Data Embedding into Images Using Wavelet Techniques and Sorting. IEEE Transactions on Image Processing 14(12), 2082–2090 (2005)

[7] Kim, H.Y.: New Public-key Authentication Watermarking for Binary Document Images Resistant to Parity Attacks. In: Proceedings of IEEE International Conference on Image Processing, 2005, September, Genoa, Italy. vol. 2, pp. 1074–1077 (2005)

[8] Li, Z.Q., Zhang, H.B.: A New Data Hiding Method in Binary Images. In: Proceedings of the First International Conference Innovative Computing, Information and Control (ICICIC 2006), August 2006, Beijing, China, vol. 3, pp. 66–69 (2006)
[9] Lu, H., Wang, J., Kot, A.C., Shi, Y.Q.: An Objective Distortion Measure for Binary Document Images Based on Human. In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec City, Canada, August 2002. vol. 4, pp. 239–242, (2002)
[10] Mielikainen, J.: LSB Matching Revisited. IEEE Signal Processing Letters 13(5), 285–287 (2006)
[11] Moskowitz, I.S., Longdon, G.E., Chang, L.: A New Paradigm Hidden in Steganography. In: Proceedings of the 2000 Workshop on New Security Paradigms (NSPW'2000), September 2001, Cork, Ireland, pp. 41–50 (2001)
[12] Ni, Z., Shi, Y.Q., Ansari, N., Su, W.: Reversible Data Hiding. IEEE Transactions on Circuits and Systems for Video Technology 16(3), 354–361 (2006)
[13] Pamboukian, S.V.D., Kim, H.Y.: Reversible Data Hiding and Reversible Authentication Watermarking for Binary Images. In: Proceedings of the Sixth Brazilian Symposium on Information and Computer System Security (SBSEG'06) August 2006 Santos, Brazil. (2006) Available at http://www.lps.usp.br/hae/sbseg2006-rdtc.pdf
[14] Potdar, V.M., Han, S., Chang, E.: Fingerprinted Secret Sharing Steganography for Robustness Against Image Cropping Attacks, In: Proceedings of the third IEEE International Conference on Industrial Informatics (INDIN'05), August 2005. Perth, Australia, pp. 10–12 (2005)
[15] Tsai, C.L., Chiang, H.F., Fan, K.C., Chung, C.D.: Reversible Data Hiding and Lossless Reconstruction of Binary Images Using Pair-wise Logical Computation Mechanism. Pattern Recognition 38(11), 1993–2006 (2005)
[16] Tseng, Y.C., Chen, Y.Y., Pan, H.K.: A Secure Data Hiding Scheme for Binary Images. IEEE Transactions on Communications 50(8), 1227–1231 (2002)
[17] Wang, C.C., Chang, C.C., Jan, J.K., Yeh, J.S.: Novel Steganographic Schemes Based on Row-Major Mapping Relation. In: Proceedings of 2006 IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP-2006), December 2006, Pasadena, California, USA, pp. 392–395 (2006)
[18] Wu, M., Liu, B.: Data Hiding in Binary Image for Authentication and Annotation. IEEE Transactions on Multimedia 6(4), 528–538 (2004)

Managing and Searching Distributed Multidimensional Annotations with Large Scale Image Data

Tian Xia[1,⋆], Fusheng Wang[2], Peiya Liu[2], and Sridharan Palanivelu[2]

[1] College of Computer and Information Science
Northeastern University, Boston, MA 02115
tianxia@ccs.neu.edu
[2] Integrated Data Systems Department
Siemens Corporate Research, Princeton, NJ 08540
{fusheng.wang,peiya.liu,sridharan.palanivelu}@siemens.com

Abstract. Advanced imaging applications are producing large-scale image data for medical diagnosis, part visualization and inspection, content retrieval, analysis, e-reporting, and so on. In many cases, service and non-textual annotations are made on top of the raw images to provide essential information on regions of interest, diseases, defects, evaluations, comments, and etc. These applications pose several challenges: i) Large data volume and the latency of data transfer over the Internet leads to a performance bottleneck; ii) The applications need advanced query support such as similarity queries on the multidimensional annotation data; and iii) Local applications need synchronizing data with a remote central database. In our work, we develop a general distributed multimedia data management system that achieves these goals by providing: i) an intelligent multimedia content caching system to support smooth local applications; ii) a loosely coupled extensible multi-indexing server to support different types of multimedia queries including similarity queries; iii) unified multimedia data access interfaces for universal data access; and iv) an integrated architecture that brings these technologies together into a robust system. The system is now successfully used to support image-based inspection and diagnosis applications such as global part inspection and knowledge database guided medical diagnosis.

1 Introduction

Nowadays, advanced imaging instruments make it very easy to produce large-scale image data for medical diagnosis, part visualization and inspection, content retrieval, analysis, e-reporting, and so on. In many applications, annotations are made on top of the raw images to provide additional information on diseases, defects, evaluations, comments, and etc. The annotations can include 2D or 3D markup shapes and additional metric data from the regions of interest, to provide

⋆ Work done while visiting Siemens Corporate Research.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 361–370, 2007.
© Springer-Verlag Berlin Heidelberg 2007

multidimensional semantic information of both markup shapes and features. The multidimensional data together with large image data can then be retrieved to support image-based inspection and diagnosis applications.

With the globalization trend of enterprises, both data and users are highly distributed. For instance, Siemens has more than 430,000 people in 130 Siemens units distributed in 190 countries. Therefore image and annotation management systems have to deal with data transfer and synchronization among distributed data sources.

The distributed environment in data management, and the use of multidimensional annotation information, pose several important and practical challenges in the annotation-based applications developed at Siemens. First, in the distributed environment, large volumes of data need to be transferred back and forth, and applications frequently access remote database tables to retrieve image related metadata and annotations. Therefore, network latency in data transfer and table accesses can lead to the *performance bottleneck*. Second, applications rely on advanced queries, such as *Similarity Queries*, to discover important information from the multidimensional annotations. A DBMS should be able to support such queries, and be *extensible* to support new queries in the future. Third, the distributed environment can cause data inconsistence, which lead to incorrect query results. Local applications need to synchronize data with a remote central database. Meanwhile, for a loosely coupled index system, index data have to get synchronized with data sources.

Unfortunately, many of these challenges, in particular the support of similarity queries, are not well supported in many existing commercial DBMSs (e.g. SAP, Oracle) and free stand-alone multidimensional indexing toolkits. The former is difficult to extend to support additional index structures, and the latter lacks a system to integrate it with databases and applications. For example, most enterprise applications and data in Siemens are managed with SAP, which does not support multidimensional access methods. While Oracle supports spatial data types, it does not support similarity queries and it is difficult to extend it to support new index structures for new queries.

On the other hand, many industrial projects have limitations in choosing the underlying database systems, due to business constraints such as customer experience/requirements and existing systems already used in the company. For example, SAP is the major database used for Siemens to manage data. For another example, we may choose a database system that the customers are most familiar with. Therefore, although some existing systems may partially address some of the above challenges (e.g., IBM DB2 provides user-defined functions and types to possibly answer similarity queries), we have to provide a general system that can be easily adapted to specific business environment.

Thus, we develop a general multidimensional and metric data management system, the *MM-DMS*, for searching diagnostic annotations with large image data. The MM-DMS provides a generic and extendible framework to provide pluggable components to support advanced queries with new access methods and intelligent caching, and can be easily adapted to different underlining database

servers. The MM-DMS consists of component-based multiple layers, and can be customized to include only user required layers. Two essential layers are the *Advanced Indexing System* and the *Intelligent Caching System*. The advanced indexing system contains a multi-index server which is loosely coupled and highly extensible. Advanced queries are supported through the advanced indexing system, and new multidimensional access methods can be easily added into the system. The intelligent caching system provides efficient and consistent performance for the distributed client sites. Based on the caching policies, it automatically caches frequently used data, "intelligently" prefetch data for possible future queries, and upload/download large image data on schedule.

Our MM-DMS provides a comprehensive solution, and can be used to support image-based inspection and diagnosis applications such as global part inspection system and knowledge database based medical diagnosis.

The rest of this paper is organized as follows. Section 2 discusses a user case of the annotation-based applications at Siemens. Section 3 presents the MM-DMS architecture. Section 4 reviews the related work, and finally Section 5 concludes this paper.

2 Application: The Global Inspection System

In this section, we discuss a user case of image and annotation management to support enterprise applications at Siemens. The application is a global inspection system to inspect turbine parts used for power generation on different sites, through 3D annotations on scanned images on the parts.

Siemens sells turbines for power generation worldwide, for example, the installation of power generation equipments covers one-third of all electric power generated in the US. There are multiple turbine parts maintenance sites, including Orlando, Berlin, etc. One major task is to inspect and report on the problems of the parts, to provide exact information for repairs and maintenance.

To effectively manage the image and annotation data, and share diagnosis knowledge globally, two types of databases are needed: a global database and a local database on each site. The global database maintains all the images and annotations from multiple distributed maintenance sites, including scanned 2D images and their calibrations, 3D scanned models and matrix data; markup and measurement data for defects and measurements.

Each local database on a client site serves two purposes: i) a local repository to receive raw images generated from image acquisition hardware; and ii) caching images and annotations for visualization, updates on annotations, and queries.

Each local database is then connected to a part visualization and inspection software – GIS, an application that manipulates images and annotations. GIS provides the following functionalities: receiving scanned images (2D images and 3D matrix and models), 2D and 3D visualization, authoring annotations, uploading data to the central database, and retrieving data from the central database to perform such operations. Figure 1 displays one image file (on the right) and the corresponding 3D model. If there are markups/annotations on the images,

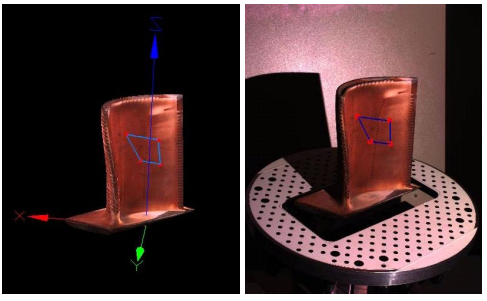


Fig. 1. An example of a 3-D annotation on a turbine blade in the Global Inspection System

they will also be retrieved from the database and overlaid on the 3D models. For example, the right picture in Figure 1 shows a markup on the image, and the left picture shows the overlaid 3D markup on the 3D model.

Markups are created on the regions with defect or damages, and are in the shape of lines, polylines, polygons, or circle, and can be in 2D or 3D. Besides, additional measurements can be done on the markup regions to provide additional inspection information. The markups can later be updated or deleted. These multidimensional markup data will then be used to provide advanced queries. For example, one common query is to find all defect on certain region of some turbine parts. Another interesting query is to find all parts with similar defect in certain region.

3 The MM-DMS System

In this section, we present a system of multidimensional and metric data management for searching diagnostic annotations with large image data, the MM-DMS system for short. We first give our motivations for the MM-DMS, and then present the details of the MM-DMS architecture.

3.1 Motivations

To support the server-client applications we discussed in the previous sections, a comprehensive system is needed to address several important and practical challenges. In the following discussion, we categorize the challenges into three kinds, namely, *Performance Bottlenecks*, *Advanced Query Support*, and *Synchronization Issue*.

Performance Bottlenecks. In our applications, there are three key factors that affect the performances.

First, the volume of image data can be very large. For example, in the Global Inspection System example, the total size of the scan images of just one turbine

cab be as large as several gigabytes. Even with data compression, the total size is still in the order of hundred megabytes. It will take minutes or even hours for the clients to download or upload the image data. Clearly, on-the-fly file operations will cause huge delays and are not preferred.

Second, operations (insert, delete, update and query) on the relational database tables are very frequent. In our applications, relational databases are used to store the detailed metadata of the images and annotations, such as scan details and image parameters. When a user is reading one set of scan data, multiple tables are also queried at the same time (using traditional SQL query) in order to display correct scan information. In most scenarios, the database response time is the accumulation of *many* table operations. Thus, each table operation needs to be efficient enough.

Third, because of the distributed environment and frequent database accesses, the network latency will largely affect the system performance. Figure 2 shows the performance differences between Princeton testing site and Berlin site. Here, we show performances on two frequently accessed tables, the scan_details and image_parameters. Since the remote server is located in Germany, the average performance of one operation in Princeton is two to three times slower than that in Berlin. Moreover, with frequent table operations, the half-second response time of one operation in Princeton is still insufficient for the whole system. Therefore, we aim at providing efficient as well as consistent performances among all sites.

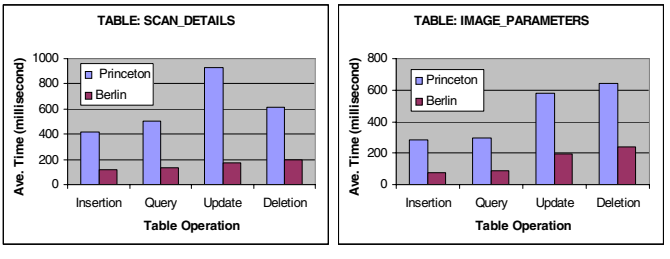


Fig. 2. Comparison between Princeton site and Berlin site of the performances of four table-operations

Advanced Query Support. One of the major tasks of our applications is the *knowledge discovery* based on the multidimensional annotations. Advanced queries are extensively used by the applications to study defect and diseases. i) *Similarity Queries.* In the Global Inspection System, the annotations/markups are defect areas on the turbine blades, and the sample similarity query is to find top-K similar defect markups within a certain region. In medical diagnosis application, similarity queries are issued to find similar Ejection Fraction (EF) contours. ii) *Similarities on Sets.* annotations can be nested, which means one annotation can contain a set of more detailed annotations. Similarity queries on sets are also asked by users. iii) Besides similarity queries, spatial/temporal queries are

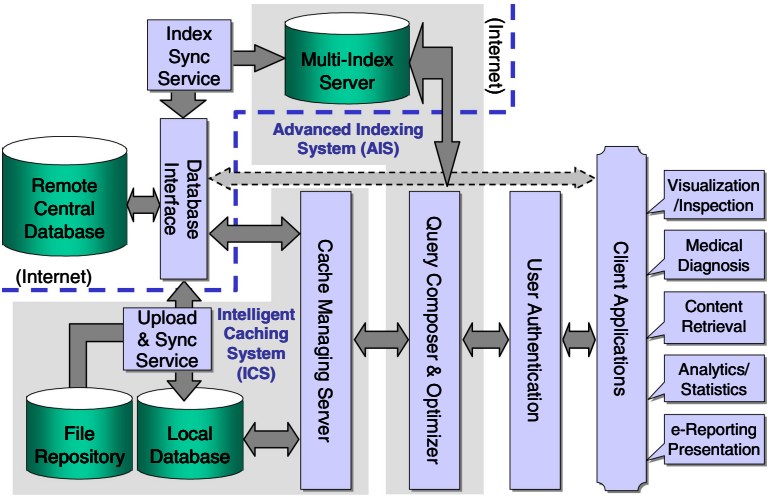


Fig. 3. System architecture of the MM-DMS system on one site

also common in our applications. For example, a user may ask to find all defect annotations on certain regions of all turbine blades, and to find the history of the defect cracks of a turbine similar to the current one.

However, current major commercial DBMS's (e.g. SAP, Oracle) have limited support of advanced similarity queries.

Synchronization Issue. To improve application performances and support advanced queries, the MM-DMS will integrate an intelligent caching system and an advanced indexing system together with commercial DBMS's. Therefore, local (caching) databases and the indexing contents need to be consistent with the central database. Also, newly acquired data sets also need to be migrated to the remote central server. Synchronization among local database, indexing contents, and remote server has to be addressed properly to ensure correct query results.

3.2 System Architecture

We propose the comprehensive MM-DMS system to address all the previous challenges. Connecting client applications with remote server, the MM-DMS integrates all data from distributed local data servers into a centralized data server. Figure 3 shows the full-functional MM-DMS system on one site.

One of the features of the MM-DMS is that it consists of multiple layers of independent components. These layers are transparent to the users and are component-based. That is, they can be easily disabled ("removed") from or enabled ("plugged") into the system without major changes in the system. As in the Figure 3, the long arrow with dashed borders illustrated the component-based idea. We discuss each component in details below.

Client Application. The client application layer shows that our MM-DMS is general and adaptive to many applications, in addition to the two core applications we mentioned before. For example, an advanced customizable reporting application can use the MM-DMS to generate user-specific inspection reports (customized by user queries).

User Authentication. User rights are stored in the server and the authentications is by default done on the server site. Nevertheless, we still provide the client-site authentication in MM-DMS. It is initialized by fetching the rights table from the server. There are two reasons that the user authentication component is used: (1) Reduce the server load and malicious attacks. Many unauthorized queries/accesses can be directly filtered on the client sites, without flooding into the server. (2) User rights are customizable on each client sites. A client site can have its own user-right system inherited from the central server.

Advanced Indexing System (AIS). The loosely-coupled and extensible Advanced Indexing System supports similarity and spatial/temporal queries. The AIS consists of two parts, the Query Composer & Optimizer and the Multi-Index Server. The multi-index server has multiple pluggable index structures for different advanced queries. In Figure 3, the multi-index server is put on the server site. While in practice, it can be put either on the server site or on client sites based on users' requirements. If the multi-index is on the server site, client sites don't have to keep duplicate copies of the index files and save the maintenance overhead. On the other hand, if the multi-index is put on the client site, customizing the indexes becomes possible. The AIS also needs to handle data/index synchronization between the central database and the index server(s). We provide an Index Sync Service on the server side to automatically catch the database changes and update the indexes.

Intelligent Caching System (ICS). To provide efficient and consistent performances for all client sites, we use intelligent caching system in the MM-DMS. It automatically caches frequently used data and downloads pre-schedule large image data. For example, the ICS can pre-download the original images corresponding to the frequently accessed thumbnail images. The ICS effectively resolves the bottlenecks, and significantly improves the whole system performance. The ICS consists of three parts, the Cache Management Server, the local database with file repository, and the Upload/Sync Service. The local database can be implemented by any database system. The file repository stores the large scan files that need to be uploaded to or downloaded from the remote server. The upload/sync service automatically migrate the newly acquired data sets to the remote system, keeping the data consistent between the local system and the remote server.

Database Interface. The local data system and the centralized data server can use their own databases that have totally different interfaces. For example, Oracle accepts SQL statement, while SAP only communicates through BAPI – a SAP

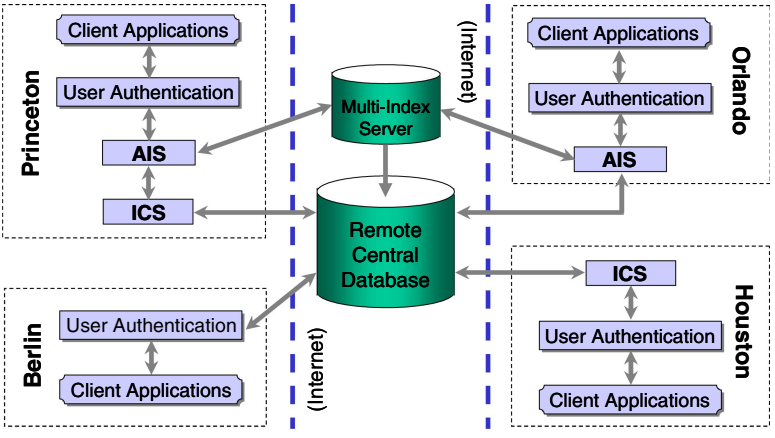


Fig. 4. Overview of the MM-DMS systems with four different sites

specific API. We implement a high-level database interface to hide details of database operations from different venders.

Figure 4 shows the high-level overview of the MM-DMS with four different sites of Siemens. Here, we omit showing the database interface component, which is integrated into the remote server. On each client site, users choose to install desired components of the MM-DMS. Since maintaining ICS and AIS (if on local sites) incurs extra synchronization and disk costs, our highly versatile design of MM-DMS provides the ability to tailor our system to satisfy various client requirements and achieves the best efficiency. Clients can install either or none of the two systems. For example, in Figure 4, the minimum system is installed on the Berlin site, where only simple queries are used and local resources are limited. Thus no overhead cost is spent on caching and indexing.

To summarize this section, we illustrate the query processing flow in our full-functional MM-DMS system. Suppose a user of a client application issues a query. The user authentication layer will decide whether the query is accepted or rejected based on the user's rights. The accepted query is then send to the query composer of the AIS. Depending on whether this is a simple SQL query or an advance similarity query, the query composer either directly sends the (SQL) query to the next level, or composes it to appropriate query object for the Multi-Index Server. Assume that it is a simple SQL query and is passed to the intelligent caching system. The ICS decides to retrieve the actual records from local cache or the remote data server.

4 Related Work

Multimedia data management has been extensively studied in the past [17]. Recently, Gao *et al.* proposed a distributed data management system for large-scale volume visualization [10]. The system is optimized for distributed storage

and visualization of large volume data, and support of markups on images is not studied. A grid-based approach [13] is proposed to manage distributed medical PACS systems, where distributed storage of DICOM images is the focus of the paper.

Similarity search based on feature vectors are most common in multimedia databases. Many feature-based approaches have been proposed in the past, such as QBIC (Query by Image Content) [9], shape matching retrieval [15,14], and etc. To efficiently index high dimensional feature vectors and support fast similarity search, White and Jain [18] proposed the SS-tree (Similarity Search Tree). The SS-tree uses spheres as bounding regions, which is the shape of a similarity query.

Supporting new access methods in DBMS is a major cost for DBMS vendors. Nevertheless, several DBMS are supporting spatial data by extending current DBMS through extenders, blades, or catridges. IBM provides three systems to support spatial data: DB2 Spatial Extender [1,6] with DB2 Grid Index, Spatial DataBlade [3] with R-trees Index[11,7], and Geodetic DataBlade [2] with R-Trees. Oracle Spatial [4] supports spatial data with R-Tree and Quad-Tree indexes. PostGIS [5] is an open source DBMS that adds support for geographic objects to the PostgreSQL with R-Tree index integration. While spatial data is well supported with these DBMS, tightly integrating additional index methods to support complex multidimensional data is a difficult goal for DBMS vendors.

There are also several free softwares that support multidimensional indexes. They are stand-alone tools and/or specific to certain applications. GiST [12] is an index structure that supports an extensible set of queries and data types. It is designed to address the data and query extensibility. XXL [8]is a java library that provides an infrastructure for implementing advanced query processing functionality. Generic Multidimensional Scientific Indexing Library (GMIL) [16] uses semantic information to build multidimensional indexes, and groups items into properly sized chunks to maximize disk I/O bandwidth.

5 Conclusions

In this paper, we present a system MM-DMS with flexible support of multidimensional indexing and intelligent caching to manage and search textual and non-textual annotations for image-based inspection and diagnosis with large distributed archived data. With pluggable components, the system can be easily extended to support new advanced queries and can be quickly adapted to different business environments. Our system consists of: i) a user authentication component to enhance security and reduce network load from unauthorized queries; ii) a loosely coupled and highly extensible multi-indexing system to support advanced similarity queries, and in particular, various methods to synchronize index data and the data sources; iii) an intelligent multimedia content caching system to intelligently prefetch/cache data through a set of user customizable caching policies; iv) unified multimedia database access interfaces, and v) an integrated component-based architecture that brings these technologies together to support multiple applications. The system is deployed in Siemens Power

Generation at both Orlando and Berlin for global inspection of power generation parts, and is being planned to support database-guided medical diagnosis with large-scale image data and their annotations.

References

1. DB2 Spatial Extender: http://www-306.ibm.com/software/data/spatial/
2. IBM Informix Geodetic DataBlade: http://www.ibm.com/software/data/informix/blades/geodetic/
3. IBM Informix Spatial DataBlade Module: www.ibm.com/software/data/informix/blades/spatial/
4. Oracle Spatial: http://www.oracle.com/technology/products/spatial/
5. PostGIS.: http://www.postgis.org/
6. Adler, D.W.: IBM DB2 Spatial Extender - Spatial data within the RDBMS. In: VLDB (2001)
7. Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*-tree: An Efficient and Robust Access Method for Points and Rectangles. In: ACM SIGMOD, pp. 322–331 (1990)
8. van den Bercken, J., Blohsfeld, B., Dittrich, J.-P., Krämer, J., Schäfer, T., Schneider, M., Seeger, B.: XXL- A Library arrproach to supporting efficient implementations of advanced Database queries. In: VLDB, pp. 39–48 (2001)
9. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems 3(3-4), 231–262 (1994)
10. Gao, J., Huang, J., Johnson, C.R., Atchley, S., Kohl, J.: Distributed Data Management for Large Volume Visualization. In: IEEE Visualization Conference, IEEE Computer Society Press, Los Alamitos (2005)
11. Guttman, A.: R-trees: A Dynamic Index Structure for Spatial Searching. In: ACM SIGMOD, pp. 47–57. ACM Press, New York (1984)
12. Hellerstein, J.M., Naughton, J.F., Pfeffer, A.: Generalized Search Trees for Database Systems. In: VLDB, pp. 562–573 (1995)
13. Huang, H.K., Zhang, A., Liu, B., Zhou, Z., Documet, J., King, N., Chan, L.W.C.: Data Grid for Large-Scale Medical Image Archive and Analysis. In: ACM Multimedia, ACM Press, New York (2005)
14. Korn, F., Sidiropoulos, N., Faloutsos, C., Siegel, E., Protopapas, Z.: Fast Nearest Neighbor Search in Medical Image Databases. In: VLDB, pp. 215–226 (1996)
15. Mehrotra, R., Gary, J.E.: Feature-Based Retrieval of Similar Shapes. In: IEEE ICDE, pp. 108–115. IEEE Computer Society Press, Los Alamitos (1993)
16. Nam, B., Sussman, A.: Improving Access to Multi-dimensional Self-describing Scientific Dataset. In: IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2003), pp. 172–179 (2003)
17. Subrahmanian, V.S.: Principles of Multimedia Database Systems. Morgan Kaufmann, San Francisco (1998)
18. White, D.A., Jain, R.: Similarity Indexing with the SS-tree. In: IEEE ICDE, pp. 516–523. IEEE Computer Society Press, Los Alamitos (1996)

A Closed-Form Solution of Reconstruction from Nonparallel Stereo Geometry Used in Image Guided System for Surgery

Jianhua Wang and Yuncai Liu

Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University, Shanghai 200240, China
{jian-hua.wang, whomliu}@sjtu.edu.cn

Abstract. In this paper, a closed-form solution of 3D reconstruction from nonparallel stereo geometry is derived theoretically. From a pair of conjugate image points in an arbitrarily configured stereo system that has been calibrated, one can directly reconstruct the 3D scene point using the closed-form solution, bypassing the process of rectifying the images or iteratively optimizing. Experiment results from both simulated data and real images validate the closed-form solution. Practical application to Image Guided System for surgery shows that the closed-form solution has advantages to improve efficiency and accuracy of 3D reconstruction from nonparallel stereo system in comparison with the conventional method that employs algorithm for the standard parallel axis stereo geometry.

1 Introduction

In the computer vision community, structure from stereo has always been a major portion of researches. The main steps involved in stereo process are calibration, correspondence, and reconstruction [1-4]. Calibration is the process to determine internal and external parameters of a camera system. Accurate calibration is necessary in order to relate image information to an external world coordinate system. Some well-known toolkits for camera calibration as well as relevant references can be found in [5]. The correspondence problem consists of determining the location in each image that is the projection of the same object in space. Some good correspondence algorithms are discussed in [6]. In this paper, we focus on the reconstruction problem.

The reconstruction problem consists of determining three dimensional structure from two given images based on known camera geometry. For conventional parallel axis stereo geometry, which includes a pair of cameras with same intrinsic parameters and parallel camera coordinate systems only separated by a horizontal distance, it is well known that the coordinates of a scene point in any of the camera frames can be easily obtained from its disparities [7-12]. In practice, however, most stereo systems are nonparallel configuration. On the one hand, it is difficult to build a standard parallel axis stereo system. On the other hand, some applications require the optical axes to be convergent, e.g. in Image Guided System for surgery. Approaches involved in

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 371–380, 2007.
© Springer-Verlag Berlin Heidelberg 2007

reconstruction from nonparallel stereo so far fall into two classes. One class back-projects the lines joining the project center and the image point in the left and right camera respectively into space and finds the intersection of the rays [7-12]. In this class, it is necessary to solve a set of equations, and an iterative method is often used to find an optimized solution. Due to the problems of initial value estimation and local minima, this class of methods is seldom used in practice. The other class rectifies arbitrary stereo image pairs to make them satisfy standard parallel axis stereo geometry and then employs the algorithm for standard parallel axis stereo geometry [13-16]. Now the second class of methods is widely used in computer vision, which is referred to as conventional method hereinafter. However, when the conventional method is used, the images have to be rectified to satisfy the requirement for standard parallel axis stereo geometry. This is not only a time consuming process, but also can bring in errors due to truncation of intermediate results. Therefore, it is desirable to have a closed-form solution of reconstruction from nonparallel stereo geometry.

It is the motivation of this paper to find a closed-form solution of reconstruction from nonparallel stereo geometry. Our research belongs to the first class of approaches. The ideas are well discussed in [1-4,7-12], but no closed-form solution has been presented. In this paper, we derive a closed-form solution of reconstruction from nonparallel stereo geometry, which can bypass the process of rectifying the images or iteratively optimizing. Experiment results from both simulated data and real images validate the closed-form solution.

The remainder of this paper is structured as follows: Section 2 gives a theoretic derivation of the closed-form solution in detail. Section 3 is some experiment results from simulated data, real images and practical application. Section 4 is our conclusion.

2 Theoretic Derivation

Fig.1 shows an arbitrarily configured stereo system. Suppose intrinsic matrices of its left camera and right camera are A_l and A_r respectively,

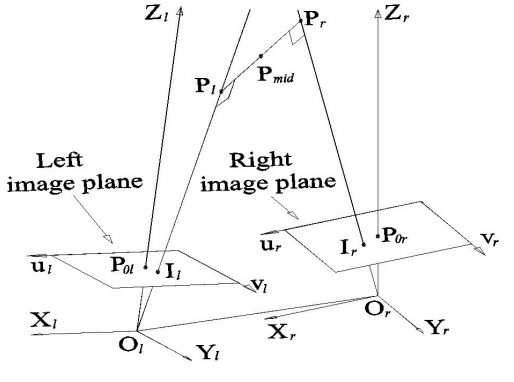


Fig. 1. Nonparallel stereo

$$A_l = \begin{bmatrix} \alpha_l & 0 & u_{l0} \\ 0 & \beta_l & v_{l0} \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

$$A_r = \begin{bmatrix} \alpha_r & 0 & u_{r0} \\ 0 & \beta_r & v_{r0} \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

and the rotational matrix and translational vector of the right camera with respect to the left camera are R and T.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{3}$$

$$T = [t_1 \quad t_2 \quad t_3] \tag{4}$$

Suppose $I_l(u_l, v_l)$ and $I_r(u_r, v_r)$ are a pair of conjugate points on the left image plane and the right image plane, with $N_l(U_l, V_l, 1)$ and $N_r(U_r, V_r, 1)$ being their normalized homogenous coordinates respectively, then

$$N_l = [U_l \quad V_l \quad 1]^T = \left[\frac{u_l - u_{l0}}{\alpha_l} \quad \frac{v_l - v_{l0}}{\beta_l} \quad 1 \right]^T \tag{5}$$

$$N_r = [U_r \quad V_r \quad 1]^T = \left[\frac{u_r - u_{r0}}{\alpha_r} \quad \frac{v_r - v_{r0}}{\beta_r} \quad 1 \right]^T \tag{6}$$

where $P_{0l}(u_{l0}, v_{l0})$ and $P_{0r}(u_{r0}, v_{r0})$ are principal points of the left camera and right camera respectively, with α_l, β_l and α_r, β_r being their principal lengths. Let us denote 3D coordinates of a point on the line $O_l I_l$ as $P_l(X_l, Y_l, Z_l)$ in the left camera frame, while 3D coordinates of a point on the line $O_r I_r$ as $P_r(X_r, Y_r, Z_r)$ in the right camera frame, then

$$P_l = [X_l \quad Y_l \quad Z_l]^T = [U_l \quad V_l \quad 1]^T Z_l \tag{7}$$

$$P_r = [X_r \quad Y_r \quad Z_r]^T = [U_r \quad V_r \quad 1]^T Z_r \tag{8}$$

3D coordinates of the point $P_r(X_r, Y_r, Z_r)$ on the line $O_r I_r$ in the right camera frame can be transformed into coordinates $P_l^r(X_l^r, Y_l^r, Z_l^r)$ in the left camera frame. The Euclid distance between the point $P_l(X_l, Y_l, Z_l)$ and $P_l^r(X_l^r, Y_l^r, Z_l^r)$ can be expressed as

$$d = \left\| P_l - P_l^r \right\| = \left\| P_l - R P_r - T \right\| \tag{9}$$

Substitute (3), (4), (7) and (8) into (9), we get

$$d^2 = (U_l Z_l - J_x Z_r - t_1)^2 + (V_l Z_l - J_y Z_r - t_2)^2 + (Z_l - J_z Z_r - t_3)^2 \tag{10}$$

where

$$\begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} U_r \\ V_r \\ 1 \end{bmatrix} \tag{11}$$

i.e. $$J = RN_r \tag{12}$$

Because d is the Euclid distance between the point $P_l(X_l, Y_l, Z_l)$ and $P_l^r(X_l^r, Y_l^r, Z_l^r)$, d^2 has only one minimum. From

$$\begin{bmatrix} \frac{\partial d^2}{\partial Z_l} \\ \frac{\partial d^2}{\partial Z_r} \end{bmatrix} = 2\begin{bmatrix} U_l & V_l & 1 \\ -J_x & -J_y & -J_z \end{bmatrix}\begin{bmatrix} U_l Z_l - J_x Z_r - t_1 \\ V_l Z_l - J_y Z_r - t_2 \\ Z_l - J_z Z_r - t_3 \end{bmatrix} = 0 \tag{13}$$

we get

$$Z_l = \frac{(N_l^T J)(J^T T) - (N_l^T T)(J^T J)}{(N_l^T J)^2 - (J^T J)(N_l^T N_l)} \tag{14}$$

$$Z_r = \frac{(N_l^T N_l)(J^T T) - (N_l^T T)(N_l^T J)}{(N_l^T J)^2 - (J^T J)(N_l^T N_l)} \tag{15}$$

Substitute (12) into (14) and (15), we can get depth Z_l and Z_r of point $P_l(X_l, Y_l, Z_l)$ and $P_r(X_r, Y_r, Z_r)$ in the left camera frame and the right camera frame respectively.

$$Z_l = \frac{(N_l^T R N_r)(N_r^T R^T T) - (N_l^T T)(N_r^T N_r)}{(N_l^T R N_r)^2 - (N_r^T N_r)(N_l^T N_l)} \tag{16}$$

$$Z_r = \frac{(N_l^T N_l)(N_r^T R^T T) - (N_l^T T)(N_l^T R N_r)}{(N_l^T R N_r)^2 - (N_r^T N_r)(N_l^T N_l)} \tag{17}$$

Intersections P_l and P_r of the common perpendicular $P_l P_r$ with rays $O_l I_l$ and $O_r I_r$ expressed in the left camera frame and the right camera frame respectively can be obtained.

$$P_l = \frac{(N_l^T R N_r)(N_r^T R^T T) - (N_l^T T)(N_r^T N_r)}{(N_l^T R N_r)^2 - (N_r^T N_r)(N_l^T N_l)} N_l \tag{18}$$

$$P_r = \frac{(N_l^T N_l)(N_r^T R^T T) - (N_l^T T)(N_l^T R N_r)}{(N_l^T R N_r)^2 - (N_r^T N_r)(N_l^T N_l)} N_r \tag{19}$$

We can take the midpoint P_{mid} of the common perpendicular $P_l P_r$ as an estimation of the scene point, then 3D coordinates of the scene point in the left camera frame and the right camera frame can be expressed as

$$\hat{P}_l = \frac{P_l + RP_r + T}{2} \tag{20}$$

$$\hat{P}_r = \frac{P_r + R^T P_l - R^T T}{2} \tag{21}$$

For an arbitrarily configured stereo system, the intrinsic parameters A_l and A_r as well as the extrinsic parameters R and T are known after calibration. From a pair of corresponding points $I_l(u_l, v_l)$ and $I_r(u_r, v_r)$, we can compute their normalized coordinates N_l and N_r using formula (5) and (6). Then the end points P_l and P_r of the common perpendicular can be found from formula (18) and (19). Finally, the scene point can be estimated with formula (20) and (21).

If there is no noise and the match between the corresponding points is exact, the end points P_l and P_r of the common perpendicular are coincident, therefore $\hat{P}_l = P_l$ and $\hat{P}_r = P_r$. Especially, for a standard parallel axis stereo geometry, $\alpha_l = \alpha_r = \alpha$, $T = [t_1 \quad 0 \quad 0]^T$ and R is a unit matrix, therefore formula (16) and (17) can be reduced as

$$Z_l = Z_r = \frac{t_1 \alpha}{u_l - u_r} \tag{22}$$

which is consistent with the result from standard parallel axis stereo geometry.

Formula (20) and (21) together with (16), (17), (18) and (19) can be used as a closed solution for reconstruction from an arbitrarily configured stereo geometry.

3 Experiments and Discussion

3.1 Experiments with Simulated Data

In the experiments with simulated data, the stereo configuration is supposed as shown in Fig.2, where the intrinsic parameters of left camera and right camera are A_l = [1000 0 390; 0 1000 270; 0 0 1] and A_r = [900 0 370; 0 900 290; 0 0 1], and the rotational matrix and translation vector of the right camera with respect to the left camera are R = [0.9563 -0.1307 0.2615; 0.1443 0.9890 -0.0334; -0.2542 0.0697 0.9646] and T = [300; -50; 80]. In front of the cameras, a three dimensional array of points are positioned, which include 10×10×10 = 1000

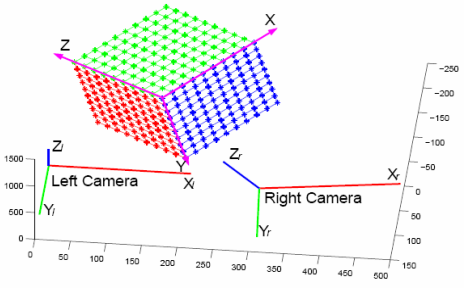
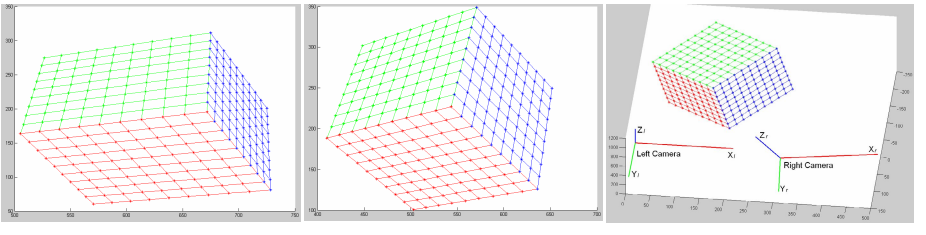


Fig. 2. Stereo configuration in experiments with simulated data

points. The rotational matrix and translation vector of the array frame XYZ with respect to the left camera is Ra = [0.7220 0.2780 0.6336; 0.2780 0.7220 -0.6336; -0.6336 0.6336 0.4440] and Ta = [100; 0; 1000]. The distance between two adjacent points in the array along X, Y and Z direction is 40mm. Therefore, coordinates of each point in the array can be determined and referred to as ground truth. We reconstruct the array of points and compare the result with the ground truth.

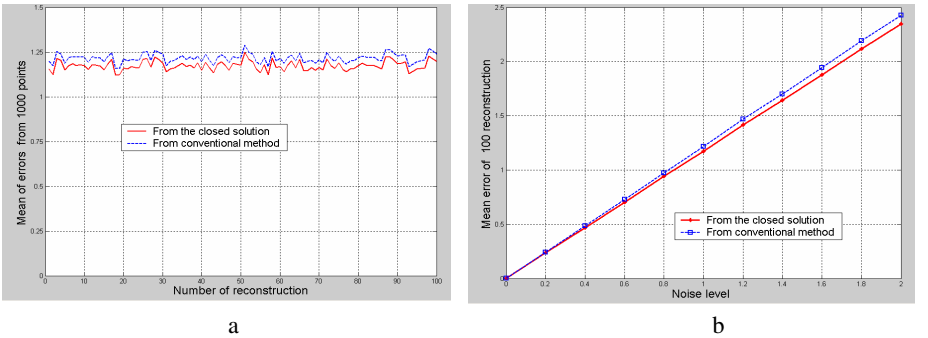
At first, the array is projected onto the image plane of the left camera and right camera respectively, forming the left image as shown in Fig.3a and the right image as shown in Fig.3b. In both the left image and right image, normally distributed noise with mean 0 and variance σ^2 is added. Then the array is reconstructed from the left image and right image with noise added using the closed-form solution, and the result is shown in Fig.3c.

In order to have a comparison, we also reconstruct the array of points using the conventional method, i.e. the method in which the images are first warped to satisfy the standard stereo geometry and then the reconstructing algorithm for the standard parallel axis stereo system is employed. We reconstruct the array of points 100 times using the closed-form solution and the conventional method respectively at noise level σ2 = 1, the average errors of all the 1000 points in the array for each reconstruction are shown in Fig.4a, from which it can be seen that the averages of errors from the closed-form solution are all less than those from the conventional method.



a b c

Fig. 3. Reconstructed array of points at noise level σ2 = 1



a b

Fig. 4. Average errors of reconstructed points in the array

At each noise level from σ2 = 0 to σ^2 = 2.0 with incremental σ^2 = 0.2, we reconstruct the array of points 100 times using the closed-form solution and the conventional method respectively, and the results are shown in Fig.4b, from which it can be seen that: (1) when the noise level is σ^2 = 0, the average errors from both the closed-form solution and the conventional method are zero, which means that the closed-form solution is correct theoretically; (2) the average errors from both the closed-form solution and the conventional method are in proportion to the noise level, which is reasonable; (3) the average errors from the closed-form solution are always less than those from the conventional method, and the difference between them is in proportion to the noise level, which illustrates that the closed-from solution is more accurate than the conventional method. The reason lies in that the closed-form solution comes from the optimizing process and therefore can combat noise, while the conventional method comes from just similar triangle without consideration of noise.

3.2 Experiments with Real Images

In experiments with real images, we take 20 pairs of images of a checkerboard using an arbitrarily configured but calibrated stereo system. The image pairs are shown in Fig.5a and Fig.5b. The intrinsic matrices of the left camera and right camera are A_l = [2960.0605 0 325.3016; 0 2942.9863 278.0018; 0 0 1] and A_r = [2984.8064 0 391.6716; 0 2968.9773 313.3299; 0 0 1]. The first order radial distortion coefficient of the left lens and right lens are K_{1l} = -1.01603 and K_{1r} = -1.10770. The rotational matrix and the translation vector of the right camera with respect to the left camera are R = [0.9381 -0.0061 0.3462; 0.0119 0.9998 -0.0145; -0.3461 0.0177 0.9380] and T = [583.5223;-3.0486; 30.5348]. There are 28 × 20 squares in the checkerboard, with the side length of each square being 14mm.

After extracting the corner points from the left image and the right image of the checkerboard, we reconstruct the coordinates of the corner points in the left camera frame. Fig.6a and Fig.6b are the left image and the right image of the sixth pair of images, showing the corner points extracted. Fig.6c is the reconstructed result from the left image and the right image as shown in Fig.6a and Fig.6b.

The ground truth of the corner points in the left camera frame are not known, but the side length of each square in the checkerboard is known. After reconstruction, we

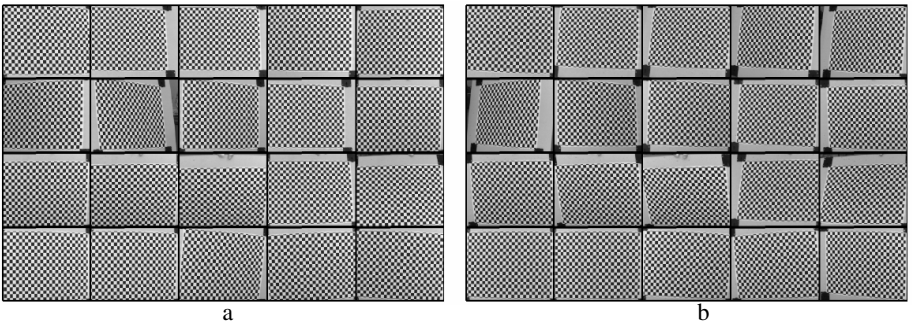
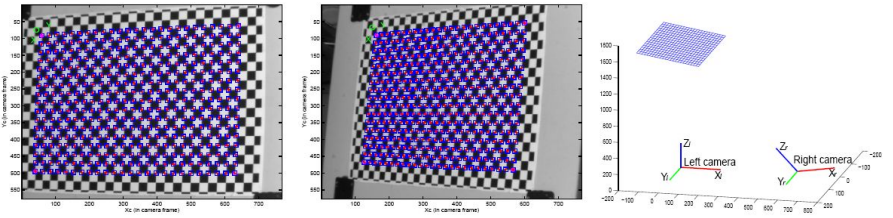
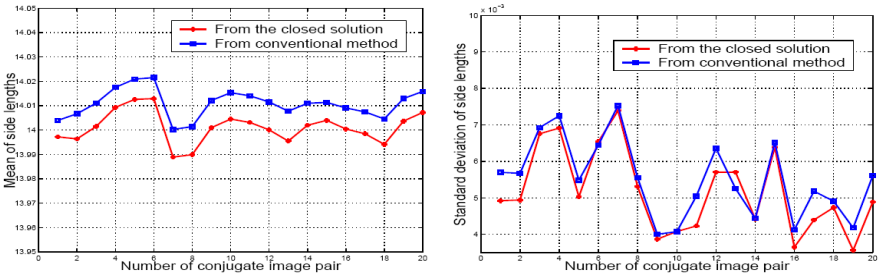


Fig. 5. Images of a checkerboard in the left and right cameras



a b c

Fig. 6. The reconstructed result from the 6th pair of images



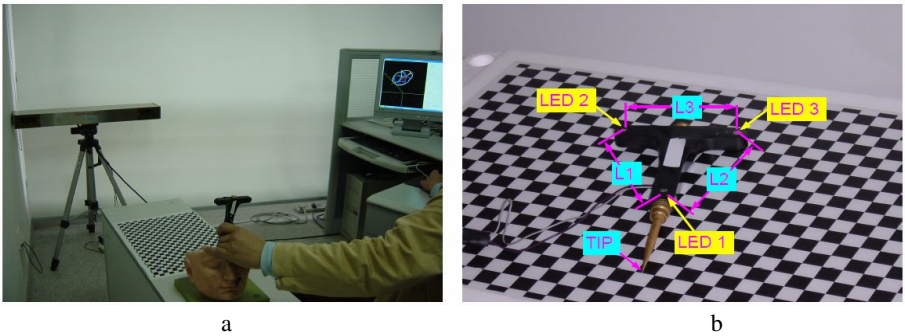
a b

Fig. 7. Average side lengths of each square reconstructed

compute each side length between two adjacent corner points, and the mean and standard deviation of the side lengths from each reconstruction of the 20 pairs images are shown in Fig.7a and Fig.7b, from which it can be seen that the error of reconstruction is small enough to be used in practical application.

3.3 Practical Application

Based on experiments of both simulated data and real images, we have used the closed-form solution in our Image Guided System for surgery (IGS) as shown in

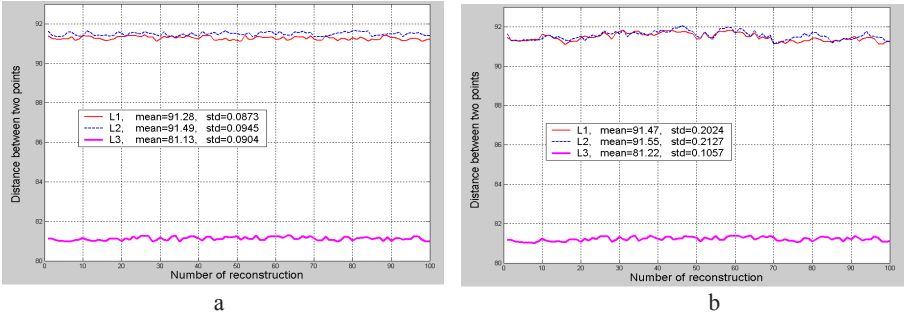


a b

Fig. 8. Image Guided System for surgery

a b c

Fig. 9. Reconstruction result from the 36th pair of images



a b

Fig. 10. Reconstruction accuracy of marker points on the probe

Fig.8a. In the IGS, the stereo is used to track the position and pose of a probe as shown in Fig.8b, on which there are three infrared LEDs. When used, the LEDs light,and the left camera and right camera capture their images. After reconstructing the centers of the three LEDs, the position and pose of the probe can be obtained.

The intrinsic and extrinsic parameters of the stereo system are the same as in the experiments with real images. Fig.9a and Fig.9b are the 36th pair of images captured in test, and Fig.9c is the reconstructed probe from the images in Fig.9a and Fig.9b.

Before we obtained the closed-form solution, the reconstruction was carried out using the conventional method. In order to have a contrast, we take 100 pairs of images of the probe at different position and pose and reconstruct the probe using the closed-form solution and the conventional method. Although the actual positions and poses of the probe with respect to the left camera are not known, the distances L1, L2 and L3 between two LEDs as shown in Fig.8b are constant. We can compute the distances L1, L2 and L3 in the 100 reconstructions and compare their variance. Fig.10a is the result from using the closed-form solution, and Fig.10b is from the conventional method. It can be seen from Fig.10a and Fig.10b that the variance from the closed-form solution is less than that from the conventional method. The result of comparison is consistent with that from simulated experiments.

4 Conclusion

In this paper, we derive a closed-form solution of reconstruction from nonparallel stereo system theoretically. From a pair of conjugate image points in an arbitrarily configured stereo system that has been calibrated, one can directly reconstruct the 3D scene point using the closed-form solution, bypassing the process of rectifying the images or iteratively optimizing. Experiment results from both simulated data and real images validate the closed-form solution. Practical application of the closed-form solution to Image Guided System for surgery shows that it has benefits to improve efficiency and accuracy of reconstruction in comparison with the conventional method employing the standard parallel axis stereo geometry.

References

1. Barnard, S.T., Fischler, M.A.: Computational Stereo. ACM Computing Surveys 14, 553–572 (1982)
2. Dhond, U.R., Aggarwal, J.K.: Structure from Stereo: A Review. IEEE Trans. Systems, Man, and Cybernetics 19, 1489–1510 (1989)
3. Koschan, A.: What is New in Computational Stereo Since 1989: A Survey of Current Stereo Papers. Technical Report 93-22, Technical Univ. of Berlin (1993)
4. Brown, M.Z., Burschka, D., Hager, G.D.: Advances in Computational Stereo. IEEE Transaction on Pattern Analysis and Machine Intelligence 25(8) (August 2003)
5. http://www.vision.caltech.edu/bouguetj/calib_doc/
6. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two- Frame Stereo Correspondence Algorithms. International Journal of Computer Vision 47(1/2/3), 7–42 (April-June 2002)
7. Jain, R., Kasturi, R., Schunck, B.G.: Machine Vision. McGraw-Hill, New York (1995)
8. Ma, S., Zhang, Z.: Computer Vision, Chinese version, CAS press (1998)
9. Guangyou, X.: Computer vision. Chinese version. TsingHua University Press, Beijing (1999)
10. Yunde, J.: Machine Vision. Chinese version. Science Press, Beijing (2000)
11. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, Cambridge (2000)
12. Forsyth, D., Ponce, J.: Computer Vision: a modern approach, United States Ed edition. Prentice Hall, Englewood Cliffs (2002)
13. Silva, L.C., Petraglia, A., Petraglia, M.R.: Stereo Vision System for Remote Monitoring and 3-D Reconstruction, In: IEEE International Symposium on Industrial Electronics, June 9-11 2003. Rio de Janeiro, Brazil (2003)
14. Kim, W.S., Ansar, A.I., Steele, R.D.: Performance Analysis and Validation of a Stereo Vision System. In: IEEE SMC 2005 (October 2005)
15. Liu, H.: Research on Robot Assisted Surgical System for Total Knee Replacement and Key Technologies. PhD. Thesis, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University (March 2005)
16. Shi, F.: Experimental System of Hand-Eye Robot Aided Surgery for Total Knee Replacement. PhD. Thesis, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University (March 2006)

SIEVE—Search Images Effectively Through Visual Elimination

Ying Liu, Dengsheng Zhang, and Guojun Lu

Gippsland School of Info Tech, Monash University, Churchill, Victoria, 3842
{dengsheng.zhang, guojun.lu}@infotech.monash.edu.au

Abstract. Existing Web image search engines index images by textual descriptions including filename, image caption, surrounding text, etc. However, the textual description available on the Web could be ambiguous or inaccurate in describing the actual image content and some images irrelevant to user's query are also returned by text-based search engines. In this paper, we propose to integrate the existing text-based image search engine with visual features, in order to improve the performance of pure text-based Web image search. The proposed algorithm is named SIEVE. Practical fusion methods are proposed to integrate SIEVE with contemporary text-based search engines. In our approach, text-based image search results for a given query are obtained first. Then, SIEVE is used to filter out those images which are semantically irrelevant to the query. Experimental results show that the image retrieval performance using SIEVE improves over Google image search significantly.

Keywords: CBIR, RBIR, SIEVE, web image searching, decision tree.

1 Introduction

Due to the explosive growth of the World Wide Web, nowadays mammoth amount of images are available on the Web. However, unlike text documents, images on the Web are not categorized properly according to their contents. This makes the Web a large unstructured image database. In order to make use of this vast database of Web images, it is desirable to have an effective Web image search engine that meets users' need [1].

As Web images usually come with HTML source code including textual descriptions, existing Web image search engines such as Google, Yahoo! and AltaVista index images by textual descriptions including filename, image caption, surrounding text, etc [1], [2], [3], [4]. Google, for example, has indexed over 425 million images and allows users to search this index using textual queries. However, the textual description available on the Web can be ambiguous or inaccurate in describing the actual image content, irrelevant images are also returned by text-based search engines. For instance, Web image filenames may be misleading, surrounding text might not describe the content of an image, or a word may have multiple meanings resulting in irrelevant retrievals by matching the right word but with the wrong meaning. Research has been done to refine the search results returned by existing text-based image search engines and deliver the images which are more relevant to user queries.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 381–390, 2007.
© Springer-Verlag Berlin Heidelberg 2007

A hierarchical clustering technique using visual, textual and link information is proposed in [5]. This algorithm can organize the results returned by text-based image search engines into different semantic categories to help users in quickly finding the desired images. Tested on top 500 images returned for a few queries by text-based image search, it is found that the re-organized image list better facilitates users in finding their desired images. This algorithm requires that the number of clusters in which images are to be categorized should be known beforehand, which is still an open research challenge. The co-ranking framework proposed in [6] aims to re-rank the Web images returned by text-based search engines, such that the images irrelevant to user query are moved to the tail of the list. The framework segregates the low-level image features into disjoint subsets (views). One-Class SVM (support vector machine) is used as the learning algorithm in each view and the output includes the ranking score for all the images considered. Then, the results from multiple views are combined to improve the final ranking of images according to their relevance to the user query. Experimental results on the top 10 images returned by Google image search shows that the framework improves the precision of Web image search. However, the scalability of the framework is questionable and further work is needed to analyse its re-ranking performance for more images returned by Google.

In previous work, we have developed content-based techniques for image database retrieval, including salient low-level image feature extraction techniques and a methodology to derive high-level concepts from such low-level image features [7], [8], [9], [10], [11]. Our experimental results on natural scenery image database confirm the effectiveness of our methods on stand alone image database where no textual information is available. However, in order to apply it to Web image search, further work is necessary to combine visual image content with the textual description available on the Web. Existing text-based search engines, such as Google, are among the most widely used web tools and have done a good job in Web image search. It is more practical and advantageous to improve the retrieval performance by further refining their image search results than developing a completely new image search engine.

In this paper, we propose to integrate the existing text-based image search engine with the region based image retrieval (RBIR) technique we have developed (SIEVE), in order to improve the performance of text-based Web image search. Practical fusion methods are proposed to integrate SIEVE into contemporary text-based search engines. The rest of the paper is organized as following. In Section 2, the SIEVE algorithm is described in details. Section 3 shows experimental results and compares SIEVE with Google image search. In Section 4, we propose three methods of integrating SIEVE with existing text based image search engines and discuss the scalability issue. Finally, Section 5 concludes the paper.

2 The SIEVE Algorithm

In this section, we present the details of our Web image search algorithm–SIEVE, which integrates image semantic learning into Web image search. The idea of using SIEVE is very similar to object classification done by a human being. First, objects of interest are roughly distinguished from other very different objects either manually or

through certain hand tools. Then, the collected objects are subject to visual inspection to confirm each object of interest from unwanted objects.

Fig. 1 shows the block diagrams of the proposed retrieval system and the SIEVE module. The input to the system is a keyword query submitted by the user, the system first comes out a ranked list of retrieved images given by the text based image search engine. The list of retrieved images by text based image search engine is then input to SIEVE for further analysis. For each image in the list, SIEVE first segments it into different regions. Then, color and texture features of each region are extracted using the methods presented in [7], [8] and [9], [10] respectively. The region color feature is the dominant color in HSV space and the region texture feature is the Gabor feature obtained using a novel padding algorithm [10]. In [11], we have developed a semantic template based decision tree reasoning algorithm and derived a set of decision rules to learn a set of concepts in natural scenery images. Using these decision rules, the low-level features of a region are mapped to semantic concepts.

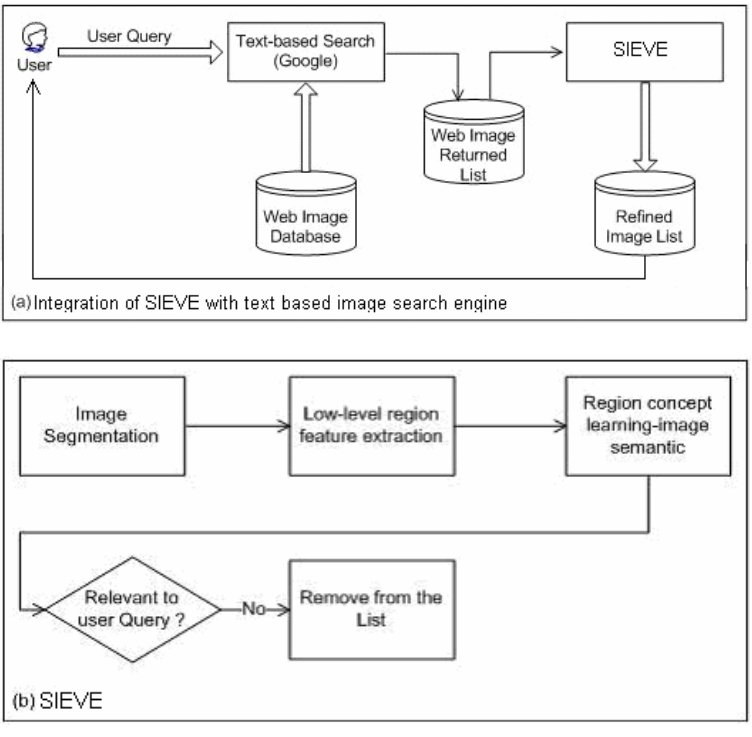


Fig. 1. Block diagrams of the retrieval system and the SIEVE module

Each query in our experiments corresponds to a high-level concept. It is assumed that each image has at least one dominant region representing its semantic meaning. Given a user query, for each image in the returned list, if any region it contains is relevant to the query concept, then the image is considered relevant to the query. If none of the region it contains is relevant to the query, the image is considered

irrelevant to the query and is eliminated from the list. Refer to the human classification analogy, the retrieval list given by the text based image search engine is the rough selection step. The SIEVE analysis is the visual inspection because it refines the list by using visual features which are also what human being perceives the images.

3 Experiments and Results Analysis

This section presents the experimental results of SIEVE for Web image search and compares its performance with Google image search results. The process simulates the integration of the SIEVE module with the Web search engine. At this moment, the system works with natural scenery images because the underlying decision tree is trained using natural images. Therefore, the queries in our experiments are limited to this class of images currently.

3.1 Web Image Collection and Performance Evaluation

To test the retrieval performance of SIEVE, 10 queries are selected, including *mountain, beach, building, firework, flower, forest, snow, sunset, tiger* and *sea*. Google image search can return up to thousands of images for a query, however, users are usually only interested in the first few pages and unlikely to take time to view the remaining pages. Therefore, for each query, the top 100 images are downloaded from the first 5 pages.

For a given query, each image in the returned list is segmented into different regions using JSEG [7], and the regions with size over 5% of the entire image are selected. Then, low-level features of these regions are extracted. Next, the semantic based decision tree method is used to learn the concept of each region in an image and decide whether the image is relevant to the query or not.

In Web image search scenario, it is difficult to know how many relevant images there are in the database for a given query. Hence, the conventional precision and recall measurement is not used, instead, the bull's eye measurement is used. The bull's eye measures the retrieval precision among the top K retrieved images.

3.2 Experimental Results

We observed that among the top 10-20 images returned by Google, most of them are relevant to the query. However, many images in the rest of the list are irrelevant due to semantic misunderstanding. Using SIEVE, the performance of Web image retrieval is significantly improved by filtering out many of these irrelevant images. Fig. 2 compares the average performance of Google image search and SIEVE for the 10 different queries we selected. Overall, the retrieval precision of SIEVE is significantly higher than that of Google. For instance, Fig. 3 and Fig. 4 list the top 25 and 20 images using query 'tiger' and 'snow' respectively. It is shown that the retrieval results of SIEVE are more relevant to the queries than those of Google.

In the results given by Google image search, there are some irrelevant images whose surrounding text contains the keyword 'tiger' and are hence selected as relevant images. For example, the seventh image in Fig. 3 is described as 'tiger pear in tyre', and the description for the twelfth image for query 'snow' in Fig. 4 is: '*Snow*

Game publishers are starting to push the envelope further and further …' SIEVE removes many of such images from the list, whose visual content are obviously irrelevant to the query.

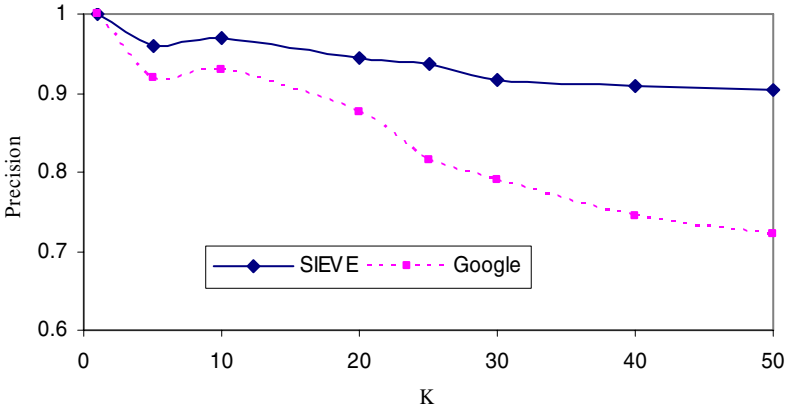


Fig. 2. Average Performance of SIEVE and Google Image Search for the 10 Queries

It is observed that in some circumstances, SIEVE removes some relevant images from the returned list. There are mainly two reasons for this. The first comes from the limitation in the decision tree as the training samples do not cover every case in the image database. Fig. 5 shows a few such images. For query 'tiger', the eighth image (a black-and-white cartoon tiger as shown in Fig. 5(a)) in the list is removed by SIEVE, as its color is very different from the training samples. Another sample is the thirty-seventh image for query 'snow', as shown in Fig. 5(b), whose color is different from most of the 'snow' images due to the dark light. Therefore it is also considered irrelevant by the decision tree.

In some cases, the decision tree may classify irrelevant images as relevant. For instance, the eleventh image for query 'snow' is a 'snow leopard', as shown in Fig. 5(c). Although this image is irrelevant to the query 'snow', it has been classified as 'snow' because a certain region in the image (the chest of the leopard) is visually similar to snow. Another example is given in Fig. 5(d), which is a movie ad, not a sunset image. SIEVE did not remove it from the list for query 'sunset', as it contains regions visually very similar to sunset and the decision tree cannot tell the difference.

For queries with clear semantic meaning, like 'sunset' and 'firework', Google can have high retrieval performance. However, the results are still improved by using SIEVE, especially when K (total number of returned images) is large. As examples, Fig. 6 and Fig. 7 list the top 20 images given by Google image search and SIEVE using query 'sunset' and 'firework', respectively.

For the query 'sunset', the fifth and twentieth images in the list are relevant to the query, they are removed by SIEVE. This is because 'sunset' regions are very small in these images, and are ignored by the system. In most cases, very small regions are not meaningful and ignoring such regions helps to find 'salient' regions which are more important. However, in some cases like the 'sunset' images mentioned above, small

regions may be semantically important. Other factors than size should be considered in future to deal with this type of situations. The seventeenth image in the list is about '*NASA - Sunset Planets. Venus, Jupiter and the Moon are gathering for a beautiful sunset sky show*', and it is irrelevant to the query. SIEVE successfully removes it from the list.

The eleventh image in Fig. 7 returned by Google does not contain a normal firework texture pattern. Thus, the decision tree does not 'recognize' it and mistakenly removes it from the list. A few other irrelevant images such as the fifteenth and eighteenth images are successfully removed from the list by SIEVE.

To summarize, the experimental results demonstrate that SIEVE effective improves the performance of Web image search by filtering out many irrelevant images from the image list returned by text-based search engine, although some relevant images may also be removed due to the limited dataset used to train the decision tree.

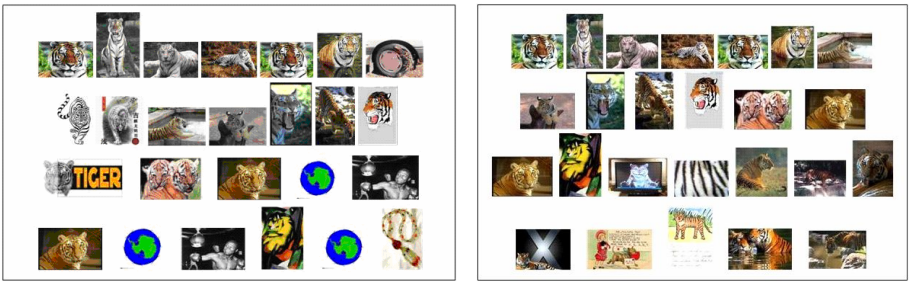


Fig. 3. Top 25 Images Returned by Google (left) and by SIEVE (right) using Query 'Tiger'

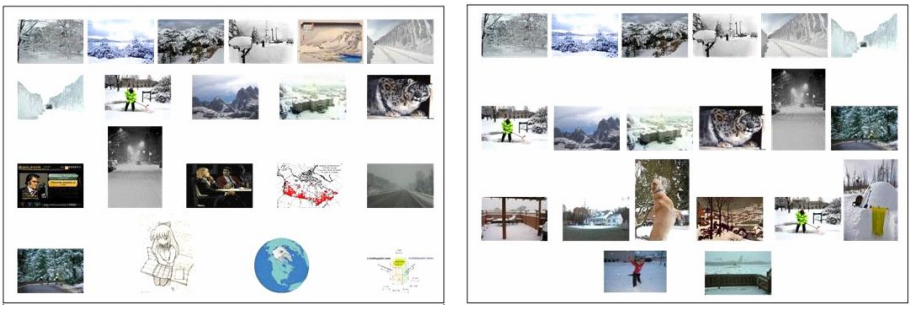


Fig. 4. Top 20 Images Returned by Google (left) and by SIEVE (right) using Query 'Snow'

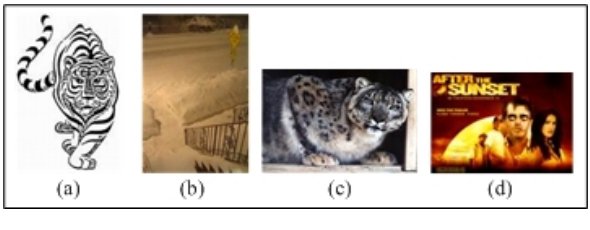


Fig. 5. Examples of Classification Errors

Fig. 6. Tope 20 Images Returned by Google (left) and SIEVE (right) using Query 'Sunset'

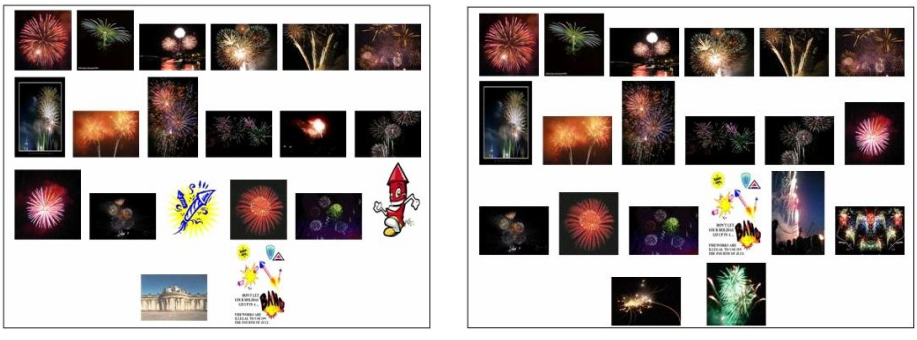


Fig. 7. Top 20 Images Returned by Google (left) and SIEVE (right) using Query 'Firework'

4 Discussions

In this section, we first suggest three approaches for integrating SIEVE with existing text-based Web image search engines. Next, we discuss how to reduce SIEVE response time for real-time Web image search by making use of both text-based index and semantic-based index. Finally, the scalability of SIEVE is discussed in the context of extending it for a wide range of user queries.

4.1 Integration of SIEVE with Text Based Image Search Engine

In the following, we propose different integration scenarios and take Google as an example of text based image search engine. In all these scenarios, Google and SIEVE are integrated in a hierarchy, with Google at the bottom layer to net the roughly similar images based on textual analysis. The results from Google are passed on to SIEVE for visual analysis.

Scenario 1 – SIEVE is integrated with Google on the server side. In this scenario, a user sends an image search query to Google using a standard Web browser. Google searches its image database using its own text-based image search algorithms. However, the resultant image list is not returned to the user immediately. Instead, the returned image list is input to SIEVE for visual analysis. SIEVE eliminates those semantically irrelevant images from the list and returns the final image list to user. In

this case, SIEVE is opaque to users, and the client side needs no modification. Users feel no difference from existing image searching experience, only to receive more accurate results due to the integration of SIEVE with Google.

Scenario 2 – SIEVE is integrated with the Web browser as a plug-in. A user query is directed by the SIEVE to Google. The results from Google are received by SIEVE but not shown to user immediately. SIEVE then performs semantic level filtering on the received image list and returns the filtered results to the user. While the server side remains unaltered in this case, the client side needs to install the plug-in in order to make use of SIEVE for semantic level filtering.

Scenario 3 – SIEVE is used as an application software. This SIEVE software directs user queries to various Web image search engines. The user has the option of selecting more than one search engines to search for images. SIEVE software takes user queries and uses Application Programmable Interfaces (APIs) provided by various search engines, such as Google API, to send user queries and obtain image search results. SIEVE receives text-based image search results from one or more search engines, filters out these images on semantic level and presents the results to the user. This approach has the advantage that users can search multiple search engines simultaneously. This minimizes the delay caused by searching through many search engines one after the other by the users themselves. Moreover, the chances of quickly finding the most relevant images from the Web are maximized. Another advantage is that the users are not dependent on Google or other Web search engines for integration. However, in this case, the users need to install the SIEVE on their machines, and they do not have the freedom to use it everywhere, contrary to the Web browser based integration solutions discussed in scenarios 1 and 2.

4.2 SIEVE Response Time

Refer to Fig. 1(b), to determine whether an image is irrelevant to a query or not, we need to first apply image segmentation to obtain image regions, and then extract low-level region features. With region features available, SIEVE then derives the high-level concept of each region according to the decision rules obtained from decision tree. As the decision rules are obtained off-line, deriving region concepts on-line can be done easily by testing the region against the decision rules. Hence, the main computation cost of SIEVE comes from image segmentation and low-level region feature extraction. Our experimental data show that, on an average, image segmentation and region feature extraction together takes about three seconds for an image, on a Pentium IV 3G MHz processor.

Considering that significant time is spent on learning the semantics of each image, a semantic index has to be built off-line for real-time Web image search. As proposed in Section 4.1, SIEVE and text-based search engine can be integrated in a hierarchy. In the long run, Google or any other text-based search engine has to make provision of storing this semantic index along with the text-based index in the database.

4.3 Scalability

The scalability of SIEVE highly depends on the performance of the image semantic learning method employed. In the current version, SIEVE is built on top of the

decision tree which is limited to natural scenery images. Although a limited concept set is used to test its performance, the decision tree can accommodate more semantic concepts, provided their corresponding distinct feature templates are available for inclusion in the training dataset. For concepts in different domains, we can generate different decision trees by using different types of attributes (image features) and defining different templates accordingly. For example, shape is not used in learning concepts for natural scenery images, but it may be important in other domains such as tools database including screwdrivers, pliers, wrench, etc. In addition, dominant color is used to describe region colors, as most natural scenery image regions have single dominant color. However, this may not be true for images in other domains such as fabric cloth for which color template including more than one color has to be defined.

To use SIEVE for general applications, one possible way is to classify user queries into different categories (domains) and learn decision rules for the concepts in each category. With a large number of user queries and Web images collected, queries can be classified by first clustering all Web images into different semantic categories (such as natural scenery, architectures and musical instruments), and then study the relationship between query keywords and image categories to map the keywords to each category.

5 Conclusions

In this paper, we have proposed to use SIEVE to improve text based Web image search. Specifically, SIEVE makes use of visual features which represent image semantic content, to eliminate irrelevant images from the image list returned by a text-based image search engine. The idea behind the technique is intuitive and simple, it is easy to be integrated with existing Web image search engines. Our experimental results show it is very promising. It has significant improvement over Google image search in terms of retrieval accuracy. Different methods of integrating SIEVE with existing text-based search engines in real time have also been proposed.

Compared with similar integration approaches in literature, the proposed SIEVE is linear to the size of the text-based result set, whereas existing approaches such as the re-ranking and grouping steps increase their complexity by a higher order.

Compared with conventional content based image retrieval systems, the proposed SIEVE is also much more efficient. It does not need to access the entire huge image database, instead, it lets the more efficient text based image search engine to net the relevant images from the image database and it then works on the small returned list of images by eliminating the irrelevant images from the netted list.

Currently, we have used only a limited concept set related to natural scenery images to test the performance of SIEVE. In the future, we will expand our concept sets to include a much larger number of categories, so as to further improve the retrieval accuracy. For concepts in different domains, we plan to use different types of features and define different semantic templates to represent the concepts. Multi-keywords query will also be considered in our future implementation.

References

[1] Thao, C., Munson, E.V.: A Relevance Model for Web Image Search. In: Proc. of International Workshop on Web Document Analysis (WDA 2003) UK, August 3, 2003 pp. 57–60 (2003)
[2] Lin, W.-H., Jin, R., Hauptmann, A.: Web Image Retrieval Re-Ranking with Relevance Model. In: Proc. of IEEE/WIC International Conference on Web Intelligence (WI'03), pp. 242–248 (2003)
[3] Google image search: http://images.google.com (accessed in December 2006.)
[4] Yahoo image search: http://images.yahoo.com (accessed in December 2006.)
[5] Cai, D., He, X., Li, Z., Ma, W.-Y., Wen, J.-R.: Hierarchical Clustering of WWW Image Search Results using Visual, Textual and Link Information. In: Proc. of ACM Inter. Conf. on Multimedia, pp. 952–959 (2004)
[6] He, J., Zhang, C., Zhao, N., Tong, H.: Boosting Web Image Search by Co-Ranking. In: Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 409–412 (2005)
[7] Liu, Y., Zhang, D.S., Lu, G., Ma, W.-Y.: Region-Based Image Retrieval with High-Level Semantic Color Names. In: Proc. of IEEE 11th International Multi-Media Modelling Conference (MMM05), January 12-14, 2005, Melbourne, Australia, pp. 180–187 (2005)
[8] Liu, Y., Zhang, D.S., Lu, G., Ma, W.-Y.: Region-based Image Retrieval with Perceptual Colors. In: Proc. of Pacific-Rim Multimedia Conference (PCM (December 2004), pp. 931–938 (2004)
[9] Liu, Y., Zhang, D.S., Lu, G., Ma, W.-Y.: Study on Texture Feature Extraction in Region-Based Image Retrieval System. In: Proc. of Inter. Multimedia Modelling Conf (MMM 2006), Beijing, pp. 264–271 (2006)
[10] Liu, Y., Ma, W., Zhang, D.S., Lu, G.: An Efficient Texture Feature Extraction Algorithm for Arbitrary-Shaped Regions. In: Proc. of IEEE 7th International Conference on Signal Processing (ICSP2004), August 31-September 4, 2004, Beijing, China vol. 2, pp. 1037–1040. (2004)
[11] Liu, Y., Zhang, D.S., Lu, G.: Deriving High-Level Concepts Using Fuzzy-ID3 Decision Tree for Image Retrieval. In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP05), March 18-23, 2005 Philadelphia, PA, USA pp. 501–504. (2005)

Virtual Community Based Secure Service Discovery and Access for 3D Video Steaming Applications

Shudong Chen, Igor Radovanovic, Johan Lukkien, Richard Verhoeven, Melissa Tjiong, and Remi Bosman

Department of Mathematics and Computer Science
Eindhoven University of Technology, the Netherlands
{shudong.chen, i.radovanovic, j.j.lukkien, p.h.f.m.verhoeven, m.tjiong, r.p.bosman}@tue.nl

Abstract. The Freeband I-Share project aims to define the mechanisms for trust, willingness, resource discovery and sharing mechanisms in virtual communities. To improve the secure and performance of a 3D video streaming application, which is a research vehicle of the I-Share project, we propose a virtual community based access control approach for secure service discovery and access (VICSDA) which groups services in virtual communities and only grants authenticated community members to discover and access these community services. There are two main contributions associated with this approach. First, different from most of the other access control approaches it adopts a dual access control mechanism which allows community services to define their local access control policy besides following the community membership policy. Second, behavior of these community services is monitored in order to guarantee a better QoS provision. Using this approach, the 3D video streaming application can be guaranteed with authentication and message confidentiality through the dual secure service discovery and access mechanism. Better application performance can also be achieved through the community member behavior audit.

Keywords: 3D video streaming, virtual community, access control.

1 Introduction

The Freeband I-Share project is a joint research project of various Dutch public universities and companies. I-Share aims to define the mechanisms for trust, willingness, resource discovery and sharing mechanisms in virtual communities, which are dynamic groups of devices and services that are willing to collaborate for the better of the whole [1]. Emphasis is on protocols for resource/service discovery and for efficient data/file/video distribution over heterogeneous networks and devices. In particular, mechanisms for distributed and layered content processing of the multi-view 3D video utilizing resources shared in virtual communities are considered as a research vehicle. The aim for this paper is to extend an existing application [2] with techniques and methods that enable establishment of confidence in secure operation of a distributed system built out of cooperating services provided by multiple parties.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 391–397, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Paper [2] presents a layered framework for 3D-TV transmission, combining multi-view and depth-based approaches in a scalable fashion. Besides texture and depth information, specific layers are added for coded occlusion data and edge-mask information to allow high-quality 3D rendering of key objects in the scene. By relying on the concept of resource sharing for the creation and streaming of virtual viewpoints in a network overlay, the range of view-points selectable by the user (FTV) is extended. An experimental 3D video transmission application running on this layered framework was implemented as a standalone program, running on a powerful platform. Both the sender and the receiver were implemented on a desktop PC running Linux OS. A problem with this is a lack of flexibility and security. It's hard to adapt the system when there are some configuration changes, for example, another type of receiver arrives. And also because there is no access control involved, the system can easily be attacked, for example, by a malicious user pretends to be a legal receiver to steal the content of the 3D video.

The rapidly increasing technology in Service Oriented Architecture (SOA) [3-6] enables the flexibility of software. In SOA, software components are wrapped into network-exposed services with explicitly described interfaces to provide functionalities; applications can be built by interconnecting services, leaving the binding until runtime. Relying on standards SOA provides a highly flexible and adaptable implementation for services and applications; eventually, it becomes possible to switch from a particular service to a different one without adaptations. Thus, to achieve high flexibility of the 3D video streaming application, it is necessary to borrow the concept of SOA and wrap the existing components into services. In this way it is much easier to develop and adapt the components independently.

However some issues still remain in current SOA systems. There are three activities in a SOA system: publish, discover and invoke. A service provider defines a service description and publishes it to a service registry which acts as an intermediary between providers and user where services are stored. A service user uses the service registry's search capability to discover service descriptions and their respective providers. Later the user will invoke service interfaces for required functionalities. Services published in a service registry [7] are supposed to be discoverable by all service users. This may lead to a privacy problem when a service provider only wants to reveal its services to trusted service users. In addition, service providers and service users should have some guarantee that the agreements in the contract, for example, promised functionalities and QoS in the service description, are supported. That is difficult when a SOA lacks monitoring and enforcement authority.

In order to achieve a flexible and secure 3D video streaming application, we propose a virtual community based access control approach for service discovery and access namely VICSDA. In this approach, we extend SOA with the concept of virtual communities to help protect the privacy of service providers and include incentive for high QoS provision [1]. Services are grouped into virtual communities. Only authenticated community members can be allowed to discover and access these services. There is a large volume of publications on access control and some of them are on secure service discovery and access [8, 9]. Compared with other approaches, our proposed approach contains two main contributions: (i) a dual access control mechanism which is that community services are autonomous to define their local access control policy; (ii) an audit functionality of the virtual community maintenance

mechanism which is designed to guarantee better QoS provision to applications through monitoring services' behavior.

The remainder of this paper is organized as follows. In section 2, we elaborate the virtual community based secure service discovery and access approach. Then we present the 3D video streaming application as an example to validate the feasibility of the proposed approach. Conclusion is drawn in Section 4.

2 Virtual Community Based Secure Service Discovery and Access

A virtual community is a dynamic contract-based aggregation whose members have commonalities and interact via shared services by means of a digital network for example the Internet. A virtual community has rules that each member has to follow. It provides services to members and has the potential to develop different applications through service cooperation and external orchestration.

The advantage of sharing functionalities and resources in type of services with other community members is that limitations of individual devices can be overcome by collaboration with others. The sharing not only benefits the individual participating devices, but also the system as a whole since functionalities, resources and communication of software can be optimally distributed. We are interested in investigating the SOA approach for composing distributed applications from cooperating services. In this process, service users do not necessarily take part in this service cooperation; it is referred to an external orchestration as an orchestrator.

Services are grouped in virtual communities through adding additional functionalities to them, for example can only be exposed to authenticated community members. Then a virtual community is an overlay network for the existing services. It provides at least the following functionalities to its members: (i) register a service provider or a service user as a member and deregistering a member; (ii) publish/discover/invoke required community services; (iii) protect the privacy of service providers and provide secure service access; (iv) maintain the member and service list. Rich functionalities of a virtual community can be making trustable recommendations for a service user during the service selection stage, and more.

Based on this virtual community overlay, we designed the dual access control approach. First, we use the member boundary of a virtual community to control the service discovery scope and filter the un-trusted or malicious access requests to services. Second, in our opinion services are autonomous entities and it is natural to allow services to define their local access control policy and to deny access requests coming from their un-trusted users. However, services' autonomy may cause the entire application fails because required service providers may stop providing functionalities. To prevent this, we grant each community service a credit value to represent its behavior. A service user can evaluate a service's QoS. This evaluation affects a service's credit value. This value may increase because of a positive evaluation and may decrease when a negative one arrives. When it is below a threshold it will be deregistered from the community. Using this maintenance mechanism, service autonomy can be restricted and then guarantee a better QoS provision to applications. Using the UML notation [10], Figure 1 depicts the essential ingredients that together compose a virtual community.

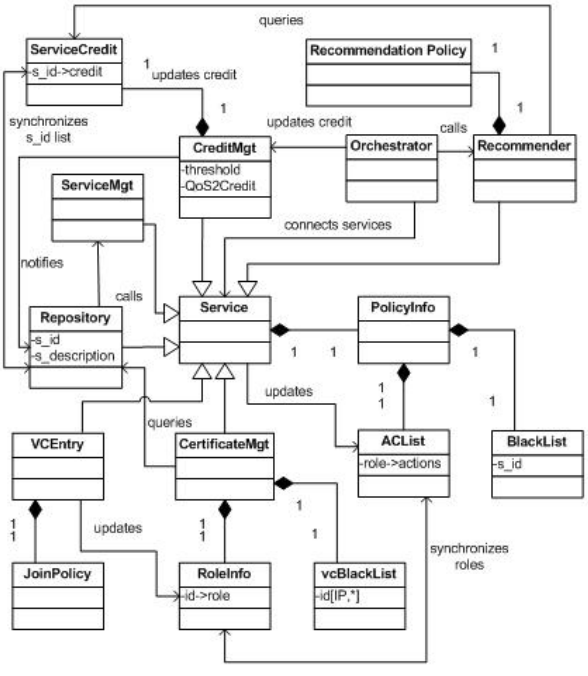


Fig. 1. Ingredients of a virtual community

VCEntry and *CertificateMgt* are accessible to any parties including users on the network and internal community members. While others, for example, the *Repository* and the *CreditMgt* are only available for community members. *VCEntry* is a service to deal with members' actions inside a community including member registration and deregistration. When one meets the community joining policy in *JoinPolicy*, it will be authorized to be a community member. Then specific roles are assigned to it by the *VCEntry*. It could be a service user to discover and access services, or a service provider to publish services, or a combination of available roles in *RoleInfo*. When a member publishes services at a virtual community *Repository,* the *ServiceMgt* will be called by the *Repository*. *ServiceMgt* can wrap a service as a community service by adding virtual community properties to it including *ACList*, *BlackList* and *Credit.* A service can specify its local access control policy by assigning capabilities to different roles in its *ACList* and by adding its un-trusted users to the *BlackList*.

During the execution of an application, an orchestrator will first discover required services and then orchestrates these services cooperate with each other. The virtual community based dual access control approach will react on each transaction. An example is when the *Orchestrator* submits the service discovery request to the *Repository*, a valid ticket which is distributed by the *CertificateMgt* will be checked by the *Repository*. This ticket is an authentication that the *Orchestrator* is a community member. The *Repository* will first verify the authenticity of the signature and the validity of the ticket. If approved, the *Repository* will execute a second round of access control based on its local policy. In this stage, the *Repository* will query its

BlackList to check whether the *Orchestrator* should be trusted. If it trusts the *Orchestrator*, the *Repository* looks up its *ACList* to grant the *Orchestrator* with corresponding authorization according to the roles indicated in the ticket. For example, the *Orchestrator* can invoke the service discovery interface but not the service list update. Only after the *Orchestrator* passes the dual access control, its service discovery request can be executed by the *Repository*.

CreditMgt is designed to facilitate the member list, community services list maintenance and the better QoS guarantee for service cooperation. It periodically compares a service's credit value with a threshold. When a credit is smaller than the threshold, the *CreditMgt* will inform the *Repository* to deregister that service.

3 Service-Based 3D Video Streaming Application

In order to validate the effect of improving the feasibility and security of the 3D-TV transmission framework through adopting this virtual community based access control approach, in this section, we describe an experimental application. This application demonstrates a 3D video generated by two cameras shown on a terminal from different viewpoints in an interactive fashion. A user either selects an arbitrary viewpoint and a viewing direction, or the user's movements are continuously tracked and the displayed content is automatically adjusted to the view points.

In particular, we integrate the proposed dual access control approach with an early version of this application and pervious software components are wrapped as community services. In the new application, the two cameras generate two video streams of an object from different viewpoints and stream them to the 3D video generation service. Then the 3D video generation service calculates the depth map files of these two videos that will be used to generate a 3D video. At the mean time, the 3D video generation service is bound to a mouse service. The mouse's event indicates a user's viewpoints switch requirements. A user can control that mouse service to express her expected viewpoints of the 3D video. These viewpoints are received by the 3D video generation service. During the video streaming the 3D video generation service continuously tracks a user's viewpoints and correspondingly adjusts the generated 3D video's content. The mouse service may work on another device for example a PDA. The working principle of the 3D video generation service and the video receiver is shown in Figure 2 [2].

All needed software components are visible as services on the network. The early 3D video generation program which is written in C++. We have modified it and wrapped it into a web service [11]. We have also developed two mouse services. Both of them are running as web services with one running on the PC and the other running on the PDA. During the 3D video streaming, particularly, the coding of the displayed content is automatically adapted to the control information coming from an external service: the mouse of the PC or the touch-screen of the PDA.

The architecture of the 3D video streaming application is depicted in Figure 3. We wrap all these services into virtual community services. So their availability is restricted to users with community credentials. We use an orchestrator to execute the 3D video streaming application. It first discovers all required services from a virtual

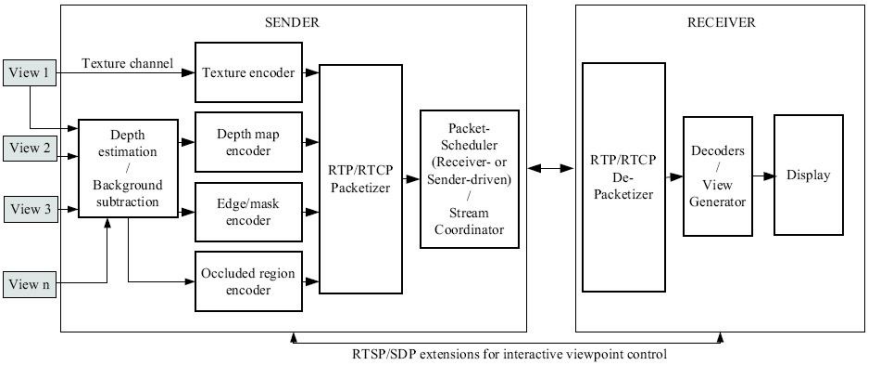


Fig. 2. Constituents of the 3D video generation service and the video receiver service

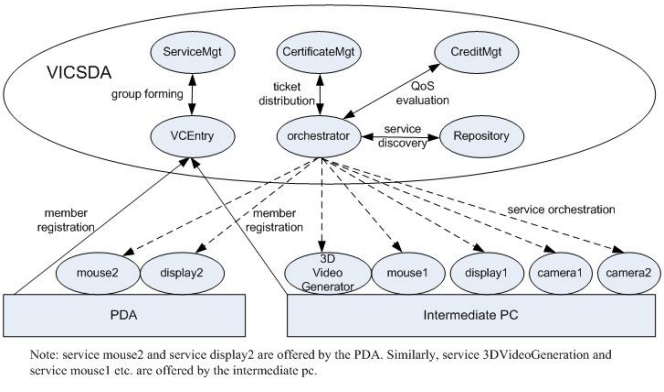


Fig. 3. Architecture of the 3D video streaming application

community repository. Then it configures and binds these services together to achieve the application. For example, it binds the 3DVideoGenerator service with one mouse service and one display service. During the course of service discovery, binding and access, the dual access control approach, which includes the ticket validation and the capability granting, is applied between every two communication parties. For example, the 3DVideoGenerator service and the mouse service have to negotiate before they are bound together.

4 Conclusions

This paper describes the virtual community based access control approach for secure service discovery and access, which is aimed to solve the flexibility and security problems of the 3D video streaming application by wrapping existing software into community services. This proposed approach contains two novel features: (i) a dual access control mechanism which allows autonomous services to define their local

access control policy besides following the community membership policy; (ii) a member behavior audit functionality which aims at better QoS provision to applications. This approach can also be applied to other service-based ubiquitous computing systems. There are immediate challenges facing us: (i) how to reduce the authentication overhead to improve the performance of the system? (ii) how to recover from the overload or fail of the *CertificateMgt* service without affecting the system's performance? Our ongoing research is aimed to solve those problems.

Acknowledgments. We thank Goran Petrovic and Peter H. N. de With for their technical feedback at the implementation stage of this work. We also thank our anonymous reviewers for their invaluable feedback which helped in improving the quality of the paper.

References

1. Freeband I-Share D1.5. Project Deliverable. http://www.win.tue.nl/san/publications/
2. Petrovic, G., de With, P.H.N.: Framework for Layered 3D Video Streaming. In: 27th Symposium on Information Theory. Noordwijk, The Netherlands (2006)
3. SOA. http://www.service-architecture.com/index.html
4. Kreger, H.: Web Services Conceptual Architecture (WSCA 1.0) http://www-306.ibm.com/software/solutions/webservices/pdf/WSCA.pdf
5. UPnP Device Architecture. Version 1.0. http://www.upnp.org/
6. Sun Microsystems. Jini technology architectural overview. White Paper http://www.sun.com/jini/whitepapers/architecture.html
7. UDDI Version 3.0. Published Specification. http://www.uddi.org/
8. Czerwinski, S., Zhao, B.Y., Hodes, T., Joseph, A., Katz, R.: An Architecture for Secure Service Discovery Service. In: 5th International Conference on Mobile Computing and Networks (MobiCom'99), Seattle, WA, pp. 24–35 (1999)
9. Pearlman, L., Welch, V., Foster, I., Kesselman, C., Tuecke, S.: A Community Authorization Service for Group Collaboration. In: 3rd International Workshop on Policies for Distributed Systems and Networks, Monterey, CA, USA (2002)
10. Fowler, M.: UML Distilled: A Brief Guide to the Standard Object Modeling Language, Version 3.0. Addison-Wesley (ISBN 0-321-19368-7)
11. Cerami, E.: Web Services Essentials: Distributed Applications with XML-RPC, SOAP, UDDI & WSDL, 5th edn. p. 304. Publisher: O'Reilly (February 2002) ISBN: 0-596-00224-6

QoS Adaptive Data Organizing and Delivery Framework for P2P Media Streaming

Longshe Huo[1], Wen Gao[1,2], Qiang Fu[1], Rui Guo[2], and Shuangjia Chen[2]

[1] Institute of Digital Media, Peking University, 100871 Beijing, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, 100080 Beijing, China
{lshuo, wgao}@pku.edu.cn, {rguo, qfu, sjchen}@idm.pku.edu.cn
http://idm.pku.edu.cn

Abstract. This paper presents a novel data organizing and delivery framework for P2P media streaming which is aware of media content. In this framework, the media data is partitioned GOP by GOP, and reorganized according to the frame-priorities within a GOP. All GOPs cached in the media buffer of a peer node can be scheduled concurrently, while the frames contained in each GOP are transmitted by order strictly according to their priorities. Analysis and primary experimental results show that the proposed techniques can provide not only self-adaptive QoS for heterogeneous network conditions, but also quick channel switching with near fixed delay time.

Keywords: P2P, media streaming, data organization, data delivery, QoS.

1 Introduction

Recently, Peer-to-Peer (P2P) has been commonly considered as the most promising technique for deploying large-scale media streaming systems over the Internet. Compared with traditional client-server based systems, such as content delivery network (CDN), P2P based system has the advantages of requiring no dedicated infrastructure and being able to self-scale as the resources of the network increase with the number of users.

From the view of data-flow topology, P2P streaming systems can be classified into three categories approximately: tree topology ([1], [2], etc.), forest (multi-tree) topology ([3], [4], etc.), and mesh topology ([5], [6], [7], etc.), among which mesh is more efficient than others though it is also more complicated to implement. With the mesh topology, at the same time a peer can select and receive media data from a set of different senders, each contributing a portion of the streaming bandwidth. The main challenges of mesh topology are how to select the appropriate set of senders and how to schedule the multi-source data transmission. Inspired by the article by Zhang et al. [7], which presented a data-driven overlay network (DONet) and a gossip-based topology maintaining algorithm, recent years mesh-style P2P streaming systems have been widely proposed and developed. For example, merely in China nowadays there are about dozens of P2P live streaming applications being deployed on the Internet [8], such as PPlive, PPstream, Roxbeam, etc.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 398–407, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Though some successes have been achieved, there are still challenging problems need to be overcome in P2P media streaming. In our opinion, the biggest problem results from the fact that most of the current work on P2P streaming systems focuses on network topology and protocol design, but pay rare attention to the media contents carried over the network. In fact, since streaming media have their distinct characteristics from normal data file, good performance can be achieved only when both the characteristics of media coding and networking are considered together perfectly.

In the aspect of data organization, current systems simply partition a streaming media, which is composed of audio and video, into a series of data blocks, and then schedule and transmit data with block as the minimal unit. First, since such a partition does not consider the boundary of audio and video frames, block-losses may lead to incorrect or failing decoding of frames which cross over block boundaries. Second, since there is no awareness of block contents, selective and priority-based dropping strategies cannot be used by schedulers to transmit important media data first in case of bandwidth shortage. Third, since audio and video data are multiplexed together, block-losses may result in both blacked screen display and mute voice for a while at the client side. However, because person is more sensitive to audio than that of video, discontinuous voice can be annoying and undesirable to human ears.

In the aspect of data delivery, most of the current systems try to provide the same and perfect QoS for all participated peers. However, due to the heterogeneity and the best effort property of the Internet, this wish is hard to be satisfied under current situations. For example, if the access bandwidth of a peer is less than the average bit rate of the media channel it required, or the sum of the uplink bandwidths of all upstream peers who provide data for this peer is less than the average bit rate, then despite UDP or TCP is used, random packet-losses may occur ether during the network or at the buffer of upstream peers. The random packet-losses may lead to incorrect decoding at the client side even partial data has been received, which means not only the waste of bandwidth resources, but also degraded media reconstruction qualities.

To address these problems, in this paper we present a novel data organizing and delivery framework for P2P media streaming system, which takes advantage of the coding characteristics of compressed media content, and can adaptively provide appropriate QoS for each peer according to its current network conditions. The proposed techniques can guarantee that every data unit received by a peer at any time is intact and decodable independent of any still un-received media data. By this way, no waste of bandwidth is involved and free-error and fluent media experience can be obtained even in the case of worse network conditions

2 Proposed Techniques

In this section, we first present the overall architecture of the P2P media streaming system, then describe in detail the proposed data organizing and delivery mechanism for this system respectively, and finally analyze the advantages of the proposed techniques.

2.1 System Architecture

As shown in Fig. 1, our proposed P2P media streaming system consists of four types of node: Streaming Server, Seed Peer, Normal Peer, and Media Player. A collection of Seed Peers and Normal Peers constitute a P2P overlay network.

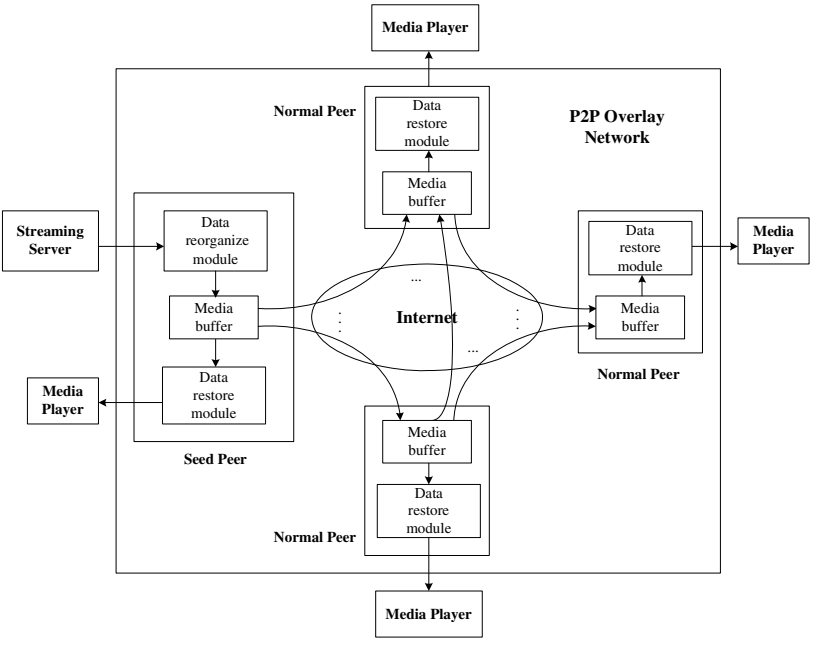


Fig. 1. Proposed architecture for P2P media streaming system

The Streaming Server serves as the media source of the whole streaming system. It is connected with Seed Peer, and provides media data for the latter. The data format and the protocols used to transfer data from the Streaming Server to the Seed Peer are the same as used in general client-server streaming systems.

All the media data flow from the Streaming Server cannot be transmitted to Normal Peers directly, but must be redistributed by a Seed Peer. Thus the Seed Peer serves as a broadcaster for the P2P overlay network. It first receives data from the Streaming Server, then reorganizes them and put into media buffer according to a specific data format which can be exchanged between different peers, and finally retransfers the buffered data to other peers which have direct connections with it. Furthermore, the Seed Peer can also acts as a Normal Peer which connects directly with a Media Player.

Each Normal Peer is a user who participates in the P2P network. On the one hand, a Normal Peer can establish connections with one or more other Normal / Seed Peers via the Internet. It first receives data from some of them and puts into its media buffer, and then relays the buffered data to others. On the other hand, each Normal

Peer also has a direct connection with a Media Player. It first restores the buffered media data back into their original format as they flow from the Streaming Server, and then sends them to the Media Player.

Media Player is the rearmost consumer of media data carried over the P2P network. Generally a Media Player is located on the same terminal which runs the Normal / Seed Peer program. The data format and the protocols used to exchange media data between peer program and Media Player are the same as used in general client-server streaming systems.

2.2 Data Organization

From above descriptions it can be seen that when the media data flow from the Streaming Server into the P2P overlay network, they should be first transformed into specific format by the data reorganize module and then cached in the media buffer in the Seed Peer; when they flow from the P2P overlay network out to a Media Player, the data restore module in a Normal / Seed Peer should be called to restore them back to their original media format; when they flow between any two Normal /Seed Peers, a common data structure defined for media buffer should be obeyed. Following we first give the proposed media buffer data structure, and then describe the processes of data reorganizing and restoring in detail.

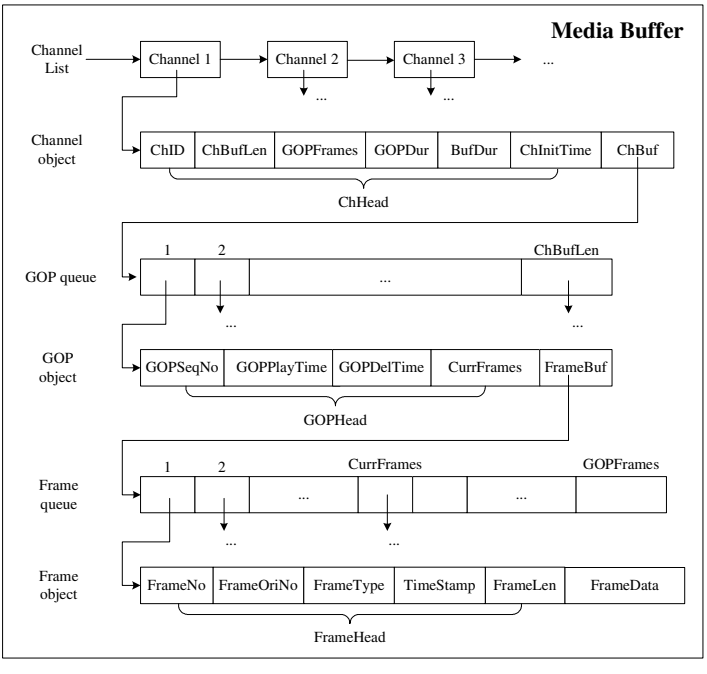


Fig. 2. Data structure of media buffer in Normal /Seed Peers

2.2.1 Media Buffer Data Structure

Based on the data dependency characteristics of video coding, within a P2P overlay network, we partition media data according to the boundaries of video GOP (group of pictures) and frame, and regard a video frame as the minimal data unit for transmitting. As shown in Fig. 2, the media buffer in each Normal / Seed Peer consists of a list of channel pointers, each pointing to a channel object. Each channel is composed of a channel head (*ChHead*) and a channel buffer (*ChBuf*), where *ChHead* includes at least the following fields:

- *ChID*: channel ID, solely identify a channel in the whole system;
- *ChBufLen*: the max number of GOPs cached in the channel buffer;
- *GOPFrames*: the original number of frames contained in a GOP of current channel;
- *GOPDur*: the preset playing duration of a GOP of current channel;
- *BufDur*: the preset playing duration of all GOPs cached in the channel buffer, it is also the time interval that a GOP should stay in the channel buffer;
- *ChInitTime*: the creation time of current channel in this Peer.

The *ChBuf* points to a FIFO (first-in-first-out) GOP queue, in which each element points to a GOP object, and each GOP object is composed of a GOP head (*GOPHead*) and a frame buffer (*FrameBuf*). The *GOPHead* contains at least the following fileds:

- *GOPSeqNo*: each GOP is assigned a sequence number;
- *GOPPlayTime*: the expected play time of this GOP in current peer;
- *GOPDelTime*: the expected time this GOP should be deleted from *ChBuf*;
- *CurrFrames*: the number of frames currently contained in *FrameBuf*.

The *FrameBuf* points to a frame queue with a size of *GOPFrames*. Each element in the queue points to a frame object. Each frame object is composed of two parts: *FrameHead* and *FrameData*, where *FrameHead* includes at least the following fields:

- *FrameNo*: each frame in a GOP is assigned a *FrameNo*, which increases inverse proportional to the priority of the frame, i.e., the frame with highest priority has a *FrameNo* of 1, and the frame with lowest priority has a *FrameNo* of *GOPFrames*.
- *FrameOriNo*: each frame in a GOP is also assigned a *FramOriNo*, which increases according to the decoding order before reorganization.
- *FrameType*: the coding type of current frame, such as I-, P-, or B-frame.
- *TimeStamp*: the presentation time stamp of current frame.
- *FrameLen*: the length of *FrameData*.

The *FrameData* contains the original media data of an intact frame.

All frame objects contained in a frame queue are ordered by their *FrameNo*. A frame object is allowed to be stayed in the frame queue only if all other frames with smaller *FrameNo* have already been stayed in the queue.

2.2.2 Data Reorganizing

The data reorganize module in the Seed Peer is responsible for receiving data from the Streaming Server and then reorganizing and putting them into the media buffer

GOP by GOP. The most important step of data reorganizing is to reorder all the frames within a GOP according to their priorities.

If a channel includes both audio and video data, in order to provide time-synchronized transmission of audio and video, and ensure the first-transmit right of audio data in case of bandwidth shortage, while reorganizing we encapsulate all the audio data corresponding with a video GOP in time into a virtual audio frame (called A-frame), and regard it as a normal video frame in the same GOP.

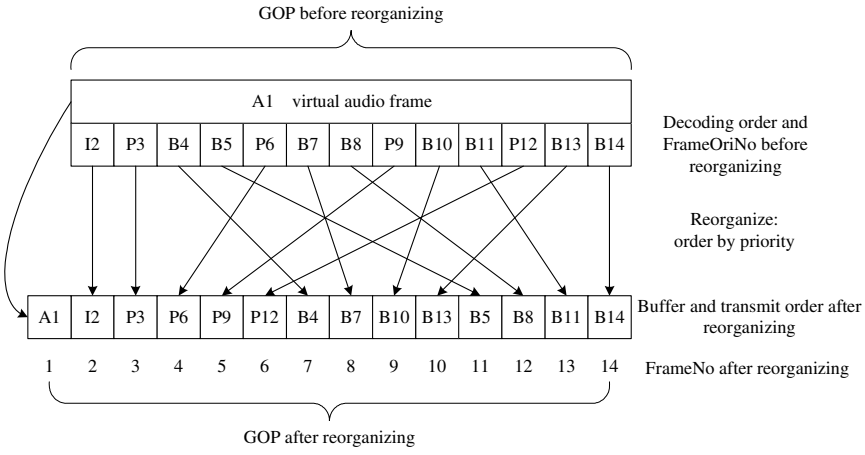


Fig. 3. Data reorganizing: reorder frames within a GOP by their priorities

As shown in Fig. 3, while reordering the frames within a GOP by priority, if there is no A-frame, then the I-frame has the highest priority, next followed by all P-frames ordered by their decoding order, next all the first B-frames directly after each P-frame ordered by their decoding order, and next again all the second B-frames directly after each P-frame ordered by their decoding order, and so on; otherwise if there is A-frame, then the A-frame has the highest priority, then followed by other video frames ordered according to the above rules.

2.2.3 Data Restoring

The data restore module in Normal / Seed Peer is responsible for restoring the data cached in the media buffer back into their original format as they flow from the Streaming Server, and then sending them to Media Player. The most important step of data restoring is to reorder all the frames within a GOP according to their original decoding order. This can be easily done by sorting all the frame objects in the frame queue with their *FrameOriNo* field in ascendant order.

2.3 Data Delivery Mechanism

Based on the above data organization mechanism, when a peer participates in the P2P overlay network and requests to play a channel already broadcasted by a Seed Peer, it should follow the following data delivery steps.

Step 1. Create a new channel object and its GOP queue in the media buffer of this Normal Peer, set the value of *ChInitTime* as the current system time, and get the values of *ChBufLen*, *GOPFrames*, *GOPDur*, and *BufDur* by communicating with other Normal / Seed Peers;

Step 2. Create a new GOP object and put it into the GOP queue, set the value of *GOPPlayTime* as the sum of *ChInitTime* and *StartDelay*, where *StartDelay* is a preset parameter of start delay time for Media Player, and set the value of *GOPDelTime* as the sum of *ChInitTime* and *BufDur*, set the value of *CurrFrames* as 0, and finally determine the value of *GOPSeqNo* by communicating with other Normal / Seed Peers;

Step 3. If the GOP queue is not full, then go to Step 4; otherwise go to Step 5;

Step 4. Create a new GOP object and put it into the GOP queue, the value of *GOPSeqNo* is increased by 1 based on the same field of the previous GOP, the values of *GOPPlayTime* and *GOPDelTime* are increased by *GOPDur* based on the same fields of the previous GOP, go to Step 3;

Step 5. By calling a multi-source P2P data scheduling algorithm, simultaneously select and receive frame object data for one or more GOP objects, whose values of *GOPPlayTime* are still greater than current system time, from the corresponding GOP queues and frame queues with the same channel located in one or more other Normal / Seed peers, and then put the received frame objects into their corresponding frame queues in local media buffer;

Step 6. At the same time, for each GOP object cached in the GOP queue, judge whether the current system time is greater than the value of *GOPPlayTime*. If the answer is yes, then stop receiving new frames for this GOP from other peers, and call the data restore module to restore all the frame objects cached in this GOP's frame queue back into their original format and send to the Media Player for decoding and rendering.

Step 7. If the current system time is greater than the value of *GOPDelTime* of the first GOP object in the GOP queue, then remove this GOP object from the GOP queue and delete it. Go to Step 3.

In Step 5 of the above process, while a peer receives media data from other Normal / Seed Peers, the following rules must be obeyed:

(1) Frame (include video frame and virtual audio frame) is the minimal data unit for data exchange between any peers;

(2) Before receiving a frame object for a GOP, it must be assured that all other frame objects of this GOP with smaller *FrameNo* have already been received and put into the GOP's frame queue intactly.

(3) Once a frame is selected to be received, all data contained in it must be transmitted to the receiver reliably and intactly. That is, no partial delivery of a frame is permitted. A frame object is either successfully delivered to the receiver, or wholly discarded at the sender. To fulfill this requirement, reliable transmission protocols such as TCP are suggested.

(4) Before putting a received frame object into its corresponding frame queue, it must be assured that all other frame objects within the same GOP with smaller *FrameNo* have already cached in the frame queue intactly.

As an example, let's assume that in a channel each GOP contains 8 frames (decoding order: A1, I2, P3, B4, B5, P6, B7, and B8), and the channel buffer length (*ChBufLen*) is 10, then Fig. 4 demonstrates a snapshot of the channel's GOP queue in the media buffer. At this moment, only the frames marked with symbol 'o' can be scheduled and received from other peers.

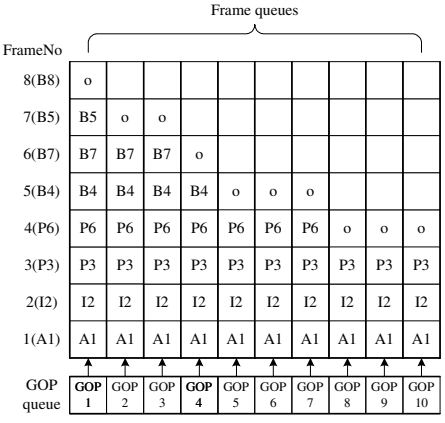


Fig. 4. Snapshot of a GOP queue in the media buffer

2.4 Analysis

The benefits of the proposed data organizing and delivery framework for P2P media streaming system can be concluded as follows.

First, since the media data is partitioned and organized according to the boundary of video GOP, and further all the frames within a GOP are reordered according to their priorities, therefore when the network bandwidth becomes insufficient, we can selectively transmit only part of the frames with higher priority, and discard others with lower priority. By this way, temporal scalable media delivery can be achieved, and QoS adaptive media streaming services can be provided for various users according to the dynamic fluctuating network conditions.

Second, since we encapsulate all the audio data corresponding with a video GOP in time into a virtual audio frame, and assign it the highest priority over all video frames in this GOP, therefore it can be guaranteed that the audio data is transmitted first before video in case of bandwidth shortage, and people can listen uninterrupted voice and music. At the same time, this technique can also achieve a good time synchronization effect between audio and video.

Third, since we regard an individual media frame as the minimal unit of buffering and transmitting, and transmit each frame within a GOP strictly according to the order of its priority using a reliable transmission protocol, therefore it can be assured that a frame is either transmitted to the receiver successfully or wholly discarded at the sender side, and at anytime the decoding of each frame which has been received by a peer should never depend on the decoding of other frames which have not been received. As a result, decoding errors caused by incomplete frames or inter-frame

dependency would never happen, and the network bandwidth resources can also be utilized sufficiently.

Fourth, in each peer, when the preset play time for a GOP is reached, the already received frames in this GOP are sent to the Media Player for decoding and playing immediately, despite whether all frames belonging to this GOP have been totally received. As s result, fast channel switching with near fixed delay can be achieved.

Finally, the proposed framework is suitable to most of the current market-prevailing media coding standards, such as AVS and MPEG series, and at the same time needs no modifications to the current used streaming servers and media players, therefore it has strong adaptability and is easy to be developed and deployed in large-scale P2P media streaming systems.

3 Experimental Results

To verify the effect and performance of the proposed framework, intelligent P2P topology construction and data scheduling algorithms based on this framework should be developed. Absent of them, in this section we only test the QoS adaptive performance in a simplified system, which has only a Seed Peer and a Normal Peer, and the data scheduler in the Normal Peer obeys a simple rule: first receive all frames with *FrameNo* of 1 for each GOP in the GOP queue, and then the frames with *FrameNo* of 2, and so on. The value of *ChBufLen* is set as 10.

We choose a segment of the movie "Inspector Gadget" with standard definition as the test video, and compress it using the AVS video codec. The average bit rate is 760kb/s, and the length of every GOP is 31 frames. For changed bandwidth between

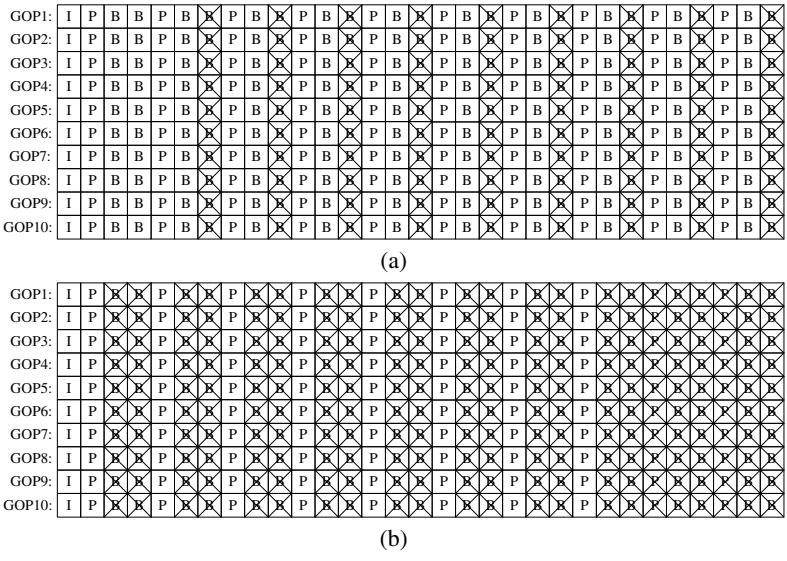


Fig. 5. Successfully received frames observed at the Media Player, the crossed symbols mean the dropped frames. (a) bandwidth = 700kb/s; (b) bandwidth = 400kb/s.

the Normal Peer and the Seed Peer, we record at the Media Player the types of all the successfully received video frames within 10 consecutive GOPs. As shown in Fig. 5, when the bandwidth is 700kb/s, in each GOP the second B-frame follows each P-frame (except the first one) is discarded; when the bandwidth decreases to 400kb/s, all B-frames and the last two P-frames in each GOP are discarded. In these situations, fluent video rendering effects can still be observed at the client side. The results are accordant to our anticipation.

4 Conclusion

In this paper, we present a novel data organizing and delivery framework for P2P media streaming, which takes into account both the characteristics of media coding and P2P networking, and can provide self-adaptive QoS for different users of the Internet. In the future, intelligent P2P topology construction and layered multi-source data scheduling algorithms which are based on this new framework and can take full advantages of this framework will be studied, and more extensive experiments both in simulative and practical operational environments should be conducted.

Acknowledgments. This work was partially supported by the National Key Technology R&D Program under grant No. 2006BAH02A10, the China Postdoctoral Science Foundation, and the“985” construction project of Peking University.

References

1. Deshpande, H., Bawa, M., Garcia-Molina, H.: Streaming Live Media over a Peer-to-Peer Network. Stanford database group technical report (2001-20) (August 2001)
2. Tran, D.A., Hua, K.A., Do, T.T.: A Peer-to-Peer Architecture for Media Streaming. IEEE Journal on Selected Areas in Communications 1, 1–13 (2004)
3. Chu, Y.-H., Rao, S.G., Zhang, H.: A Case for End System Multicast. ACM SIGMETRICS Performance Evaluation Review 1, 1–12 (2000)
4. Castro, M., Druschel, P., Kermarrec, A.-M., Nandi, A., Rowstron, A., Singh, A.: SplitStream: High-bandwidth content distribution in a cooperative environment. In: Proc. the International Workshop on Peer-to-Peer Systems, February 2003, Berkeley, CA (2003)
5. Heffeeda, M., Habib, A., Botev, B., Xu, D., Bhargava, B.: PROMISE: Peer-to-Peer Media Streaming Using CollectCast. In: Proc. ACM Multimedia, November 2003, Berkeley, CA (2003)
6. Jiang, X., Dong, Y., Xu, D., Bhargava, B.: GnuStream: A P2P media streaming prototype. In: Proc. IEEE International Conference on Multimedia and Expo, July 2003, Baltimore, USA (2003)
7. Zhang, X., Liu, J., Li, B., Yum, T.-S.P.: Coolstreaming/DONet: A Data-driven Overlay Network for Live Media Streaming. In: Proc. IEEE INFOCOM’05, March 2005, Miami, FL, USA (2005)
8. Gao, W., Huo, L., Fu, Q.: Recent Advances in Peer-to-Peer Media Streaming Systems. China Comminications 5, 52–57 (2006)

Adaptive Interpolation for Error Concealment in H.264 Using Directional Histograms

Toan Nguyen, Mi Seon Park, Won Taek Lim, Deok Jae Choi, Guee Sang Lee, and Jae Myung Yoo

Dept. of Computer Science, Chonnam National University
300 Yongbong-dong, Buk-gu, Gwangju, Korea,
toan_mulmi@hotmail.com, sunnydix@hanmail.net,
{djchoi, gslee}@chonnam.ac.kr

Abstract. Error Concealment methods for Intra frames in H.264 reconstruct the missing macroblock by computing weighted average of the boundary pixels of the neighboring blocks. However, the simple averaging of pixel values leads to blurring and degrades the picture quality severely. To solve this problem, an adaptive interpolation considering directions of neighboring blocks is proposed. Directional interpolation is adaptively chosen by a threshold computed from the distribution of local directions in areas around the lost block. Experiments show improvement of picture quality of about 0.5~2.0 dB compared to existing methods.

Keywords: error concealment, H.264, adaptive interpolation, histograms, intra frames.

1 Introduction

Recently H.264/AVC[1] has become the most promising video compression standard for a large range of video consumer applications, such as television broadcasting, streaming multimedia and video conferencing[4]. Its algorithms mainly rely on two techniques: macroblock-based motion compensation for reduction of temporal redundancies[5] and transform domain coding for reduction of spatial redundancies. However, it introduces more sensitivity to channel disturbances[7]. To conceal the error in temporal and spatial domain, various techniques have been proposed. For spatial domain error concealment, mainly for intra frames of video sequences, boundary pixel values are interpolated linearly by computing the weighted average. But, the interpolation with simple averaging of pixels in H.264/AVC leads to the blurring problem and results in severe degradation of picture quality. In this paper, we propose an adaptive interpolation method considering edges and their directions in local patterns around the lost block, which is represented in a histogram.

The rest of this paper is organized as follows. In section 2, existing interpolation techniques are discussed and an adaptive interpolation method is presented in section 3. In section 4, experimental results are given with an analysis. Section 5 draws the conclusion.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 408–415, 2007.
© Springer-Verlag Berlin Heidelberg 2007

2 Spatial Error Concealment for H.264: Weighted Interpolation

Spatial concealment algorithm for INTRA frame in H.264 is usually based on weighted pixel interpolation which recovers each pixel of the lost block from neighboring pixels, as shown in Fig.1. Assuming that the neighboring blocks are decoded correctly, *weighted averaging*[3] computes pixel value $\tilde{Y}(x, y)$ by interpolation as

$$\tilde{Y}(x, y) = \frac{\sum_{n=1}^{4} Y^n \times [15 - d_n]}{\sum_{n=1}^{4} d_n}, \tag{1}$$

where $\tilde{Y}$ is the reconstructed pixel value, Y^1, Y^2, Y^3, Y^4 are the one pixel wide boundary of each neighboring blocks and d_1, d_2, d_3, d_4 are the distance between $\tilde{Y}$ and Y^1, Y^2, Y^3, Y^4.

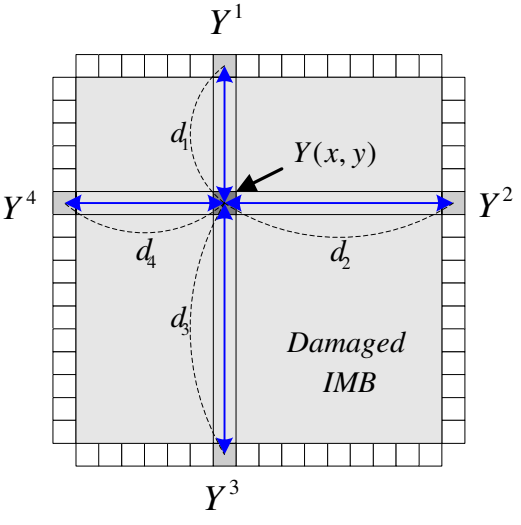


Fig. 1. Spatial Error Concealment in H.264

Weighted averaging performs well only when the missing block is in a smooth region, otherwise it makes blurring or visible artifacts. To overcome the drawback of weighted averaging, several approaches have been proposed including an iterative method based on projection onto convex sets[8], directional interpolation and its variants[9, 10]. However these approaches either suffers from high complexity or does not adapt into varying situations, thus may not properly work in real time applications.

3 Proposed Algorithm

3.1 Edge Estimation of the Lost Block

To estimate the edge direction of missing macroblock, the neighboring blocks are checked if they are corrupted or not. If four neighboring blocks are decoded without any errors, 16x8 subblocks for edge estimation are selected as upMB and downMB and 8x16 areas are selected as leftMB and rightMB in Fig.2 to evaluate the directions of edges in four neighboring blocks.

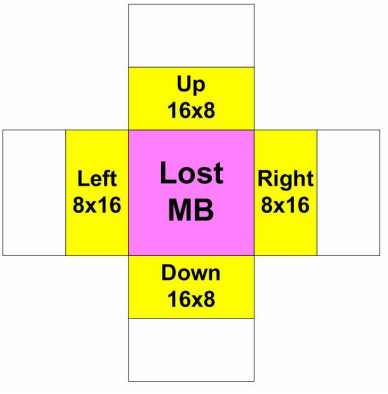


Fig. 2. Neighboring MB for Edge Estimation

Gradient filters are employed to detect the edge in a simple and fast way[9]. We use Prewitt operators:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } G_y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}. \quad (2)$$

The main edge's magnitude(M) and direction(D) are given by

$$M = \sqrt{G_x^2 + G_y^2}\ ,\ D = \arctan(\frac{G_y}{G_x}). \quad (3)$$

The directions D are classified into 8 categories as shown in Fig.3. For each estimation subblocks of upMB, downMB, leftMB and rightMB, the edges are detected for all pixels which can be applied to Prewitt operator and their directions with magnitude are recorded in a counter for the computation of statistical tendency. The direction which yields the largest count is selected as the main direction of the subblock among eight directions.

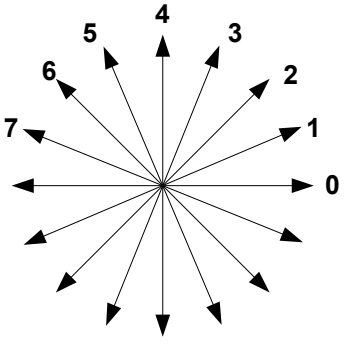


Fig. 3. Neighboring MB for Edge Estimation

3.2 Histogram of Neighboring Directions

The `BestDir` is the direction of the main edge with the largest count and weighted interpolation in the `BestDir` is taken on each missing pixel. If the `BestDir` is not stronger than other directions, that is, all of edge counts are almost the same, weighted interpolation in the `BestDir` seldom leads to improved quality of the images. To selectively apply the directional interpolation, we use the direction histogram and measure the threshold by which the simple weighted interpolation or the directional interpolation is chosen.

$$Dist\,(d_{best}) = \frac{\sum_{i=0}^{7} f_i \times w_i}{\sum_{i=0}^{7} f_i} \tag{4}$$

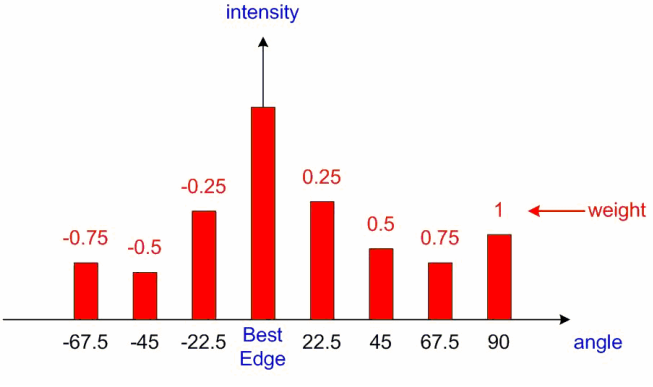


Fig. 4. Direction Histogram of Neighbor MB

The threshold $Dist(d_{best})$ is the fitness function to decide the threshold by which the interpolation type is selected between the average interpolation and the directional interpolation. f_i denotes the count of the i-th edge among all of edges and w_i is weighted coefficient of the i-th component as shown in Fig.4.

3.3 Adaptive Interpolation

The missing macroblocks are concealed with either the average interpolation in H.264 or the directional interpolation by a threshold which is compared with $Dist(d_{best})$. Fig.5 shows examples where `BestDir` indicates (a) the image pattern has clearly a conspicuous edge component, or (b) there are no special directional patterns in the image. In the case of Fig. 5 (a), directional interpolation is used while in Fig. 5(b) weighted interpolation is used. By selectively using the interpolation method by the image characteristic, the direction histogram can generate images with better quality than the case where only a simple weighted interpolation is used.

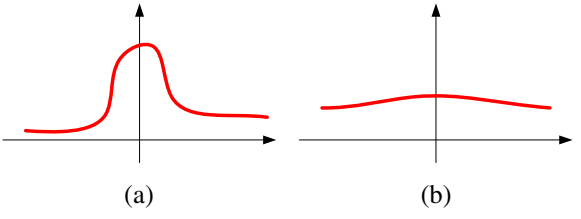


Fig. 5. Two types of Direction Histogram

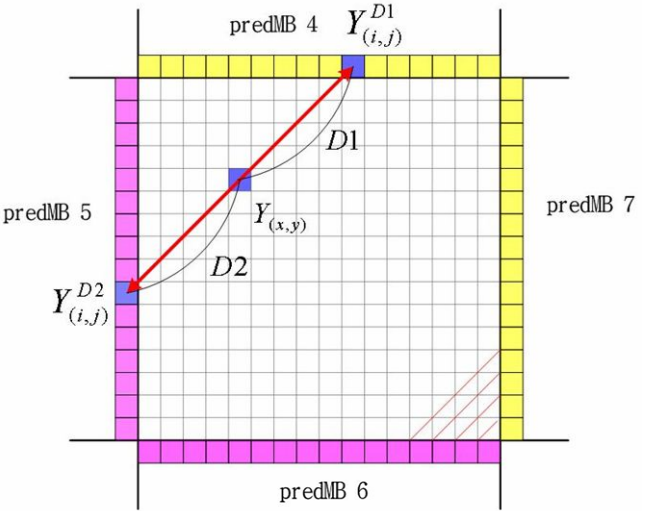


Fig. 6. Directional Interpolation

$Dist(d_{best}) < Threshold$ is the case of (a) in Fig.5 and the directional interpolation by the `BestDir` works as shown in Fig.6. Fig.6 shows an example of `BestDir` =2. (b) of Fig.5 is in $Dist(d_{best}) > Threshold$, and then the average interpolation works as shown in Fig.2. The pixel value in the missing macroblocks is obtained by

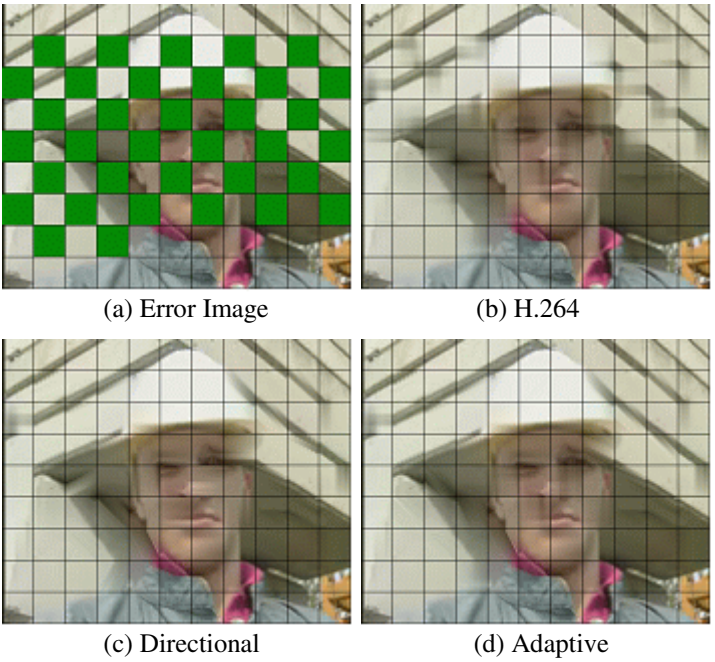
$$\tilde{Y}(x,y) = \left(\frac{1}{D_1 + D_2}\right) \times \left(Y_{(i,j)}^{D1} \times D_2 + Y_{(i,j)}^{D2} \times D_1\right) \tag{5}$$

$\tilde{Y}$ is the reconstructed pixel value. $Y_{(i,j)}^{D1}$ and $Y_{(i,j)}^{D2}$ are the pixels on the boundaries along the `BestDir` and D_1, D_2 are the distances.

4 Test Results

A new error concealment method has been proposed and implemented into the Joint Model (JM) version 8.6 H.264 software [11]. QCIF videos are encoded with slicing mode 1 and the performance was evaluated with an error of 4 to 11 lost packets. Fig.7 shows the results of INTRA frame of Foreman sequence by each methods and Fig.8 show the results of Claire.

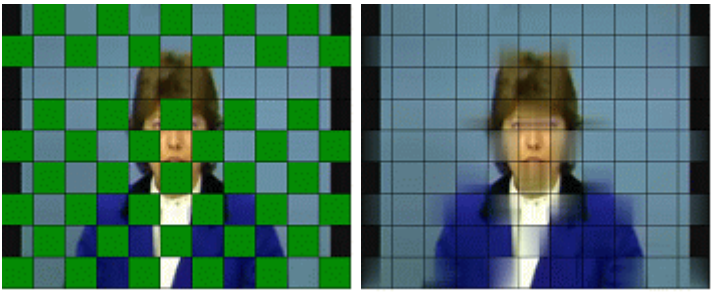
We compared PSNR performance of the proposed method with others for 150 frames of 10 QCIF sequences to compare with existing methods as shown in Fig. 11. Fig.9 and Fig.10 show the PSNR graphs of Foreman and Claire images.



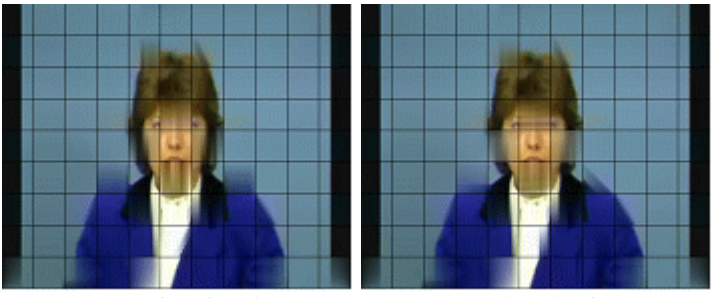
(a) Error Image (b) H.264

(c) Directional (d) Adaptive

Fig. 7. Result with Foreman.qcif



(a) Error Image (b) H.264



(c) Directional (d) Adaptive

Fig. 8. Result with Claire.qcif

Table 1. PSNR results

Video	H.264 PSNR (A)	Xu PSNR (B)	Proposed PSNR (C)	(B)-(A)	(C)-(A)	(C)-(B)
Akiyo	27.28	27.62	27.83	0.33	0.54	0.21
Carphone	26.65	27.22	27.35	0.57	0.70	0.13
Claire	24.75	26.24	26.53	1.49	1.78	0.28
Container	25.90	25.20	25.92	-0.69	0.02	0.71
Foreman	25.99	27.72	28.25	1.73	2.26	0.52
Mother	28.08	28.63	28.80	0.54	0.72	0.17
News	27.33	26.70	27.49	-0.62	0.16	0.78
Salesman	28.15	27.15	28.17	-1.00	0.02	1.01
Silent	28.25	27.74	28.52	-0.50	0.27	0.77
Suzie	31.06	30.61	31.26	-0.45	0.19	0.64

5 Conclusion

In this paper, we proposed an adaptive method for interpolation in error concealment of intra frames of H.264. Instead of applying only a simple weighted interpolation with pixels lying on the boundary of the lost block in the direction of horizontal and vertical line of pixel to be restored. Our method objectively improved error video

more higher 0.5~2 dB PSNR than the simple weighted, or average, interpolation and Xu's method[2].

Acknowledgments

This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government.

References

1. Thomas, W., Gary, S., Ajay, L.: Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC). Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT – G050r1, Geneva, Switzerland (May 2003)
2. Yanling, X., Yuanhua, Z.: H.264 Video Communication Based Refined Error Concealment Schemes. Consumer Electronics, IEEE Transactions 50(4), 1135–1141 (2004)
3. Sun, M.T., Reibman, A.R.: Compressed Video over Networks. Signal Processing and Communications Series, Marcel Dekker, New York (2001)
4. Lee, Y.C., Altunbasak, Y., Mersereau, R.M.: Multiframe error concealment for MPEG-coded video delivery over error-prone networks. Image Processing, IEEE Transactions 11(11), 1314–1331 (2002)
5. Jung, Y.H., Kim, Y.G., Choe, Y.S.: Robust error concealment algorithm using iterative weighted boundary matching criterion. Image Processing. In: Proceedings. 2000 International Conference. vol. 3, pp. 384–387 (2000)
6. Yue, W., Yu, S.G.: Joint source-channel decoding for H.264 coded video stream. Consumer Electronics. IEEE Transactions 51(4), 1273–1276 (2005)
7. Wenger, S.: H.264/AVC over IP. Circuits and Systems for Video Technology. IEEE Transactions 13(7), 645–656 (2003)
8. Sun, H., Kwok, W.: Concealment of Damaged Block Transform Coded Images Using Projections onto Convex Sets. IEEE Transactions on Image Processing 4(4) (April 1995)
9. Suh, J.W., Ho, Y.S.: Error concealment based on directional interpolation. Consumer Electronics, IEEE Transactions 43(3), 295–302 (1997)
10. Nemethova, O., Al-Moghrabi, A., Rupp, M.: Flexible Error Concealment for H.264 Based on Directional Interpolation. In: Wireless Networks, Communications and Mobile Computing, 2005 International Conference, vol. 2, pp. 1255–1260 (2005)
11. H.264/AVC Software Coordination: JM Software. ver. 8.6, available in http://iphome.hhi.de/suehring/tml/

Players and Ball Detection in Soccer Videos Based on Color Segmentation and Shape Analysis

Yu Huang, Joan Llach, and Sitaram Bhagavathy

Thomson Corporate Research,
2 Independence Way, Princeton, NJ 08540, USA
{yu.huang2, joan.llach, sitaram.bhagavathy}@thomson.net

Abstract. This paper proposes a scheme to detect and locate the players and the ball on the grass playfield in soccer videos. We put forward a shape analysis-based approach to identify the players and the ball from the roughly extracted foreground, which is obtained by a trained, color histogram-based playfield detector and connected component analysis. We employ Euclidean distance transform to extract skeletons for every foreground blob, and then perform shape analysis to remove false alarms (non-player and non-ball blobs) and cut-off the artifacts (mostly due to playfield lines) based on skeleton pruning and reverse Euclidean distance transform. Results are given to demonstrate the proposed algorithm works well in soccer video clips.

Keywords: Object detection, shape analysis, distance transform, skeleton pruning.

1 Introduction

The players and the ball are the most important objects in soccer videos. Detection and tracking of them are motivated by various applications, such as event detection, tactics analysis, automatic summarization and object-based compression [17].

Methods of locating the ball as well as players in soccer videos can be split in two groups: the first group makes use of fixed cameras (usually calibrated in advance) in a controlled environment [9]; the second group uses only regular broadcasting videos [3]. While the former can provide better performance, the latter is more flexible. In this paper, we focus on those efforts made in the second group.

In [3], a pioneering work in soccer video analysis was reported. The ball is detected by both its chromatic and morphological features. The ball is assumed to be among the white regions and its circularity (an indicator of resemblance of the shape to a circle) bigger than a given threshold. For the players, their uniform colors are recognized based on peaks (other than green) in the color histogram.

In [11], the authors propose a modified Circle Hough Transform to detect the ball on selected frames. Their method, however, cannot handle occlusions and requires the ball to be homogeneous, which greatly limit its application. In [10], there is an application of ball detection for video compression. Color histogram back projection

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 416–425, 2007.
© Springer-Verlag Berlin Heidelberg 2007

and intensity template matching are used to detect the ball. It is claimed that trajectory knowledge is used for prediction, but no details are given in the paper.

Unlike the object-based algorithms, a trajectory-based algorithm was put forward in [18] for detecting and tracking the ball. In this off-line framework, how to obtain the ball candidates, remove the false alarms and verify the candidates becomes critical. Similarly, trajectory constraints are also used in [7], where the difference is that a Viterbi algorithm rather than a Kalman filter is employed to verify the ball's trajectory candidates.

Since soccer is a spectator sport, the play fields, the lines, the ball and the clothing of the players are designed to be visually distinctive in color. Therefore, some approaches tried to alleviate detection difficulties by finding the grass playfield first [1, 8, 15, 16], using color segmentation and post-processing with morphological operations, such as connected component analysis, in order to limit the search area. However, how to remove false alarms (e.g. small areas that look like the ball) and cut-off artifacts, especially when player blobs merge with the field lines, is still a challenging problem.

A simple way to represent the playfield color is using a constant mean color value that is obtained through prior statistics over a large data set. The color distance between a pixel and the mean value of the field is used to determine whether it belongs to the field or not [15]. Since the soccer field is roughly green colored, the hue component defined in Smith's hexagonal cone model can be used to detect green-colored pixels within a certain range [16].

Statistically, single Gaussian or mixture of Gaussian (MoG) can be learned to represent a color model of the playfield and adapted by the Expectation-Maximization (EM) algorithm incrementally [8]. A non-parametric color model for the playfield is the histogram, where the dominant color is typically used for playfield detection assuming the field has a uniform shade of green and occupies the largest area in each frame [1].

More efforts in shape analysis are made for foreground blobs obtained by background subtraction, but most of the techniques rely on silhouette and region features [4]. Recently, skeleton analysis has been employed for object recognition [5, 12], shadow removal [13] and shape filtering [2, 14].

This paper proposes a shape analysis-based approach to identify the players and the ball from the roughly extracted foreground, which is obtained by a trained color histogram-based playfield detector and connected component analysis. We employ Euclidean distance transform to extract skeletons for every foreground blob, and then perform shape analysis to remove false alarms (non-players and non-ball blobs) and cut-off the artifacts (mostly due to playfield lines) based on skeleton pruning and reverse Euclidean distance transform.

2 Playfield Pixel Detector

Soccer is normally played on a grass field. In order to detect the players and the ball, a useful first step is to detect the pixels that form the playfield. In our work, we employ a simple but powerful histogram learning technique to detect the playfield pixels. This

framework was proposed by Jones et al. [6] for skin detection. Color models are learned for playfield pixels and non-playfield (background) pixels, using a training set of soccer videos. The color model is an RGB color histogram with N bins per channel in the RGB color space.

The playfield and non-playfield models are learned as follows. The pixels in the training set videos are labeled as playfield pixels and non-playfield pixels, either manually or using a semi-supervised method. Based on its RGB color vector, each labeled playfield pixel is placed into the appropriate rgb bin of the playfield histogram. A similar process is carried out for the pixels labeled as non-playfield. The histogram counts are converted into a discrete probability distribution:

$$P(rgb|\,\text{playfield}) = \frac{f(rgb)}{T_f}, \quad P(rgb|\,\text{non - playfield}) = \frac{n(rgb)}{T_n}, \tag{1}$$

where $f[rgb]$ is the pixel count in bin rgb of the playfield histogram, $n[rgb]$ is the corresponding count from the non-playfield histogram, and T_f, and T_n are the total counts contained in the playfield and non-playfield histograms, respectively.

A playfield pixel classifier is derived through the standard likelihood ratio approach [6]. A particular RGB value is labeled playfield if

$$\frac{P(rgb|\,\text{playfield})}{P(rgb|\,\text{non - playfield})} \geq \theta, \tag{2}$$

where $\theta \geq 0$ is a threshold which can be adjusted to trade-off between correct detections and false positives.

The number of bins per channel, N, and the detection threshold, θ, can be chosen based on the receiver operating characteristic (ROC) curve. The ROC curve shows the relationship between correct detections and false detections as a function of the detection threshold θ. ROC curves are computed using a test set of soccer videos (usually the labeled data set is separated into a training set and a test set).

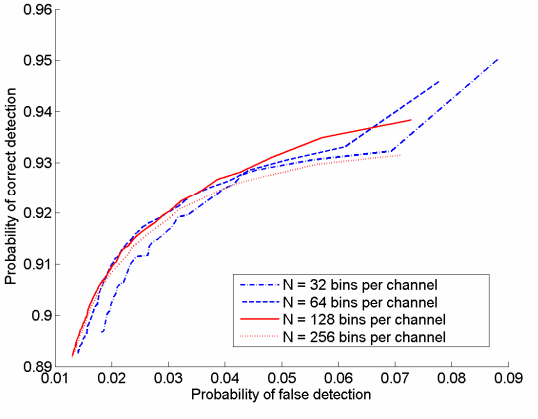


Fig. 1. ROC curves for playfield pixel classification with respect to N

Fig. 1 shows a family of ROC curves for different values of N. The y-axis gives the fraction of pixels labeled as playfield that were classified correctly, and the x-axis gives the fraction of non-playfield pixels that were wrongly classified as playfield. The higher the area under an ROC curve, the better the overall detection performance. Accordingly, the color histogram model with $N = 128$ bins is found to give the best results. After choosing N, a suitable operating point is selected on the corresponding ROC curve, to give the "best" (application dependent) trade-off between the probability of correct detections and the probability of false positives. The value of θ corresponding to the operating point is then used for playfield pixel classification through (2). For our experiments, we have chosen $\theta = 0.1$ and $N = 128$, and the corresponding probability of correct detection is 0.935 and the probability of false detection is 0.057.

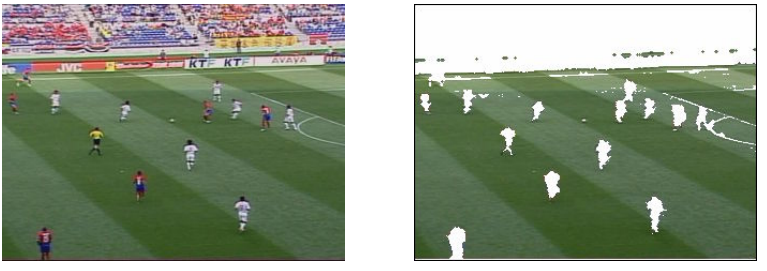


Fig. 2. Playfield detection result: the original frame and the detected playfield pixels

Fig. 2 shows the result of the playfield detection method described above. The original frame is displayed on the left. The detected playfield pixels are shown on the

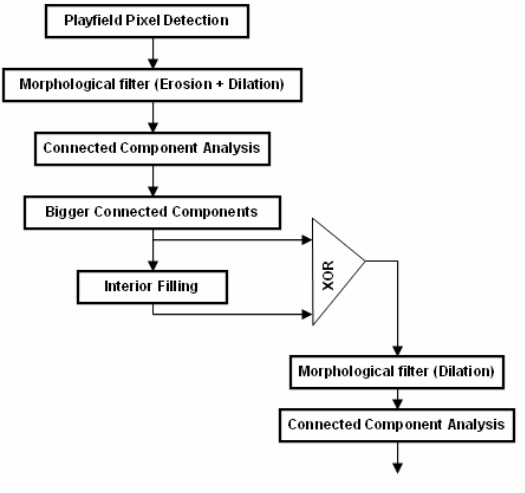


Fig. 3. Foreground blob extraction in soccer video

right. A morphological opening operation was applied after detecting playfield pixels in order to remove small false positive blobs. It can be observed from Fig. 2 that the method is successful in detecting playfield pixels under significant variations in color (dark and light stripes) and illumination (sunlit and shaded areas).

3 Playfield Extraction and Foreground Blob Detection

Fig. 3 illustrates the diagram of foreground extraction. Going through the playfield pixel detector, each frame yields a binary mask. Morphological filtering (erosion and then dilation) is applied to eliminate the noise.

Connected components analysis (CCA) scans the binary mask and groups its pixels into components based on pixel connectivity. The playfield is supposed to consist of several connected components, each of them bigger in size than a given area threshold. Thus, the playfield mask is obtained by filling the interior of each bigger component.

Ideally, the non-playfield pixels inside the extracted playfield areas should be the foreground pixels that can be grouped into different foreground blobs by CCA. Since the ball is typically quite small (less than 30 pixels in size for QVGA content), we cannot use erosion before grouping pixels into different blobs, but we can still use dilation. Therefore, noisy areas have to be removed later by shape analysis and true object appearance model.

4 Shape Analysis for Foreground Cleaning

Region properties for each blob are given by a few basic shape descriptors, such as perimeter P, area A, major/minor axes C_L / C_S, roundness $F = P^2 / (4\pi A)$ and eccentricity $E = C_L / C_S$.

Skeletons provide an intuitive, compact representation of a shape. One of its important features is separation of the shape's topological properties from its geometric properties. Here we employ distance transform to extract skeleton and generate additional shape descriptors for use.

4.1 Distance Transform

Given a foreground binary mask W of size *m x n*, $\overline{W}$ denotes the complementary of W, i.e. the set of background pixels. Its *squared distance transform* (SDT) [2] is given by $H = \{h(x,y)\}$, i.e.

$$h(x, y) = \min\{(x-i)^2 + (y-j)^2; 0 \le \mathrm{i} < \mathrm{m}, 0 \le j < n, (i, j) \in \overline{W}\}. \quad (3)$$

Given a set of foreground pixels *V* and a picture $F = \{f(x,y)\}$ of size *m x n*, such that $f(x,y)$ is set to the SDT value $h(x, y)$ when the pixel (x,y) belongs to *V* and 0 otherwise. Then *reverse Euclidean distance transform* (REDT) [2] of *V* consists in obtaining the set of points *W* such that

$$W = \{(x, y) \mid \max\{f(i, j) - (x - i)^2 - (y - j)^2\} > 0, \quad 0 \le i < m, 0 \le j < n, (i, j) \in F\}. \quad (4)$$

Thus, if we compute the map $H^* = \{h^*(x,y)\}$ such that

$$h^*(x, y) = \max\{f(i, j) - (x - i)^2 - (y - j)^2, 0 \le i < m, 0 \le j < n, (i, j) \in F\}, \quad (5)$$

then we obtain W by extracting from H^* all pixels of strictly positive values.

4.2 Skeleton Descriptors

Given SDT of the shape $H = \{h(x,y)\}$, its skeleton *Ske* is defined by [2]

$$Ske = \{(x, y) \mid \exists (i, j), (x - i)^2 + (y - j)^2 < h(x, y),$$

$$\text{AND} \max_{(u,v)} \{h(u, v) - (x - u)^2 - (y - v)^2\} = h(x, y) - (i - x)^2 - (j - y)^2\}. \quad (6)$$

Comparing with the definition of REDT, we see that applying REDT to the SDT of the shape, in which only the upper envelope elliptic paraboloids are marked, can yield the *Ske*. In Fig. 4, a skeleton of player blob is shown.

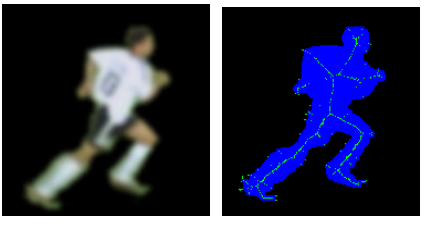


Fig. 4. Skeleton results: the player blob and its skeleton (in green)

As explained by [2], the Power diagram is a special kind of Voronoi diagram and the Power labeling assigns to each point of the plane the index of the cell containing it in the Power diagram (more details are given in [2]). When we apply REDT to get the skeleton, simultaneously the Power labeling for each skeleton point is generated as well.

Eventually, we get those additional descriptors for shape analysis as: length of the skeleton L_s, covering C_{ske}, (i.e. the number of pixels in Power labeling associated with the skeleton), maximal thickness of the skeleton $d_{\max}$, (i. e. radius of the maximal ball which covers the blob, correspondingly the center of the maximal ball is one of the skeleton pixel) elongation G= $A/(2d_{\max})^2$ and aspect ratio T= $L_s / d_{\max}$ etc.

4.3 White Pixel Proportion in Skeleton's Covering

The ball and the field lines in wide view shots are nearly white. For each skeleton point, the white pixel proportion in its covering (Power labeling) is a good indicator for artifact detection. A simple method to identify white pixels in each foreground blob is

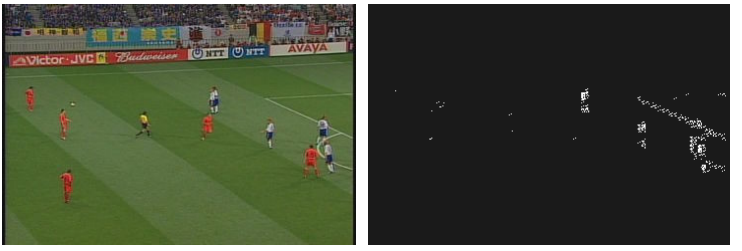


Fig. 5. White pixel detection: the original frame and the detected white pixels

$$X(x,y)=\begin{cases}1, (r(x,y)-1/3)^2+(b(x,y)-1/3)^2 < a \text{ AND I(x,y)} > \text{b} \\ 0, \qquad\qquad otherwise\end{cases}, \tag{7}$$

where $r(x,y)$ and $b(x,y)$ are normalized red and blue components for pixel (x,y), $I(x,y)$ denotes the intensity value in the range [0,255], and $X(x,y)$ is the white pixel mask. Fig. 5 shows results of white pixel detection. Below we use it for ball detection and player blobs' skeleton pruning.

4.4 Ball and Player (Referee) Localization

Apparently, there is a predefined range for the ball blob's area A according to the camera configuration. Meanwhile, the proportion of detected white pixels $X(x,y)$ by (4) in the blob indicates the possibility of being a ball. Roundness F and eccentricity E for a blob candidate should be close to 1.0, different from disconnected segments of field lines. Eventually, detection of the ball blob is carried out by

$$B=\begin{cases}1, p_{\text{W}} > \text{r}_{\text{w}} \text{ AND } \text{ar}_{\min} < \text{A} < \text{ar}_{\max} \text{ AND } F < f_{\max} \text{ AND E} < e_{\max} \\ 0, \qquad\qquad otherwise\end{cases}, \tag{8}$$

where $p_W = C\{(x,y) \mid X(x,y)=1\}/A$, C{.} is counting the number of white pixels. (All the thresholds are empirically determined.)

For an isolated player (or referee) blob, its shape also can also be approximated by an ellipse, limiting its elongation and aspect ratio. These constraints can help removing some fragments of field lines. In addition, since field lines look nearly white in color, so can use the proportion of white pixels as an index for non-player blobs. Similarly, an area range for the player (or referee) blob is predefined according to the camera configuration.

4.5 Skeleton Pruning

When the player (referee) blob merges with fragments of field lines, shape analysis becomes complex. We propose a skeleton pruning-based method to cut-off those artifacts.

First, the width of a field line is less than that of a player body in the wide view shot; so at every skeleton point we first check whether its thickness is small compared

with the maximum skeleton thickness, i.e. smaller than $c \cdot d_{\max}$, where c is thickness factor. If it is true, then we calculate average values of RGB components for all pixels covered by this skeleton point (Power labeling). The observed result is that the average RGB components are close to those of the white color when this skeleton point corresponds to a fragment of field lines.

To avoid excessive pruning, we add the distance constraints: the pruned skeleton is relatively far away from the human body centroid which corresponds to the skeleton point with the maximal covering value, i.e. its horizontal distance is bigger than $d_h \cdot d_{\max}$ and its vertical distance $d_v \cdot d_{\max}$, where d_v, d_h are distance factors.

5 Experimental Results

We test our proposed algorithm with three video clips (about 300 frames for each clip) and the results are quite encouraging. Some detection results are shown in Fig. 6-8. For each figure, the extracted grass field is given at the bottom left, the initial segmented foreground is shown at bottom right, the cleaned foreground blobs by shape analysis and skeleton pruning (skeleton is in green and the blob is in blue) are displayed at top left and the final detected ball and players is shown at top right, where each detected object is enclosed by an ellipse in yellow.

It is shown most of false alarms from field lines are removed. Artifacts in player blobs due to merging with field lines are cut-off. Our proposed method overperforms previous methods as [1, 16]. The method in [1] can't detect the ball and have to set the ball's location manually; it can't remove artifacts in players' localization and require

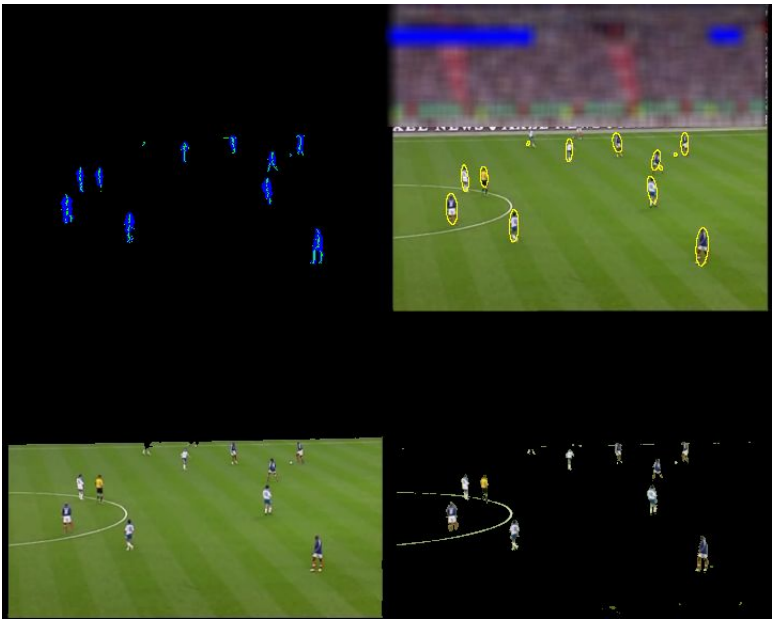


Fig. 6. Detection results in one frame of video 1

Fig. 7. Detection results in one frame of video 2

Fig. 8. Detection results in one frame of video 3

color histogram models learned beforehand to detect players' presence. The system in [1] focus on player detection only and it fails in finding players when they merge with field lines, though tracking modules are employed to handle overlapping/occlusions.

However we still find in our results, players standing at the border of the field might be missed since we extract the playfield region first; sometimes the player's feet or legs are not extracted by the playfield detector because some of their pixels look field-green due to the course graininess.

6 Conclusions and Future Work

In this paper, a shape analysis-based soccer ball and the players detection method has been proposed. We propose a learned color histogram model to detect the playfield pixels and group them into a playfield region. Then, the foreground blobs are

extracted with morphological processing. Shape analysis and skeleton pruning are performed to remove false alarms (non-players/referees and non-ball) and cut-off the artifacts (mostly due to playfield lines).

In future, we will work on how to separate the players (referees) when they overlap and partially occlude each other. Besides, we will consider incorporating the proposed detector with a ball tracking system and a player/referee tracking system.

References

1. Choi, S., Seo, Y., et al.: Where are the ball and players? Soccer game analysis with color-based tracking and image mosaic. In: Int. Conf. on Image Analysis and Processing (September 1997)
2. Coeurjolly, D., Montanvert, A.: Optimal Separable Algorithms to Compute the Reverse Euclidean Distance Transformation and Discrete Medial Axis in Arbitrary Dimension. IEEE T-PAMI 29(3), 437–448 (2007)
3. Gong, Y., Sin, L.T., Chuan, C.H., Zhang, H., Sakauchi, M.: Automatic parsing of TV soccer programs. In: Proc. Multimedia Computing & Systems, pp. 167–174 (1995)
4. Haritaoglu, I., Harwood, D., Davis, L.S.: Hydra: multiple people detection and tracking using silhouettes. In: 2nd IEEE Workshop on Visual Surveillance (1999)
5. He, L., Han, C.Y., Wee, W.G.: Object recognition and recovery by skeleton graph matching. In: IEEE ICME'06 (2006)
6. Jones, M., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. In: IEEE CVPR'99, June 1999, pp. 274–280 (1999)
7. Liang, D., et al.: A scheme for ball detection and tracking in broadcast soccer video. In: PCM 2005, Korea (2005)
8. Liu, Y., Jiang, S., Ye, Q., Gao, W., Huang, Q.: Playfield Detection Using Adaptive GMM and Its Application. In: IEEE ICASSP '05, March 2005, pp. 421– 424 (2005)
9. Needham, C.J., Boyle, R.D.: Tracking multiple sports players through occlusion, congestion and scale. In: BMVC'01 vol. 1, pp. 93–102 (2001)
10. Nementhova, O., Zahumensky, M., Rupp, M.: Preprocessing of ball game video sequences for robust transmission over mobile network. In: CDMA International Conference (CIC), Seoul (2004)
11. D'Orazio, T., et al.: A Ball Detection Algorithm for Broadcast Soccer Image Sequences. In: IAPR ICPR'02 (2002)
12. Ozcanli, O.C., Tamrakar, A., Kimia, B.B.: Augmenting Shape with Appearance in Vehicle Category Recognition. In: IEEE CVPR'06, June 2006, pp. 935–942 (2006)
13. Renno, J., Orwell, J., Thirde, D., Jones, G.A.: Shadow Classification and Evaluation for Soccer Player Detection. In: BMVC'04, September 2004, Kingston, (2004)
14. Tam, R.C., Heidrich, W.: Feature-Preserving Medial Axis Noise Removal. In: ECCV'02, pp. 672–686 (2002)
15. Tong, X., et al.: An Effective and Fast Soccer Ball Detection and Tracking Method. In: ICPR'04, pp. 795–798 (2004)
16. Utsumi, O., Miura, K., Ide, I., Sakai, S., Tanaka, H.: An object detection method for describing soccer games from video. In: IEEE ICME '02 (2002)
17. Wang, J.R., Parameswaran, N.: Survey of Sports Video Analysis: research issues and applications. In: Pan Sydney Area Workshop on Visual Information Processing (VIP03) (2003)
18. Yu, X., Xu, C., Tian, Q., Leong, H.W.: A ball tracking framework for broadcast soccer video. In: ICME'03 (2003)

Efficient Image Retrieval Using Conceptualization of Annotated Images

Miyoung Cho[1], Chang Choi[1], Hanil Kim[2], Jungpil Shin[3], and Pankoo Kim[4,*]

[1] Dept. of Computer Science and Engineering
Chosun University, 375 Seosuk-dong Dong-Ku Gwangju 501-759, Korea
irune@chosun.ac.kr, enduranceaura@gmail.com
[2] Dept. of Computer Education
Cheju National University, 66 Jejudaehakno Jeju 690-756, Korea
hikim@cheju.ac.kr
[3] Graduate School of Computer Science and Engineering
Aizu University, Aizu-Wakamatsu City, Fukushima, 965-8580, Japan
jpshin@u-aizu.ac.jp
[4] Dept. of CSE, Chosun University, Korea
pkkim@chosun.ac.kr

Abstract. As the amount of visual information is rapidly increasing, users want to find the more semantic information easily. Most retrieval systems by low-level features(such as color, texture) could not satisfy user's demand. To interpret semantic of image, many researchers use keywords as textual annotation. However, it's the image retrieval without ranking by text matching which is the simplest way to retrieval according to keyword's existence or nonexistence. In this paper, we propose conceptualization by similarity measure using relations among keywords for efficient image retrieval. We experiment annotated image retrieval by lowering the unrelated keyword's weight value and raising important keyword's one.

1 Introduction

The amount of visual information is rapidly increasing due to digital cameras and mobile telephones equipped with such devices. Also this proliferation of visual information has led to creating a demand for new methodologies to retrieve specific information with easy-to-use manner. To satisfy the demand, recent research has generated significant improvements in Visual Information Retrieval(VIR). Despite these extensive research activities in computer vision and image processing, content-based image retrieval is not yet capable of complete understanding of semantics in visual information. That is, matching based on color, size, texture and shape are within the state of the art.

Specially, semantic interpretation of image is incomplete without some mechanism for understanding semantic content that is not directly visible. For this reason, human assisted content-annotation through textual description (i.e, a keyword, or a simple

* Corresponding author.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 426–433, 2007.
© Springer-Verlag Berlin Heidelberg 2007

sentence) is the one of most common methods, and provides a means for exploiting syntactic, semantic as well as lexical information. Textual annotations may convey name of a visual object, property of a visual object, event happening in visual context. Problems in text (i.e., keywords) based image retrieval require exact remembrance of annotated word for retrieval. However, it's the image retrieval without ranking by text matching which is the simplest way to retrieval according to keyword's existence or nonexistence.

There are missing images that are related to query but have irrelative(noisy) keywords. So, noisy keywords make retrieval results too bad. Keywords in an annotated image are only arrangement of words, while conceptualization represents relation between words as concept is the least unit of meaning. To improve results, we apply the conceptualization by the similarity measure between keywords to image retrieval.

The organization of the paper is as follows. In the next section, we describe background of study by defining noisy keyword and important keyword. In section 3, we discuss similarity measures using WordNet and describe conceptualization using representing relation between keywords. And, we measure semantic similarity between annotated images. In section 4, we apply to image retrieval and compare with results before and after excluding noisy keyword. And the conclusion section summarizes the paper.

2 Background of Study

As we mentioned, semantic interpretation of image can be described by a set of keywords. For example, two images in figure 1 have many keywords including 'animal', 'fowls' and so on. However, it's hard to say that 'branch' in (a) and 'line' in (b) well describe the meaning of each image. In addition to, 'fowl' in two figures is important. That is, we can't get the each keyword's priority (that is, weight value) in keyword-based image retrieval only using the set of keywords.

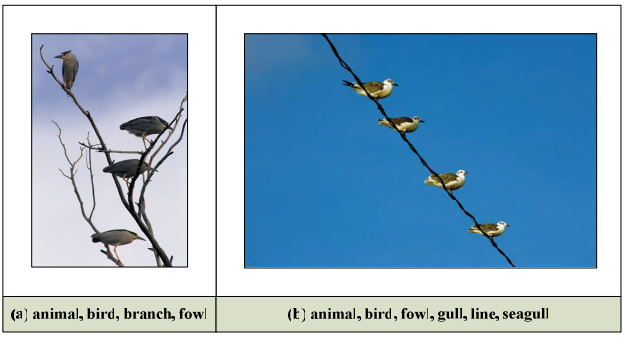


Fig. 1. The annotated images

In this paper, we define noisy keywords as irrelevant keywords and main keywords as important keywords. We discard noisy keywords from an annotated image and raise main keyword's weight value for efficient image retrieval. To consider

keyword's importance, we will utilize similarity keywords based on semantic distance. Prior to this, we introduce WordNet and investigate various semantic similarity measures between keywords in section 3.1. Each semantic similarity measure tried to find the distance between keywords using several different approaches.

3 Conceptualization in the Annotated Images

3.1 Similarity Measure Between Keywords

We will use the structure and content of WordNet for measuring semantic similarity between two keywords. WordNet is a freely available lexical database for English whose design is inspired by current psycholinguistic theories of human lexical memory[8]. English words are organized into synonym sets, so-called synsets, and each represents one underlying lexical concept. The nominal part of WordNet can be reviewed as a tangled hierarchy of hypo/hypernymy(IS_A) relations among synsets. The relations between synsets are semantic ones and the relations between words are lexical.

The state of the art can be classified to the three different categories such as: Node-based approach, edge-based approach and gloss-based approach. First, node-based approach uses the information content by Resnik[1, 5]. He proposed the following similarity measure between conceptual entities.

$$sim(c_i, c_j) = \max_{c \in S(c_i, c_j)} [-\log p(c)] \tag{1}$$

where, $S(c_i, c_j)$ is the set of concepts that subsume both c_i and c_j. To maximize the representative, the similarity value is the information content value of the node whose $H(c)$ value is the largest among those super classes.

Second, the edge based approach is a more natural and direct way of evaluating semantic similarity in WordNet. It estimates the distance (e.g. edge length) between nodes. In determining the overall edge based similarity, most methods just simply sum up all the edge weights along the shortest path. To convert the distance measure to a similarity measure, one may simply subtract the path length from the maximum possible path length[2].

Last, Lesk method is the representative algorithm of gloss-based approach. Lesk proposed that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. Banerjee and Pedersen extended this idea to use WordNet as the dictionary for the word definitions[6]. This notion was further extended to use the rich network of relationships between concepts present is WordNet. That is, this method not only considers the gloss of target word but also augments with the shared glosses by looking over all relations including hypernym, hyponym, meronym, holonym, troponym.

The described similarity measures have characteristics respectively and indicate different performance according to research field. That is, those approaches have inherent strength and weakness. To decide the similarity measurement, we use correlation coefficient to comparable similarity methods with M&C means by Miller and

Charles. In this paper, we use the first approach(that is, node-based method using information content) because it has the highest correlation value. Figure 2 illustrates a conceptual organization of a conceptual entity 'automobile' and their information content in WordNet.

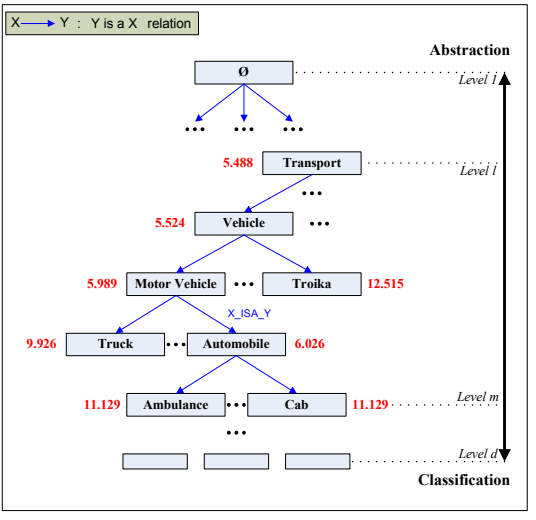


Fig. 2. Information content in the WordNet

Figure 3 shows a semantic representation of relationship using similarity measure between keywords in an annotated image. The each keyword is represented in the form of 'keyword#n#number'.

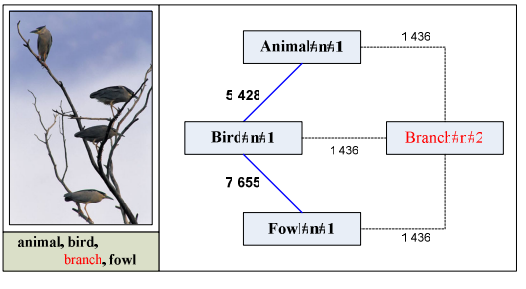


Fig. 3. Similarity measure between keywords

Where, 'n' indicates noun as parts of speech. Although WordNet contains not only noun but also verb, adjective and adverb, we only consider noun. The number is the meaning number of the word. We need to distinguish meaning because the word has many meaning. For example, 'branch#n#2' is as noun, its meaning is "a division of a stem, or secondary stem arising from the main stem of a plant" as second meaning.

As you can see in the figure 3, ‘bird’ concept is semantically close to concept of ‘fowl’. On the other hand, similarity among ‘branch#n#2’ and others is small and distance is the furthest. In other words, ‘branch’ keyword’s weight value is the smallest in keywords.

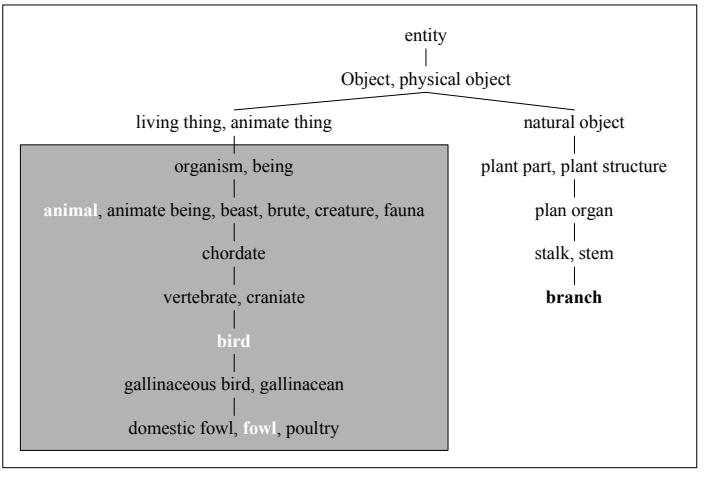


Fig. 4. Exclusion of noisy keyword(‘branch’)

Figure 4 indicates the position of each keyword and hierarchical structure by IS_A relation between keywords. ‘fowl’ is a kind of ‘bird’, ‘bird’ is also a kind of ‘animal’. That is, ‘animal’, ‘bird’ and ’fowl’ as organism’s hyponym share top nodes in compare with ‘branch’. So, ‘branch’ is a noisy keyword and is discarded from set of keywords.

3.2 Conceptual Comparison with Annotated Images

In this subsection, we will illustrate the calculation of the conceptual distance among annotated images. First of all, we look up images which include same keywords. If images consist of other keywords, we search common keywords and calculate

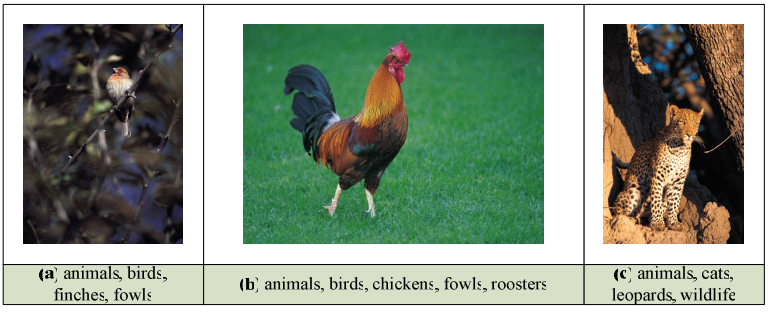
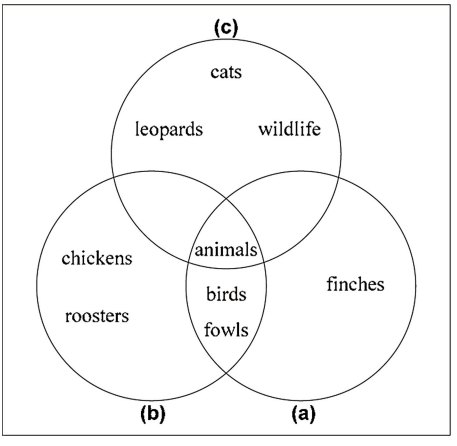


Fig. 5. Sample images

semantic distance among other keywords. For example, in figure 5, if you want to get conceptual similarity among 3 images, measure similarity between keywords except for common keywords(animals).

The similarity value between the first image and the second one is 7.655. In this calculation, both images have different keywords but they are the same conceptually. Although the third image has a common keyword(animal), the similarity value between the first image and the third one is 4.660. So, it's a further than the second image.



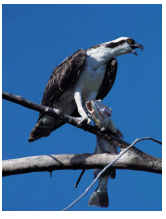
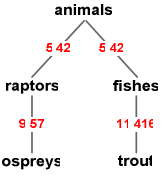
Keywords pair		Similarity
finches,	chickens	7.655
finches,	roosters	7.655
cats,	birds	5.820
cats,	finches	5.820
cats,	fowls	5.820
leopards,	birds	5.820
leopards,	finches	5.820
leopards,	fowls	5.820
wildlife,	birds	2.340
wildlife,	finches	2.340
wildlife,	fowls	2.340

Fig. 6. The comparison with annotated images

4 Experiment

In the experiment part, we used a sample image database containing nearly 2,000 selected images. The images we used in our experiments came from Microsoft's 'Design Gallery Live(http://dgl.microsoft.com)'. These images were mostly color pictures with non-uniform background. Each image has four or the more annotations. First,

Table 1. Examples of conceptualization

Image	keyword	conceptualization
	Animals Fishes ospreys Raptors troul	

we measure similarity among keywords for conceptualization. Table 1 shows sample annotated images to represent exclusion of noisy keyword by conceptualization.

After the conceptualization, we search by options selected for these experiments were:

1. Retrieval before exclusion of noisy keywords
2. Retrieval after exclusion of noisy keywords

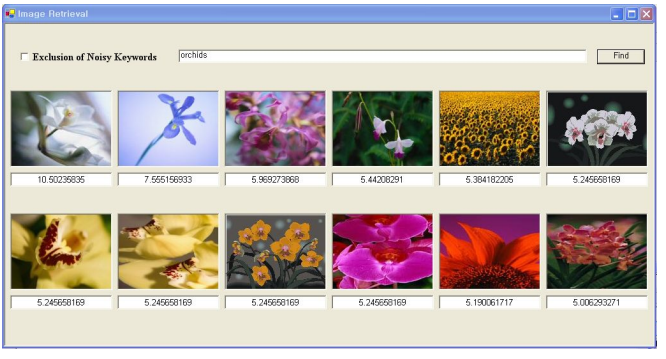


Fig. 7. The result before exclusion of noisy keyword

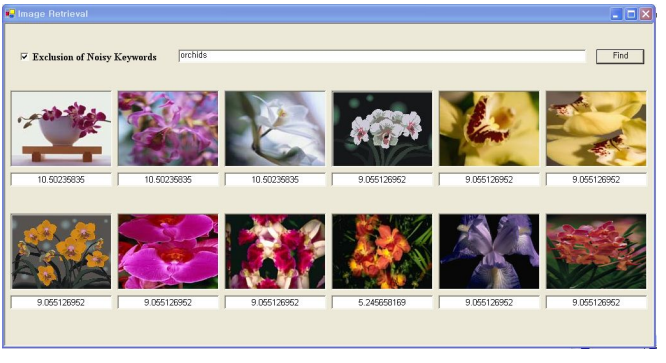


Fig. 8. The result after exclusion of noisy keyword

The first is a search for images that look like the image that contains the keyword 'orchid'. The search option is set such that we use the ontology with semantic similarity measurement before exclusion of noisy keywords, and the results of the retrieval are presented in Fig. 7. The value of each image is an average of the similarity measures between the annotated keyword of the image and the query keyword. Otherwise, Fig. 8 represents the result after exclusion of noisy keywords.

Both the 5^{th} and the 11^{th} images shown in Fig. 7, the images were relegated to out of bound in favor of other images that the system found to be closer matches. That is, system rearranges the images after exclusion of noisy keywords.

5 Conclusion

In this paper, we propose the efficient image retrieval through conceptualization (excluding of noisy keywords or raising weight value of main keywords). That is, we discarded irrelative keywords using similarity measure by Resnik's based on WordNet and applied to concept-based image retrieval.

In the future works, we will add new keywords(missing keywords because of annotator's different point of view) and apply to other multimedia retrieval just like video, audio and so on using textual metadata.

Acknowledgement

"This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Advancement)" (IITA-2006-C1090-0603-0040)

References

1. Resnik, P.: Using information content to evaluate semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453. Montreal (1995)
2. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, C. (ed.) WordNet: An electronic lexical database, pp. 265–283. The MIT Press, Cambridge, MA (1998)
3. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, Madison, WI (1998)
4. Hirst, G., St-Onge, D.: Lexical Chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An electronic lexical database, pp. 305–332. The MIT Press, Cambridge, MA (1998)
5. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applications to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
6. Banerjee, S., Pedersen, T.: Extended Gloss Overlaps as a Measure of Semantic Relatedness. In: the Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico (2003)
7. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: the Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together, Trento, Italy (2006)
8. http://wordnet.princeton.edu/
9. http://search.cpan.org/dist/WordNet-Similarity/
10. Yohan, J., Latifur, K., Lei, W., Mamoun, A.: Image Annotation by Combining Multiple Evidence & WordNet. In: Proceedings of the 13th annual ACM international conference on Multimedia table of contents, Hilton, Singapore (2005)

3-D Camera Modeling and Its Applications in Sports Broadcast Video Analysis

Jungong Han[1] and Peter H.N. de With[1,2]

[1] University of Technology Eindhoven, P.O.Box 513, 5600MB Eindhoven
{jg.han@tue.nl}
[2] LogicaCMG, TSE, PO Box 7089, 5605JB Eindhoven, The Netherlands

Abstract. This paper concentrates on a unified 3-D camera modeling technique, which can be applied to analyze several sports types. To this end, we have based our modeling on collecting feature points from two perpendicular planes: the ground plane and the net plane, as they exist in most court-net sports. A two-step algorithm is employed to extract and distinguish the feature lines and points from these two planes for determining the camera calibration parameters. Our proposed modeling enables a defined mapping from real 3-D scene coordinates to image coordinates. The use of this modeling helps in the improvement of many emerging applications, such as moving-player segmentation and 3-D scene adaptation. We evaluate the promising performance of the proposed modeling for a variety of court-net sports videos containing badminton, tennis and volleyball, and also demonstrate its capability on the case study of player segmentation and 3-D scene adaptation.

Keywords: Camera Calibration, Court-Net Detection, Sports Analysis, Application.

1 Introduction

Automatic content analysis of sports video is an interesting and challenging area for computer vision, since it enables new applications such as automatic tactic analysis, or virtual sports game generation. For such applications, it is necessary to know the position of the important objects, e.g., player and ball, in a 3-D coordinate system, rather than the position in the image. For this purpose, 3-D modeling based on camera calibration has to be established from the video input to determine the mapping of coordinates.

Research on camera modeling for sports video analysis can be broadly divided into two generations of the systems. In the first generation, literatures intend to find a 2-D to 2-D planar mapping, leading to the correspondences between a 2-D image plane and 2-D playing field plane. In [1], the authors describe a calibration algorithm for tennis courts, using a simplified camera model that only considers the camera tilt angle, the camera distance from the court field, and the focal-length. In [2], a more robust detection of the court (for soccer videos) is introduced, but the problem of complex calibration is still existing, since finding

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 434–443, 2007.
© Springer-Verlag Berlin Heidelberg 2007

calibration parameters based on court lines requires a computationally expensive initialization using an exhaustive search through the parameter space. Our previous work [3] proposed a more generic camera-calibration algorithm that can be adapted to all sports where the court consists of a sufficient number of straight lines. In general, this kind of approaches does not provide 3-D height information of the object, which is a key parameter in some applications. The second generation of camera calibration aims to provide all 3-D information of the objects. In [4], the authors estimate the ball's position on the basis of a simple camera model, but it suffers from two difficulties: (1) the algorithm has to find two objects manually, which are perpendicular to the playing field and have similar view depths, and (2) the method has to assume that the player's height is known. In [5], Liu *et al.* propose a self-calibration method to extract 3-D information in soccer video. Despite the initial features, lines need to be assigned manually, the method for computing intrinsic parameters is rather simple, which is unlikely to obtain the height information with sufficient accuracy. To sum up, the research in this field still requires to resolve two problems: (1) the technique should be a general algorithm that can be adapted to more than one sports game and more than one application, and (2) the algorithm should be fully automatic, robust against noise like lighting conditions and occlusions, it should be accurate enough, and support real-time execution.

This paper concentrates on a unified 3-D camera modeling technique for analyzing several sports types, intending to address the two challenges mentioned above. Our system is original in the two aspects. First, we propose a novel 3-D modeling, in which both the ground plane and the net plane are exploited. The use of a two-step algorithm helps to recognize a set of features at well-known positions within these two planes. Our modeling enables a mapping from 3-D coordinates to the image coordinates. Secondly, unlike most systems that directly adopt existing techniques in computer vision to analyze sports video, e.g., Sudhir *et al.* [1] employ a motion detection method to segment the players, we present here two new approaches building upon our 3-D modeling, thus leading to more accurate analysis results.

2 Camera Calibration Theorem

The task of the camera calibration is to provide a geometric transformation that maps the points in the real-world coordinates to the image domain. Since the real court-net model is a 3-D scene but the displayed image is a 2-D planar, this mapping can be written as a 3×4 transformation matrix $\mathbf{M}$, transforming a point $\mathbf{p} = (x, y, z, 1)^\top$ in real-world coordinates to image coordinates $\mathbf{p'} = (u, v, 1)^\top$ by $\mathbf{p'} = \mathbf{Mp}$, which is equivalent to

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \tag{1}$$

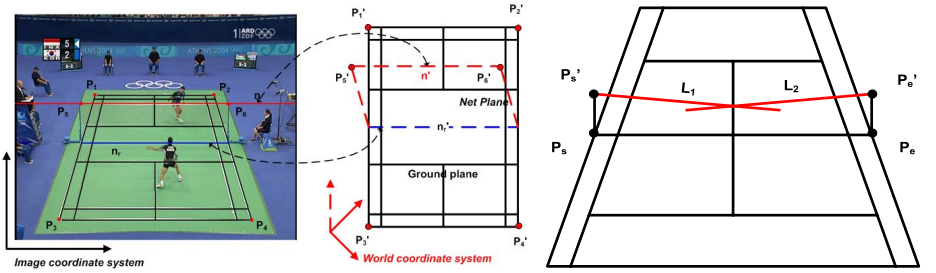


Fig. 1. (a): The planes, lines and points are selected in the image and the correspondences in the standard model are determined. Six intersection points are used for calibration. (b) More difficult case where the top net-line is broken into two pieces, which have different angles with the ground plane.

Since **M** is scaling invariant, eleven free parameters have to be determined. They can be calculated from six points whose positions are known in both the 3-D coordinates and the image. Note that these six points need not be fixed, but can be selected case by case, as a few points may be occluded in some views. Instead of using point features directly, we base our camera-calibration algorithm on lines, since detecting the accurate position of a specific point on a court is much more difficult than to estimate the position of line segments. Moreover, the detection of lines is more robust in the sense that they are hardly occluded completely. Another requirement is that these lines and points should be selected from *two* planes, which are perpendicular to each other. The second plane is required to provide the height parameter. Fortunately, the ground plane and the net plane are very suited to satisfy this requirement. Additionally, the court lines and net line can be used to characterize these two planes. Fig. 1(a) shows an example in which the points and planes used to calculate the calibration parameters are indicated. Clearly, the task so far is to identify and locate the lines on both the ground plane and the net plane. However, it may not be easily realized, since there is a problem that court lines and net line have similar properties, such as length, width and color, in the picture. For this reason, we propose a two-step algorithm to recognize them separately instead of attempting to detect them simultaneously.

2.1 Court-Line Recognition

The challenges for finding the court lines among the detected white lines are covered in two aspects. First, the detected white lines have similar features, like color, width and so on. Second, some court lines may be partly or completely occluded in the input image. We use here the layout (configuration) of the court, which must be exactly equal in both the standard court model and the image. The basic approach of our algorithm is to evaluate the distance between each layout formed by different white lines and the configuration of a standard court

model. The configuration with the best match is selected as the final solution. Note that we do not require an integrated court configuration, because we can even perform our algorithm in the case where we only have two vertical lines and two horizontal lines. Having such correspondences, a 2-D to 2-D projective mapping matrix **H** is established, transforming a point in real-world coordinates to image coordinates. With this mapping, each court line in the image can be located and related to the standard court model, although some court lines may be occluded. The entire system can be found in our previous work [3], comprising court-line pixel detection, line detector, modeling fitting and model tracking.

2.2 Net-Line Detection

This stage reviews many possible lines in the input image, which can be considered as the candidates for the net lines. Detection of a net line is carried out in three steps. The first step is to exclude the court lines from the detection, since the court lines look similar to the net line in the image. Using the matrix **H**, each court line in the image can be located (see the black lines of Fig. 1(a)). The second step imposes three constraints that must be satisfied for a line in order to be classified as a net line. Prior to introducing the constraints, two relevant definitions are first given: (1) we define n'_r in the right image of Fig. 1(a) as the bottom reference line of the net line, because it is the projection line of the net onto the ground plane, (2) accordingly, n_r corresponding to n'_r, is regarded as the reference line of the net in the picture. Based on them, the constraints are explained below.

1. **Location.** The net line must be close enough to the reference line, depending on the sports type. If the line equation of the bottom reference line n_r is $y = s_r \cdot x + b$, the net line must be located within the interval $[0; +\sigma]$ around the reference line, where parameter σ controls the search area, and is changed according to the chosen sports type.

2. **Length.** We require that the net line has sufficient length. Experiments resulted in the requirement that the net line must be longer than 1/3 of the length of n_r in the image.

3. **Slope.** The orientation of the net line should be (nearly) parallel to the reference line, since line n_r is the projection line of the net on the ground plane. In other words, the net line should be selected from the lines whose slope parameter is similar to that of n_r. Suppose that only m lines satisfy the first two constraints, and the normalized length of i^{th} line is $L(i)$. The net-line likelihood parameter $\mathcal{P}(i)$ is calculated by:

$$\mathcal{P}(i) = L(i) \cdot G(s_i), \quad 0 \leq i \leq m-1. \tag{2}$$

Here, $G(s_i)$ is a normalized *Gaussian* function centered at s_r that represents the slope parameter of n_r.

In principle, the line with the highest $\mathcal{P}(i)$ should be regarded as the net line. However, two cases are often happening in practice in test images. One case is that only one net line is visible. Another possibility is that there are still two net

lines visible, as the net, in particular tennis, is actually in the middle connected to the ground with a binding cable, so that it consists of two slightly different line segments. To make our design generic, we hope that there is only one net line in the image relating to the net line in the 3-D model eventually. In other words, we need a refinement procedure to generate the final net line for the case with two line segments. To this end, we start with a judgement scheme that determines which case the current frame belongs to. More specifically, the first two lines having highest $\mathcal{P}(i)$ are acquired, and its intersection can be computed. If the intersection is nearby the center point of the reference line n_r in horizontal direction, it should be the second case. Otherwise, it must be the first case.

For the first case, the line with the highest $\mathcal{P}(i)$ is definitely the net line. For the second case, the refinement procedure is required, being explained through a practical example (see Fig. 1(b) for a reference). Suppose that two lines pass the constraints, the left one is l_1, and the right one is denoted as l_2. Meanwhile, it is possible to find two *end* points on n_r, where the left one is regarded as $\mathbf{p_s}$ and the right is $\mathbf{p_e}$. Subsequently, two projection points of $\mathbf{p}_s$ on l_1 as well as $\mathbf{p}_e$ on l_2, are computed, thus forming the final net line through connecting these two projection points. When the net-court lines are labeled, six points, where two are from net line and four are from court lines, are utilized to determine the camera parameters.

3 Applications Using 3-D Camera Modeling

3.1 Player Segmentation

We base our player segmentation on a change-detection technique, but focus more on building a high-quality background integrated by the game properties of court-net sports. We have found that in most court-net video sequences, a regular frame containing the playing field mainly includes three parts: the court (playing field), the area surrounding the court and the area of the audience. Normally, the moving area of the players is limited to the field inside the court and partially the surrounding area. Moreover, the color of the court is uniform, as is also the case for the surrounding area. The above observations motivate us to separately construct background models for the court and the area surrounding the court. This approach automatically links us to the 3-D camera modeling from the previous section, which helps to build the background model independently.

Let N be the number of pixels belonging to the court area. We use the RGB color components for background modeling. Thus, each component has a histogram, which is modeled by a *Gaussian* distribution. We consider the $H_R(u)$ as the histogram of the R component of the pixels situated within the court area. The peak of each histogram is the color value most frequently found within the corresponding component, thus, is expected to be the background color. Using a window of width $2W$ centered at the peak point of the histogram, we compute the mean and variance of the *Gaussian* distribution of that component. Note that we compute the mean and variance for each *Gaussian* distribution using a local (windowed) histogram space, rather than the whole histogram space. The advantage is that histogram noise resulting from player body and shadow sample values is partially removed.

Once obtaining this background model, we compute the background statistics of the RGB values at each pixel location, (x, y), including the mean $\mu_B(x, y)$ and the covariances $\sum_B(x, y)$ of the RGB values. Given an input image, we compute the Mahanalobis distance of the pixel value from the estimated background using the two parameters computed above. Based on the Mahanalobis distance, pixels are classified into foreground and background. Unlike our previous method [6] that uses a fixed, user-defined threshold to classify the background and foreground pixels, our new algorithm adaptively decides on the foreground and background pixels for each frame by estimating the foreground and background likelihood using the iterative expectation maximization (EM) procedure. At each pixel, indexed by k, we have a two-class problem ($Foreground = w_1, Background = w_2$) based on the observations: Mahanalobis distance, d_k. More specifically, we need to estimate the posterior $p(w_i|d_k)$ for each pixel. Given by Bayesian's rule, this posterior equals to

$$p(w_i|d_k) = \frac{p(d_k|w_i, \mu_i, \sigma_i)p(w_i)}{p(d_k)}. \tag{3}$$

Here, $p(d_k) = \sum_{i=1}^{2} p(w_i)p(d_k|w_i, \mu_i, \sigma_i))$, which is represented by Gaussian Mixture Model (GMM). In addition, $p(d_k|w_i, \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(d_k-\mu_i)^2}{2\sigma_i^2}}$. Now, the problem reduces to estimate $p(w_i)$, μ_i and σ_i, which can be iteratively estimated using the standard EM update equations. The estimates from different iterations are distinguished using the superscript:

$$p^{(n+1)}(w_i) = \frac{1}{N} \sum_{k=1}^{N} p^{(n)}(w_i|d_k), \tag{4}$$

$$\mu_i^{(n+1)} = \frac{\sum_{k=1}^{N} p^{(n)}(w_i|d_k)d_k}{\sum_{k=1}^{N} p^{(n)}(w_i|d_k)}, \tag{5}$$

$$\sigma_i^{(n+1)} = \frac{\sum_{k=1}^{N} p^{(n)}(w_i|d_k)(d_k - \mu_i)^2}{\sum_{k=1}^{N} p^{(n)}(w_i|d_k)}, \tag{6}$$

$$p^{(n+1)}(w_i|d_k) = \frac{p(d_k|w_i, \mu_i^{(n)}, \sigma_i^{(n)})p^{(n)}(w_i)}{\sum_{i=1}^{2} p(d_k|w_i, \mu_i^{(n)}, \sigma_i^{(n)})p^{(n)}(w_i)}. \tag{7}$$

The EM process is initialized by choosing class posterior labels based on the observed distance; the larger the Mahanalobis distance of a pixel, the higher the initial posterior probability of being from the foreground.

$$p^{(0)}(w_1|d_k) = min(1.0, d_k/255), \tag{8}$$

$$p^{(0)}(w_2|d_k) = 1 - p^{(0)}(w_1|d_k). \tag{9}$$

We found that, with this initialization strategy, the process stabilizes fairly quickly, within 15 or so iterations. There are several postprocessing steps on the binary map computed by EM, such as shadow area detection, small noisy region elimination [6]. Fig. 2 portrays the segmentation results.

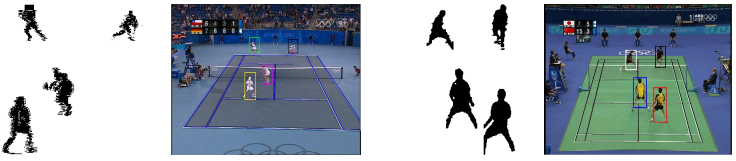


Fig. 2. Segmentation results: binary maps and corresponding player detection results

3.2 Application for 3-D Scene Adaptation

A video content-adaptation technique is essential for generating virtual games. Unlike conventional methods that insert virtual 2-D scenes like advertisements into the image [7], our adaptation technique is capable of inserting 3-D scenes into the image by means of 3-D camera modeling. Additionally, our technique enables to first design a virtual 3-D scene that is formed by a number of 2-D pictures in the 3-D model, and then mapped to the resulting picture, rather than a straightforward insertion of a virtual scene into the image. Such a supervised generation of a virtual scene can be expressed mathematically by:

$$V_i = \mathbf{M}V_m, \tag{10}$$

where V_i denotes the virtual scene in the image, V_m represents the virtual scene in the 3-D model, and $\mathbf{M}$ is a transformation matrix. Once obtaining the virtual scene and finding the suitable place for inserting in the 3-D model, it is possible to generate a virtual picture using Eqn. (10). Though this procedure sounds simple, unfortunately, generating V_i has to overcome two practical problems. First, since the parameters in $\mathbf{M}$ are not exact integers, the coordinates after transformation cannot be integer either. For an image, only integer coordinates are considered, so that we need a coordinate-rounding function. However, there are always a number of pixels in V_i that cannot find correspondences within V_m, no matter what rounding function is adopted. In other words, direct use of Eqn. (10) is unlikely to generate a perfect virtual picture. Instead, we have developed a two-step approach. In the first step, we acquire four boundary points of the inserted scene (normally, it is a rectangular), which are expressed by the corner points $p(0,0)$, $p(w,0)$, $p(0,h)$ and $p(w,h)$. Here, w means the width of the inserted scene and h is the height of the scene. When we find a suitable place for inserting this virtual scene into the 3-D model, the four points then become $\hat{p}(x_1,y_1,z_1)$, $\hat{p}(x_2,y_2,z_2)$, $\hat{p}(x_3,y_3,z_3)$ and $\hat{p}(x_4,y_4,z_4)$, accordingly. In the second step, Eqn. (1) allows to compute the correspondences of these four points in the image domain, which can be denoted as $p'(u_1,v_1)$, $p'(u_2,v_2)$, $p'(u_3,u_3)$ and $p'(u_4,v_4)$. Therefore, a bridge connecting p from the 3-D scene to p' in the 2-D picture can be established through $\hat{p}$ and $\mathbf{M}$. Next, we describe this *bridge* by a homography mapping $\hat{H}$, resulting in the transformation $p = \hat{H}p'$. Having this transformation, every pixel in the virtual-enhanced 2-D picture is able to find its correspondence in the inserted 3-D scene without any "hole". To highlight our method, we compare our proposal with the method that directly uses Eqn. (10).

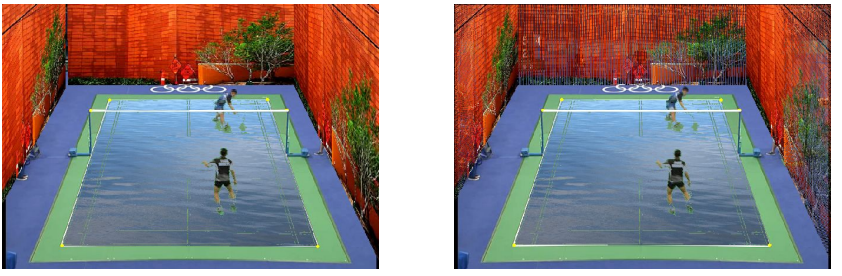


Fig. 3. Results of 3-D virtual scene generation. Left: our method. Right: the method that directly uses Eqn. (10).

The results are illustrated in Fig. 3, where the left is the virtual picture generated by our method and the right one is acquired by performing the alternative method. When looking at the walls of the virtual scene, our proposal shows no distortion of the wall textures.

The second problem that we have to deal with is the position fluctuation of the inserted scene caused by small errors in the camera parameters. These errors cannot be avoided even if the camera did not move, as it is impossible to keep the detected feature points for camera calibration unchanged frame by frame, especially in the presence of fluctuating lighting conditions. This small error does not influence the video content analysis, but it can be problematic when generating a virtual scene in the picture. The error results in a fluctuating position of the virtual picture in the resulting picture, which can be annoying. To address this problem, we employ a Double Exponential Smoothing (DES) filter to keep the camera parameters relatively stable over an interval of frames. Suppose there are n camera parameters, forming an observed camera parameter vector $\mathbf{C} = [m_1, m_2, ..., m_n]$. The DES smoothing filter is defined by

$$\begin{cases} \mathbf{S}_t^T = \alpha \cdot \mathbf{C}_t^T \ + \ (1-\alpha) \cdot (\mathbf{S}_{t-1}^T + \mathbf{B}_{t-1}^T) \ , \\ \mathbf{B}_t^T = \gamma \cdot (\mathbf{S}_t^T - \mathbf{S}_{t-1}^T) \ + \ (1-\gamma) \cdot \mathbf{B}_{t-1}^T \ , \ \text{for } t = 3, 4, ... \end{cases} \tag{11}$$

where $\mathbf{C}_t$ is the observed parameter vector in the t^{th} frame, and $\mathbf{S}_t$ refers to the vector after smoothing, $\mathbf{B}_t$ accumulates the changes of the parameter vector in the previous frames, and α and γ are two weighting parameters controlling the smoothness, which can be determined by training. The observed vectors obtained in the first frame and the second frame are used to initialize Eqn. (11).

4 Experimental Results

We tested our algorithms on six video sequences (in total more than 1 hour of video) that are extracted from regular television broadcasts. Three of them are tennis games on different court classes, two are badminton games, and one is a volleyball game. All sequences have 720×576 resolution, and 25 Hz frame rate.

Table 1. Court-net detection and camera calibration

Type	Court line	Net line	Camera calibration	Computation
badminton	98.7%	96.2%	96.2%	43 ms/frame
tennis	96.1%	95.8%	93.2%	49 ms/frame
volleyball	95.4%	91.8%	90.1%	55 ms/frame

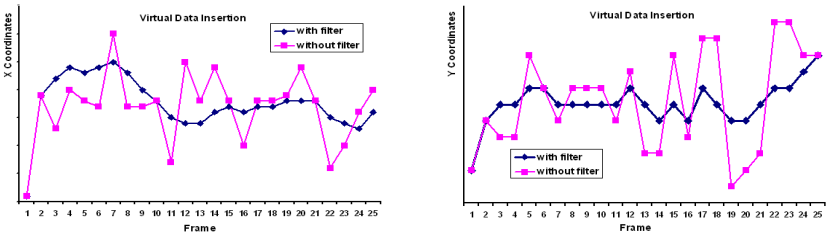


Fig. 4. The average X and Y coordinates of inserting scene in the pictures

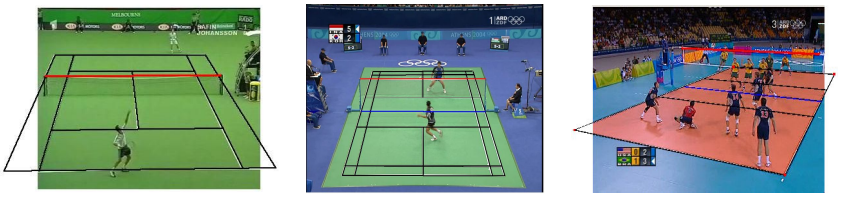


Fig. 5. Detection of the court-net lines

For our test set, the algorithm can find the courts and the net line reliably if the minimum required amount of court lines (two horizontal and two vertical lines) and the net line are clearly and fully visible in the image. Fig. 5 shows example pictures, where even crowded scenes have been selected. Table 1 shows the evaluation results of our algorithm. It can be seen that the calibration is correct for more than 90% of the sequences on the average. Note that the evaluation method requires ground-truth data, which is satisfied by the manual selection of a few *feature* points (i.e., intersection of two lines) in the 3-D model. Theoretically, the corresponding points in the image domain should also be intersections. We transform these feature points to the image domain using our calibration parameters, and compare them with the ground-truth data. The most common miscalculation in the calibration was caused by the case, where the net line is very close to one of the court lines and is mistakenly assigned to a court line. More important, our 3-D modeling algorithm is executed on a P-IV 3GHz PC, achieving a near *real-time* speed. The average computation for one frame is 43–55 ms, depending on the scene complexity of the frame.

We have implemented the two applications mentioned above. For player segmentation, we obtained a 96% accurate rate, where the criterion is that the main

body of the player is included in the detection window. For 3-D scene adaptation, an example is illustrated, in which a 3-D virtual room formed by three walls are inserted into the picture (see the left picture of Fig. 3). Since the adaptation is under the supervision of our 3-D modeling, the major contents of the video, such as court-net and players, are still preserved. Additionally, Fig. 4 demonstrates the position of a virtual scene in the picture, where the left graph draws average X coordinates of the pixels in the virtual scene on a frame-by-frame basis, and the right graph represents the average corresponding Y coordinates. The figure shows a comparison between the approach inserting a virtual scene without any filter and the same approach with our DES filter. Obviously, our proposal efficiently alleviates the position fluctuation phenomenon from Section 3.2.

5 Conclusions

We have presented a generic 3-D camera modeling for court-net sports analysis, in which both the ground plane and the perpendicular net plane are exploited. The main improvement is that we can now establish the mapping from 3-D coordinates to a 2-D image, instead of a 2-D to 2-D plane mapping. Besides this, we employ GMM model based on the backgrounds built by our 3-D camera modeling to label the foreground and background pixels, achieving a more accurate player-segmentation algorithm. Moreover, we are able to remove the fluctuation in the positioning of inserted data by using a DES filter. The performance of the system for various court-net sports exceeds 90% accuracy on camera calibration. Possible enhancements of the algorithm would be the inclusion of curved lines into the court model, which would allow calibration in cases where not enough straight lines are visible.

References

1. Sudhir, G., Lee, C., Jain, K.: Automatic classification of tennis video for high-level content-based retrieval. In: Proc. IEEE Int. Works. Cont. Based acce. of Imag. Video Data. pp. 81–90 (1998)
2. Kim, H., Hong, K.: Robust image mosaicing of soccer videos using self-calibration and line tracking. Pattern Analysis and Applications 4(1), 9–19 (2001)
3. Farin, D., Han, J., de With, P.: Fast camera-calibration for the analysis of sports sequences. In: Proc. IEEE Int. Conf. Mult. Expo (ICME) (July 2005)
4. Kim, T., Seo, Y., Hong, K.: Physics-based 3D position analysis of a soccer ball from monocular image sequence. In: Proc. ICCV, pp. 721–726 (1998)
5. Liu, Y., Liang, D., Huang, Q., Gao, W.: Extracting 3D information from broadcast soccer video. Image and Vision Computing 24, 1146–1162 (2006)
6. Han, J., Farin, D., de With, P., Lao, W.: Real-time video content analysis tool for consumer media storage system. IEEE Trans on Consumer Electronics 52(3), 870–878 (2006)
7. Xu, C., Wan, K., Bui, S., Tian, T.: Implanting virtual advistement into broadcast soccer video. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 264–271. Springer, Heidelberg (2004)

Face Recognition by Matching 2D and 3D Geodesic Distances

S. Berretti[1], A. Del Bimbo[1], P. Pala[1,⋆], and F.J. Silva Mata[2,⋆⋆]

[1] Dipartimento di Sistemi e Informatica
University of Firenze
Firenze, Italy
[2] Centro de Aplicaciones de Tecnologias de Avanzada
Havana, Cuba

Abstract. Face recognition has been addressed both in 2D, using still images or video sequences, and in 3D using three-dimensional face models. In this paper, we propose an original framework which provides a description capable to support 3D-3D face recognition as well as to directly compare 2D face images against 3D face models. This representation is extracted by measuring geodesic distances in 3D and 2D. In 3D, the geodesic distance between two points on a surface is computed as the length of the shortest path connecting the two points on the model. In 2D, the geodesic distance between two pixels is computed based on the differences of gray level intensities along the segment connecting the two pixels in the image. Experimental results are reported for 3D-3D and 2D-3D face recognition, in order to demonstrate the viability of the proposed approach.

1 Introduction

Human face identification has been addressed mainly focussing on detection and recognition of faces in 2D still images and videos [1]. Recently, three-dimensional (3D) facial data has been proposed to improve the effectiveness of face recognition systems [2]. This is motivated by the fact that solutions based on 3D face models are less sensible, if not invariant, to lighting conditions and pose variations. This opens the way to systems supporting face recognition by combining information extracted from 2D images with the information of facial characteristics derived from 3D models. Motivated by these considerations, an increasing number of face recognition approaches using 3D data has been proposed.

An interesting class of these approaches rely on multiple imaging modalities in which information extracted from 3D shapes and 2D images of the face are combined together to attain better recognition results. For example, in [3], face

⋆ This work is partially supported by the Information Society Technologies (IST) Program of the European Commission as part of the DELOS Network of Excellence on Digital Libraries (Contract G038-507618).

⋆⋆ The work was performed while the author was at the Dipartimento Sistemi e Informatica, University of Firenze, Firenze, Italy.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 444–453, 2007.
© Springer-Verlag Berlin Heidelberg 2007

recognition in videos is obtained under variations in pose and lighting by using 3D face models. In this approach, 3D models are used to capture a set of projection images taken from different points of view. Similarity between a target image and 3D models is computed by matching the image with the projection images taken from the models. Though multi-modal approaches are oriented to the combination of 2D and 3D information, the solutions proposed so far are mainly based on the separate use of 2D and 3D information. 2D and 3D matching are conducted independently, while results are combined together. In other solutions, 3D models are mostly used as an intermediate source to obtain 2D images representing particular views of the model to be used in the 2D-2D match with face images. As a particular case, in [4] faces are compared using morphable model parameters obtained by fitting a 3D morphable model to two matching face images.

In perspective, a main goal for concrete applications aiming to provide recognition/authentication, is to directly compare 2D face images of unknown subjects (*probes*), taken from one or more cameras, with the 3D face models (*gallery*) of known individuals. This requires that specific representations for 2D face images and 3D face models be extracted in a form that permits their direct comparison. A plausible way to face the problem is to use (*geodesic*) distances to model 3D and 2D face information.

In 3D, the use of distances to capture 3D facial information is directly motivated by the *face anthropometry*, the biological science dedicated to the measurement of the human face. In particular, in order to describe faces, Farkas proposed a total of 47 landmark points, and performed measurements between these points using geodesic distances, Euclidean distances and angles [5]. Until recently, the measurement process could only be carried out by experienced anthropometrists by hand, but recent works have investigated 3D scanners as an alternative to manual measurement [6].

In 2D, shading plays an important role in human perception of surface shape. Artists have long used lighting and shading to convey vivid illusions of depth in paintings. Researchers in human vision have attempted to understand and simulate the mechanisms by which our eyes and brain actually use the shading information to recover the 3D shapes. In computer vision, the idea of using the gradual variation of shading in an image to recover 3D shape, dates back to the first studies on shape-from-shading [7]. However, in these solutions the final objective is to reconstruct the 3D shape of the entire object [8]. More related to our work are researches on computing geodesic distances in 2D images. For example, in [9] geodesic sampling is used treating a 2D image as a surface embedded in a 3D space. In this framework, image intensity is weighted relative to the distance in the x-y plane of the image. It is shown as this weight increases, geodesic distances on the embedded surface are less affected by image deformations so that, in the limit, distances are deformation invariant. Geodesic distance has also been used in object recognition. For example, in [10] it is used to build bending invariant signatures for real surfaces.

In this paper, we propose an original framework to represent 2D and 3D face information by using geodesic distances. The final goal is to define a

representation which can be extracted from 2D images as well as from 3D models and directly used to compare 2D face images against 3D face models. In this way, the additional structural information associated to 3D models, can be used without any 3D to 2D projection in order to compare 3D with 2D. This is basically obtained by computing geodesic distances with respect to a reference (*fiducial*) point. In 3D, the geodesic distance of a vertex is computed as the length of the path computed on the face model surface and connecting the vertex to the fiducial point. In 2D, the geodesic distance from a pixel to the fiducial point is computed based on the differences of image gray level intensities. According to this, two are the main contributions of this work. First, we propose to use the loci of iso-geodesic distance computed from a fiducial point for capturing 3D face information. The information carried out by quantizing these loci into *iso-geodesic stripes* is then represented in a compact form by extracting the basic 3D shape of each stripe and evaluating the spatial relationships between every pair of stripes [11]. This representation can be used to effectively perform 3D-3D face recognition. Then, we propose a 2D analogous of the 3D representation, by defining the 2D geodesic distance based on image intensity of the face image. In this way, data extracted from 2D face images taken through a camera are matched to 3D information extracted from 3D face models.

The paper is organized in three Sections and Conclusions. In Sect.2, geodesic distances computed on 3D face models, given as mesh of vertices, are proposed as foundation of a 3D face recognition approach. In Sect.3, we propose the 2D equivalent for the geodesic distance computation, which is directly performed in the intensity domain of 2D face images. Based on this modeling approach, in Sect.4 we present 3D face recognition results, and show the possibility to directly compare 2D and 3D geodesic distances. Finally, conclusions and future research directions are outlined in Sect.5.

2 3D Geodesic Distances

The idea of defining smooth real valued functions on the surface of a 3D model in order to capture its characteristics dates back to the *Morse*'s theory [12]. In this theory, differential properties of the function are used to make explicit the topological properties of a surface. Depending on the choice of the function, different characteristics of the surface can be highlighted.

In our case, the function in a generic point of the model surface is defined as the geodesic distance between the point and the *pronasale* (i.e., the most protruded point of the apex nasi), that we use as fiducial point of the face. This is known to be easily detectable and stable under different expressions [13]. In the following, we will refer to this point as *nose tip*.

We assume the face surface S be approximated through a discrete mesh $\mathcal{M}$ with n vertices $v_1, \ldots, v_n$, and v_f be the fiducial vertex located at the nose tip. Every vertex v_i of the mesh is associated with an estimate $\mu_{3d}(v_i, v_f)$, of the geodesic distance between vertices v_i and v_f. Since we assume that models are approximated by polygonal meshes, computation of the geodesic distance on the piecewise planar mesh is accomplished through the *Dijkstra*'s algorithm.

This approximates the actual geodesic distance with the length of the shortest piecewise linear path on mesh vertices:

$$\mu_{3d}(v_i, v_f) = \min_P L(P(v_i, v_f)) \tag{1}$$

where $P(v_i, v_f)$ is the path from v_i to v_f, defined as an ordered sequence of adjacent vertices $(v_{p(1)}, \ldots, v_{p(k)})$, and $L(P(v_i, v_f))$ is the length of the path measured as the sum of the Euclidean distances between adjacent vertex pairs. In the presence of non-triangular meshes, or non-regular meshes, approximation of the actual geodesic distance may be inaccurate. To avoid this, meshes undergo to a pre-processing step accounting for triangularization and regularization.

Geodesic distances computed on 3D models are used for two purposes. The first objective is to prove geodesic distances computed along paths originated from the nose tip and oriented along a radial direction at a specific angle (radius), can capture salient face information, and can be directly compared to the corresponding 2D geodesic distances. To this end, given a radius, a set of equally-spaced samples is selected along the projection of the radius on the x-y plane of the 3D normal reference system (i.e., the x-axe is oriented from left to right, the y-axe from bottom to top, and the z-axe is oriented towards the observer). This permits to capture the differences occurring on the model for equal increments along the projected radial direction. Fig.1(a), shows a 3D face model where the radius originated in the nose tip and oriented along the radial direction at 0 degree is evidenced. In (b), the geodesic distance computed along this direction is shown by reporting the distance sampled at 70 equally-spaced points along the radius. It can be observed as the plot captures information on the position and the intensity of the occlusion occurring at the base of the nose.

The second objective, is to prove that geodesic distances convey sufficient information to perform face recognition directly in 3D. In particular, we propose the use of the shape and spatial arrangement of iso-geodesic stripes to represent the 3D face information.

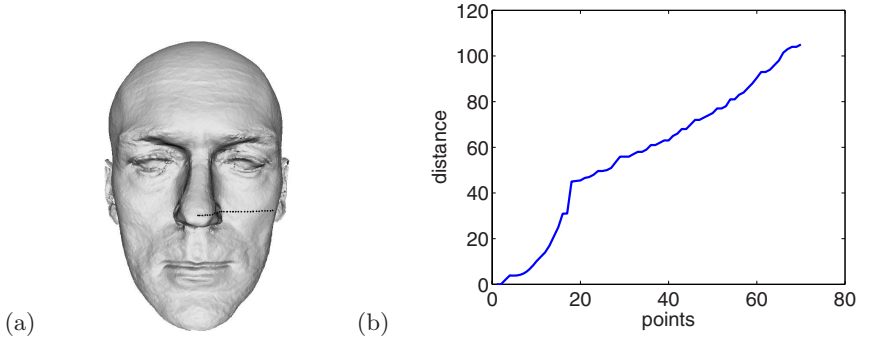


Fig. 1. (a) A 3D face model with the radius for the direction at 0 degree; (b) The geodesic distance corresponding to the radial direction shown in (a)

To this end, normalized values $\bar{\mu}_{3d}(v_i, v_f)$, of the distance are obtained dividing $\mu_{3d}(v_i, v_f)$ by the Euclidean eyes-to-nose distance. This is defined identifying the two fiducial points located at the *endocanthion* (i.e., the point at the inner commissure of the eye fissure), and considering the sum of the Euclidean distance between the two endocanthion, and the two distances between endocanthion and the nose tip. This normalization guarantees invariance of the distance with respect to scaling of the face model. Furthermore, since the Euclidean eye-to-nose distance is invariant to face expressions, this normalization does not bias values of the function under expression changes.

Once values of $\bar{\mu}_{3d}(v_i, v_f)$ are computed for every surface vertex, iso-geodesic stripes are identified by quantizing the range of $\bar{\mu}_{3d}$ values into N intervals $c_1, \ldots, c_N$. Accordingly, N stripes are identified on the model surface, the i-th stripe corresponding to the set of surface vertices for which the value of the normalized distance falls within the limits of interval c_i.

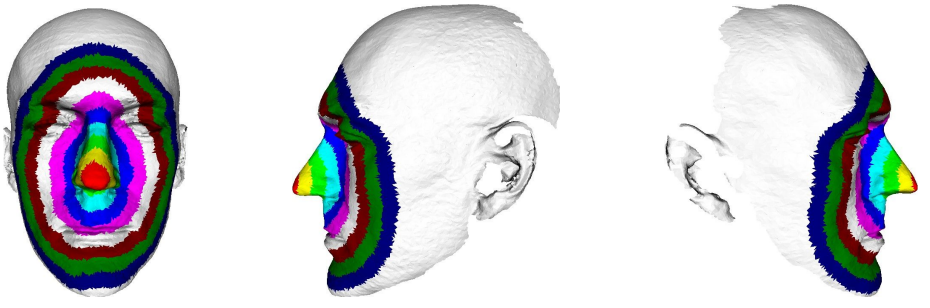


Fig. 2. Iso-geodesic stripes of the face model reported in Fig.1(a). Since stripes are surfaces in the 3D space, frontal and side views of the model are shown (different colors are used to evidence different stripes).

As an example, Fig.2 shows the first 10 stripes identified on a 3D face model. Modeling shape of iso-geodesic stripes and their spatial arrangement, makes it possible to perform face recognition directly using 3D models. More details on this modeling approach are given in [11].

3 2D Geodesic Distances

2D images of a face capture information which is related to the intensity variations represented by the values of image pixels. Our objective, is to compute geodesic distances in 2D images so that they can be compared against geodesic distances obtained from the correspondent 3D models. In doing so, we aim to find and demonstrate the existence of a relationship between these distances, so as to enable the use of this relationship as a method to verify similarity between 2D and 3D faces for recognition or authentication purposes.

According to this, in order to compute the 2D equivalent of the 3D geodesic distance, we consider the following experimental set up. 3D face models, like

that shown in Fig.1(a), are acquired with a 3D laser scanner[1]. The 640x480 color image which is also provided as output by the scanner, is used as the 2D representation of the face model. Results are reported for frontal face images, but the proposed approach can be extended to the case of different face views (e.g., side or top views). In order to effectively compare 3D and 2D geodesics, these must be computed along corresponding directions. To this end, the nose tip both in 2D and 3D should be accurately detected. To separate errors arising from a wrong detection of the nose tip, from errors related to the differences between 2D and 3D geodesic measures, the nose tip is manually detected in 2D, while in 3D it is automatically located following the approach proposed in [13].

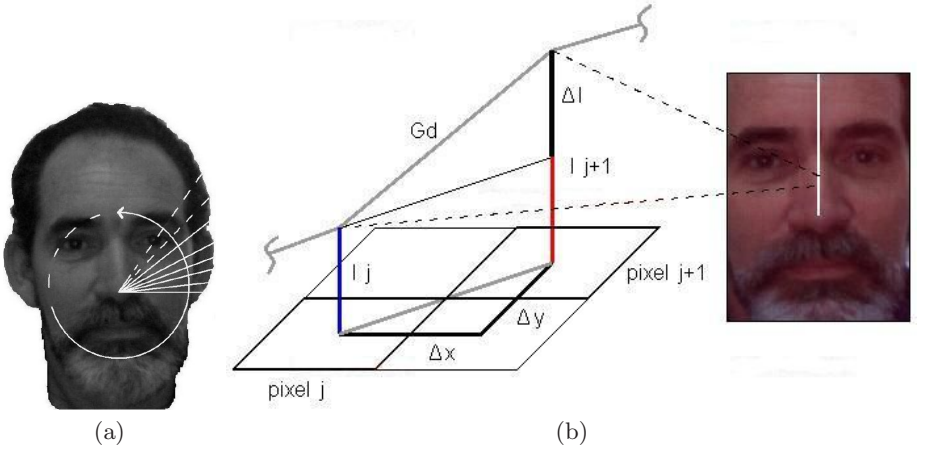


Fig. 3. (a) Multiple radii on the image; (b) The pseudo-geodesic distance computed according to the intensity values of the image

The intensity based geodesic distance is computed along radial paths from the nose tip to the border of the face, and oriented along the radial direction (see Fig.3(a)). More specifically, the geodesic distances are taken in radial trajectories going from the nose tip, both in the image and in the 3D model. In the case of 2D images, the geodesic distances (that we call *pseudo-geodesic* distances) are computed according to the intensity variations of image pixels. The following equation is used to compute the 2D geodesic distance:

$$\mu_{2d}^{j} = \sum_{i=1}^{j} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + \Delta I^{2/\max(1,w(\Delta I))}_{(x_i,y_i)(x_{i-1},y_{i-1})}} \qquad (2)$$

where: j is the index of the pixel along the radius; x_i and y_i are the pixels coordinates of the radial trajectory, and ΔI is the difference of intensities between adjacent image pixels (x_i, y_i) and (x_{i-1}, y_{i-1}). The exponent of ΔI is decremented proportionally to the value of the intensities difference in the case

[1] The Konica Minolta© Vivid VI-910.

it exceeds a given threshold. If the difference of intensity between two image pixels is zero (i.e., $\Delta I = 0$), μ^j_{2d} is reduced to the Euclidean distance between the pixels. In Fig.3(b), a visual representation of Eq.(2) is given.

4 Experimental Results

Experiments addressed the use of geodesic distances in 3D face recognition and in comparing 2D face images to 3D face models.

3D face recognition experiments, using the iso-geodesic stripes approach of Sect.2, have been carried out using the Gavab 3D face database[2]. This includes three-dimensional face models of 61 individuals (45 males and 16 females). For each person, 7 different models are taken—differing in terms of acquisition viewpoint or facial expression—resulting in 427 facial models. In particular, for each subject there are 2 frontal and 2 slightly rotated models with neutral facial expression, and 3 frontal models in which the person laughs, smiles or exhibits a random expression. Models are coded in the VRML format at one-to-one scale.

For each individual, one of the two scans with frontal view and neutral expression is used as reference model and included in the gallery. All other scans of a subject are used as probes. According to this, we conducted a set of recognition experiments using 366 probes (with neutral and non-neutral facial expression) on a gallery of 61 models. Each probe is compared against all the gallery models producing a result list of gallery models ranked in increasing order of scored distance from the probe. The effectiveness of recognition is measured as rank-1 recognition rate.

As discussed in Sect.2, geodesic distances computed on models surface are normalized, so that homologous stripes in different models represent corresponding distances from the nose tip. This permits to use the same stripe extent in the processing of every face model. A second parameter which is relevant in comparing the representations of two face models is the number of stripes used in the match. Of course, this number is strictly related to the extent of the stripes. In our setting, we found that a match comprising the first 9 stripes provides the best results for the used stripe extent of 0.08. Using this setting, rank-1 recognition rates are reported in Tab.1 for the proposed Iso-Geodesic Stripes approach (*IGS* for short), and for a 3D face matching solution which uses the Iterative Closest Point (ICP) registration algorithm [14]. Results evidence that the proposed approach is able to improve results of the ICP, both for neutral and non-neutral facial expressions.

A preliminary set of tests has been conducted to validate the correspondence between geodesic distances computed in 2D and 3D. Due to space limitation, we report only one example, illustrating the methodology of experimentation. The data used in the experiment is reported in Fig.4. In Fig.4(a), the 3D model of a subject is shown. The 3D geodesic distances from the nose tip computed on radial directions of this model, are assumed as reference in the experiment. In Fig.4(b)-(c), two images of the same subject whose 3D model is shown in

[2] Publicly available at `http://gavab.escet.urjc.es/`

Table 1. Rank-1 recognition rates for probes with neutral and non-neutral expression

	rank-1 recognition rate	
	neutral frontal	non-neutral expressions
IGS	98.36%	90.16%
ICP	70.49%	62.3%

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

Fig. 4. (a) A 3D face model; (b)-(c) Images of the individual in (a), acquired at different time and with different cameras; (d) Image of a different person

(a) are reported. These are acquired at different time, with different capture devices and illumination conditions. In Fig.4(d), the image of a different subject acquired with the same device and illumination condition of (c) is provided.

Geodesic distances have been computed both for the 3D model and for the images along 72 different radial directions (sampling angles in the range [0,355] by 5 degrees). Since 3D geodesic distances are computed on the real 3D facial data, the corresponding geodesic distances are assumed as references. The goal is to demonstrate that 2D geodesic distances computed for images of the same subject, even with different devices and illumination conditions (the images in Fig.4(b)-(c)), are more similar to the reference distances than distances computed for a different subject (the image in Fig.4(d)). This can be assimilated to an authentication scenario, where 3D models are used in the gallery, while probes are images captured by a camera in a constrained environment (e.g., frontal view, controlled illumination, etc.). In this way, the claimed identity can be verified by directly comparing 2D and 3D data.

The agreement between 2D and 3D geodesics is evaluated by an error coefficient $error^{(j,angle,view)} = \mu_{3d}^{(j,angle,view)}/\mu_{2d}^{(j,angle,view)}$, measuring the ratio between the 3D and 2D geodesic distances along a particular radius (where, in $\langle j, angle, view\rangle$, j is the index of the sample along the radius at $angle$ degrees of a particular face $view$).

The error is used to measure the correspondence between 2D and 3D geodesics. In particular, first the reference error e^{ref} originated in comparing 2D geodesics of the image acquired by the scanner and the corresponding 3D geodesics of the 3D model is computed, providing a baseline for the other error measures. Then,

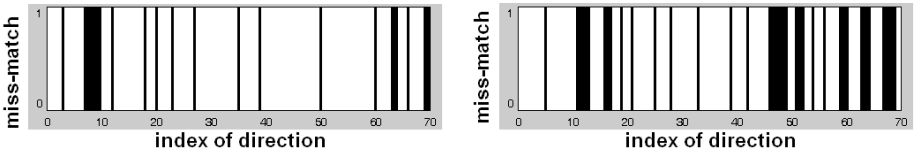


Fig. 5. Miss-matches in comparing geodesic errors: (a) $e^{c,a}$ w.r.t. e^{ref}; (b) $e^{d,a}$ w.r.t. e^{ref}. Each vertical bar indicates an error.

errors $e^{c,a}$ and $e^{d,a}$ in comparing 2D geodesics for the images in Fig.4(c)-(d), to the 3D geodesics of the 3D model in Fig.4(a) are computed. Each of these errors is obtained as the average on 72 radial directions. As an example, Fig.6 shows the error for every direction in comparing 2D geodesics of the image in Fig.4(d), to the 3D geodesics of the 3D model of Fig.4(a).

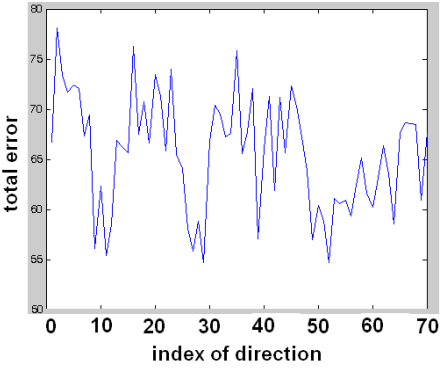


Fig. 6. Total error for every radial direction in comparing 2D geodesics of the image in Fig.4(d), to the 3D geodesics of the 3D model of Fig.4(a)

A bar plot representation is used to show the miss-match of $e^{c,a}$ and $e^{d,a}$ with respect to the reference error e^{ref}. Results are reported in Fig.5(a)-(b). In this representation, every vertical bar indicates a difference between error plots, thus representing a miss-match between the 2D image and the 3D model. According to this, plots evidence that the subject of Fig.4(c) is assimilated to the 3D model of Fig.4(a) more than the subject in Fig.4(d).

5 Conclusions

In this paper, we have proposed and experimented an approach for representing and comparing 2D and 3D face information. To this end, an original framework has been developed which provides two main contributions. First, 3D face models are described by iso-geodesic stripes that can be used for 3D face recognition.

Then, a method to represent and compare 2D and 3D faces, by computing geodesic distances from the nose tip along radial directions is proposed. The viability of the approach has been validated in a set of recognition experiments involving 3D models comparison and 2D-3D matching. Future work will address a larger scale experimentation for the 2D-3D face authentication scenario.

Acknowledgments

We would like to thank Daniele Molina, and Konica Minolta© for providing the 3D face scans shown in the paper.

References

1. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Survey 35(4), 399–458 (2003)
2. Bowyer, K., Chang, K., Flynn, P.: A survey of approaches to three dimensional face recognition. In: Proc. International Conference on Pattern Recognition, Cambridge, UK, pp. 358–361 (2004)
3. Park, U., Chen, H., Jain, A.: 3d model-assisted face recognition in video. In: Proc. Canadian Conference on Computer and Robot Vision, pp. 322–329 (2005)
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. on Pattern Analysis and Machine Intelligence 25(9), 1063–1074 (2003)
5. Farkas, L.: Anthropometry of the Head and Face. Raven Press (1994)
6. Bush, K., Antonyshyn, O.: 3-dimensional facial anthropometry using a lasersurface scannervalidation of the technique. Plastic and reconstructive surgery 98(2), 226–235 (1996)
7. Horn, B.: Understanding image intensities. Artificial Intelligence 8(2), 1250–1267 (1977)
8. Zhang, R., Tsai, P.S., Cryer, J., Shah, M.: Shape from shading: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(8), 690–706 (1999)
9. Ling, H., Jacobs, D.: Deformation invariant image matching. In: Proc. International Conference on Computer Vision. vol. II, Beijing, China pp. 1466–1473 (2005)
10. Elad(Elbaz), A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(10), 1285–1295 (2003)
11. Berretti, S., Del Bimbo, A., Pala, P.: Description and retrieval of 3d face models using iso-geodesic stripes. In: 8th ACM SIGMM International Workshop on Multimedia Information Retrieval, Santa Barbara, CA, pp. 13–22 (2006)
12. Milnor, J.: Morse Theory. Princeton University Press, Princeton, NJ (1963)
13. Chang, K., Bowyer, K., Flynn, P.: Multiple nose region matching for 3d face recognition under varying facial expression. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(6), 1695–1700 (2006)
14. Besl, P., Mc Kay, N.: A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(2), 239–256 (1992)

A Blind Watermarking Scheme Based on Visual Model for Copyright Security

Cong Jin, Liang-Gang Pan, and Ting Su

Department of Computer Science, Central China Normal University, Wuhan 430079, P.R. China
jincong@mail.ccnu.edu.cn

Abstract. A novel blind watermarking scheme is developed based on discrete wavelet transform (DWT) in this paper. In order to make the watermarking imperceptibility and robustness, the watermarking is embedded in the average of wavelet blocks using the visual model based on the human visual system (HVS). Low-pass wavelet coefficients *n* least significant bits (LSBs) are adjusted in concert with the average. Simulation results show that the proposed scheme is imperceptibility and robustness against many attacks such as JPEG compression, adding noise, rescaling, cropping, rotation, filtering, etc.

1 Introduction

With the explosive growth of the Internet in recent years, a number of information servers are available for people to access various multimedia contents, such as digital images, video and audio. As digital data can easily be copied and distributed, security of the copyright of multimedia data becomes one of the most important topics in the Internet world. As a potential and effective way to solve this problem, digital watermarking becomes a very active research area of signal and information processing. Digital watermarking is a signal embedded in a digital image or video sequence. In general, it is required that the embedded watermarking should be not only transparent to human observers, but also robust enough so that it cannot be destroyed or removed after some processing or attacks.

Many watermarking algorithms have been proposed to address this issue of copyright security. Generally, the watermarking embedding process for digital images can be accomplished in either the spatial or frequency domain. A better compromise between robustness and transparency can be obtained using the frequency domain. Frequency domain watermarking techniques could be based on spatially local or global transforms. In the published literature, a famous frequency domain scheme for digital watermarking is proposed by Cox *et al.*[1]. In recent years, with the development of JPEG2000 and MPEG-4, the application of DCT is limited due to the fact that the wavelet transform is playing an important role in JPEG2000 and MPEG-4. Therefore, the wavelet transform is one of the promising techniques for digital watermarking [2-5].

In this paper, we propose a new blind watermarking scheme based on DWT. The watermarking signal is embedded in low-pass wavelet coefficients *n* LSBs. Unlike

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 454–463, 2007.
© Springer-Verlag Berlin Heidelberg 2007

most watermarking schemes, our watermarking is not embedded by modulating individual wavelet coefficient but by modulating the average of coefficients in the wavelet blocks. Watermarking detection is blind. Simulation results show that the proposed scheme achieves a good compromise between the robustness and imperceptibility.

2 Watermarking Embedding

Given a set of wavelet coefficients, it has been observed that the average has a smaller change than individual coefficient. Thus, the watermarking embedded in the average of the wavelet blocks is more robust than the individual coefficient. In this paper, we choose the low-pass wavelet coefficient as the carrier of watermarking signal.

The low-pass wavelet coefficient matrix *Ca* can be obtained by 2D wavelet decomposition and it can be divided into wavelet blocks with size *M*×*N* and $I_i(k)$ is the *i*th wavelet coefficient in the *k*th wavelet block. Naturally, *M*×*N* is the number of coefficients in a wavelet block and variable $i \in [1, M \times N]$ .The n LSBs of $I_i(k)$ is defined as

$$\hat{I}_i(k) = \mathrm{mod}\left(I_i(k), 2^n \right) \tag{1}$$

Where, $n \in \{3, 4, 5\}$ is for the sake of tradeoff between the imperceptibility and robustness. The average of the wavelet block is defined as follows:

$$Average(k) = \frac{\sum_{i=1}^{M \times N} \hat{I}_i(k)}{M \times N} \tag{2}$$

If a few of $I_i(k)$ are changed by Ω due to some of distortions, the average of the wavelet block will only have a small change.

Supposing that $I_i'(k)$ is the *i*th wavelet coefficient in the *k*th wavelet block after embedding watermarking, $\hat{I}_i'(k)$ is the n LSBs of $I_i'(k)$ and $Average'(k)$ is the average of $\hat{I}_i'(k)$ in the *k*th wavelet block accordingly. The watermarking $W(k) \in \{-1, 1\}$ is consisted by a binary pseudo random sequence, and is embedded by adjusting the average of wavelet blocks in follow way

$$Average'(k) \in \begin{cases} [0, 2^{n-1}), & if\ W(k) = -1 \\ [2^{n-1}, 2^n), & if\ W(k) = 1 \end{cases} \tag{3}$$

In order to further enhance the robustness of the watermarking, we expect $Average'(k)$ could more trends to 2^{n-2} or $2^n - 2^{n-2}$. As show in Fig.1, there is maximal distance between integers 2^{n-2} and $2^n - 2^{n-2}$.

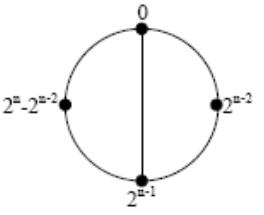


Fig. 1. The distance of *n* bit binary number in circle of integer mod 2^n

Before adjusting $Average(k)$, there are three pending questions. First, whether or not each individual coefficient *n* LSBs need adjust, if it is not, how to mark them; Second, how to control embedded strength if $Average(k)$ has satisfied Eq.3 before doing any adjustment; Third, how to make use of visual model when adjust individual coefficient *n* LSBs. These questions will be discussed in follow orderly.

In early work on DCT based watermarks, each wavelet coefficient is adjusted to make the average equal to 2^{n-2} or $2^n - 2^{n-2}$. But in proposed scheme, as show in Eq.3, the average trends to 2^{n-2} or $2^n - 2^{n-2}$; at the same time, some, but not each $\hat{I}_i(k)$ are chosen. Hence, a flag function must be set to decide $\hat{I}_i(k)$ whether need adjust or not. The flag rule is defined as follows:

$$F_i(k) = sign\left(\left(2^{n-1} - \hat{I}_i(k)\right) \times W(k)\right) \tag{4}$$

Where, $sign(x) = \begin{cases} 1, & if \quad x \geq 0 \\ 0, & if \quad x < 0 \end{cases}$. Detail result of flag function is showed in Table 1.

Table 1. The detailed results of $F_i(K)$ and $S(K)$

$W(k)$	$2^{n-1} - \hat{I}_i(k)$	2^{n-1}-$Average(k)$	$F_i(K)$	$S(K)$
-1	>0	>0	0	0
-1	≤ 0	≤ 0	1	1
1	>0	>0	1	1
1	≤ 0	≤ 0	0	0

It is possible that $Average(k)$ has satisfied Eq.3 before doing any adjustment. In this case, no adjust should be needed in these wavelet blocks. Considering the robustness of the watermarking, a very light adjustment is still indispensable. Contrarily, if $Average(k)$ is not satisfied Eq.3, a relative strong adjustment is required. So an estimation between $2^{n-1} - Average(k)$ and $W(k)$ to determine adaptable strength is necessary when embedding watermarking. The strength function $S(k)$ is defined as

$$S(k) = sign(X(k)) \tag{5}$$

Where $X(k)=\left(2^{n-1}-Average(k)\right)\times W(k)$. Detail result of strength function is also showed in Table 1.

To adapt the watermarking sequence to the local properties of the wavelet block, we use the visual model based on HVS in the watermarking system. This visual model takes into account the brightness sensitivity and texture sensitivity of the wavelet block to noise. The visual model function $Vm(k)$ is defined as follows:

$$Vm(k)=brightness(k)\times texture(k)^{\beta} \tag{6}$$

Where,

$$brightness(k)=\frac{\sum_{i=1}^{M\times N} I_i(k)}{M\times N}, \text{ and } texture(k)=\frac{\sum_{i=1}^{M\times N}\left[brightness(k)-I_i(k)\right]^2}{M\times N} \tag{7}$$

Where β control the degree of texture sensitivity. This visual model function indicates that the human eye is less sensitive to noise in the highly bright and the highly textured image areas. Hence, the wavelet blocks are divided into two parts depend on the value of $Vm(k)$: high activity wavelet block and low activity wavelet block. For simplicity, the threshold Tc is set to the average of $Vm(k)$.The following function can be applied to distinguish high or low activity wavelet block.

$$T(k)=sign\left(Vm(k)-Tc\right) \tag{8}$$

Considering the tradeoff between robustness and imperceptibility, the proposed watermark embedding algorithm can be formulated as follows:

$$\hat{I}_i'(k)=\hat{I}_i(k)+\alpha W(k)F_i(k)\left[2^{n-2-S(k)}+T(k)\times 2^{n-3}\right] \tag{9}$$

Where, α is a scaling factor to control the strength of the inserted watermarking.

Based on the above discussion, the n LSBs of wavelet coefficients have been adjusted by using the Eq.9. Naturally, their average has been updated depending on the requirement of $W(k)$. In the other word, the watermark has been embedded.

3 Watermarking Extraction and Detection

From the process of the watermarking embedding, we can obtain the watermarked objects by applying the function of Eq.3. Thus, for a given watermarked object, the watermarking can be extracted as

$$W'(k)=\begin{cases} -1, & if\ Average(k)\in\left[0,\ 2^{n-1}\right) \\ 1, & if\ Average(k)\in\left[2^{n-1},2^{n}\right) \end{cases} \tag{10}$$

Watermarking detection is based on the classical detection theory. In order to detect the watermarking W' from watermarked object, we firstly evaluate the detector response (or similarity of W' and W) as

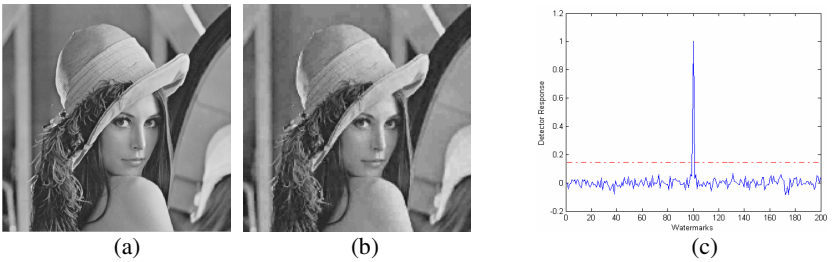
$$\rho(W', W) = \frac{\sum_{k=1}^{L} W'(k) \times W(k)}{\sum_{k=1}^{L} \|W'(k)\|^2} = \frac{\sum_{k=1}^{L} W'(k) \times W(k)}{L} \tag{11}$$

Where L is the length of the watermarking sequence (the number of wavelet block) and $W'(k) \in \{-1, 1\}$. The T_ρ is detection threshold. If $\rho \geq T_\rho$, the watermarking is present, otherwise absent.

4 Experiments Results

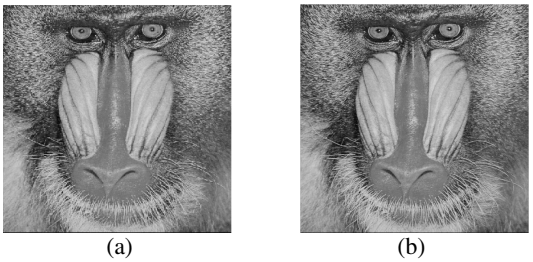
In experiments, two standard images '*Lena*' and '*Baboon*' with size 512×512 are used. The parameters are presented as follows: the threshold $n = 5$, $T_\rho = 0.15$, $\beta = 0.318$, wavelet-level=2, wavelet-type='*db*1' and scaling factor $\alpha \in [1.25, 1.73]$. We choose $\alpha = 1.45$ in order to compare expediently. For simplicity, we let $M = N = 4$. On this occasion, there are 1024 bits can be embedded in an 512×512 image, *i.e.*, $L = 1024$. In order to test the performance of the proposed scheme, 200 watermarks were randomly generated.

For *Lena* image, the PSNR result between the original object and the watermarked object is 42.4003 dB. As shown in Fig.2, the watermarking is imperceptibility and the



(a) (b) (c)

Fig. 2. (a) Original image *Lena*. (b) Watermarked object *Lena* (PSNR= 42.4003 dB). (c) Detector response of the watermarked object *Lena* for 200 randomly generated watermarking.



(a) (b)

Fig. 3. (a) Original object *baboon*. (b) Watermarked object *baboon* (PSNR= 41.6149 dB).

Table 2. PSNR between original image and watermarked object

Image	*Lena*	*Baboon*	*Man*	*Couple*	*Goldhill*	*Boat*
PSNR	42.4003	41.6149	40.9172	42.7490	41.9125	42.8583
Response	1.0	1.0	1.0	1.0	1.0	1.0

watermarked object appears visually identical to the object without watermarking. Fig.2(c) shows the response of the watermarking detector to 200 randomly generated watermarks of which only one matches the watermarking present in Fig.2.(b). The response to the correct watermarking is much higher than the responses to incorrect watermarking. The experiment result for *baboon* image is showed in Fig.3. Additional experiment results for adding images are listed in Table 2.

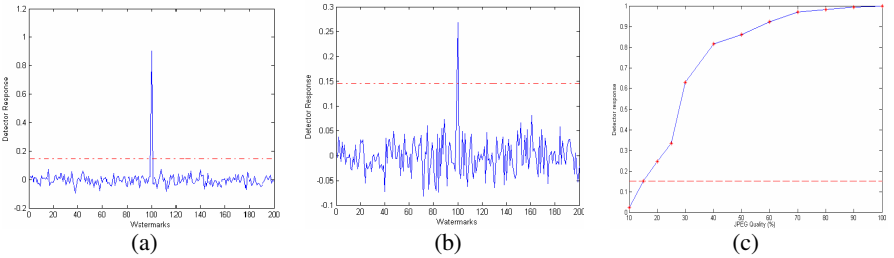
To evaluate the robustness of proposed scheme against unintentional and intentional attacks, we test the watermarked object with JPEG compression, adding noise, filtering, rescaling, cropping and rotation.

4.1 JPEG Compression

JPEG is a widely used compression format and the watermark should be resistant to it. As shown in Fig.5(c), with the decreasing of the quality of the JPEG compressed object, the response of the watermarking detector also decreases. We have found that the proposed watermarking can survive even when quality factor is as low as 20% (see Fig.5(b)), although the object is visibly distorted (see Fig.4(b)).

(a) (b)

Fig. 4. JPEG compressed of the watermarked object *Lena*. (a) 60% quality. (b) 20% quality.



(a) (b) (c)

Fig. 5. Detector response to JPEG compression of the watermarked object *Lena*. (a) 60% quality, (b) 20% quality, and (c) Watermarking detector response on the JPEG compressed of watermarked object *Lena*.

4.2 Adding Noise

Noise is one of common distorts in the image processing. In the experiment, we add 15% Salt & Pepper noise and 0.1% Gaussian noise into the watermarked object as shown in Fig.6. The watermarking can be retrieved successfully, and the responses of the watermarking detector are 0.3591 and 0.1536, respectively.

(a) (b)

Fig. 6. Watermarked object *Lena* after noise adding. (a) 15% Salt & Pepper noise adding. (b) 0.1% Gaussian noise adding.

4.3 Filtering

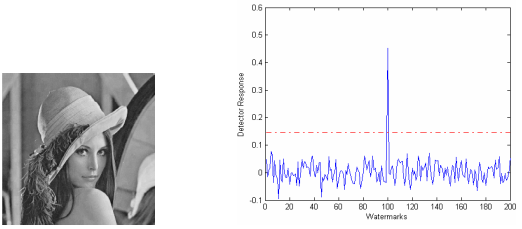
Filtering is also one of the common image processing. The watermarked object was filtered with 3×3 blur filter and 5×5 median filter (see Fig.7). The responses of the watermark detector are 0.9017 and 0.6428 respectively. These responses are well above the threshold Tc , even if the objects appear degraded.

(a) (b)

Fig. 7. Watermarked object *Lena* after filtering. (a) 3×3 blur filtering. (b) 5×5 median filtering.

4.4 Rescaling

Scaling is also very easy to perform during the editing of digital images. So the watermarking technique must be robust to the scaling attack. We test our scheme in the case of scaling the watermarked object by 0.5×0.5. The experiment results show the watermarking can still be retrieved as shown in Fig.8 with the detector response 0.4802.

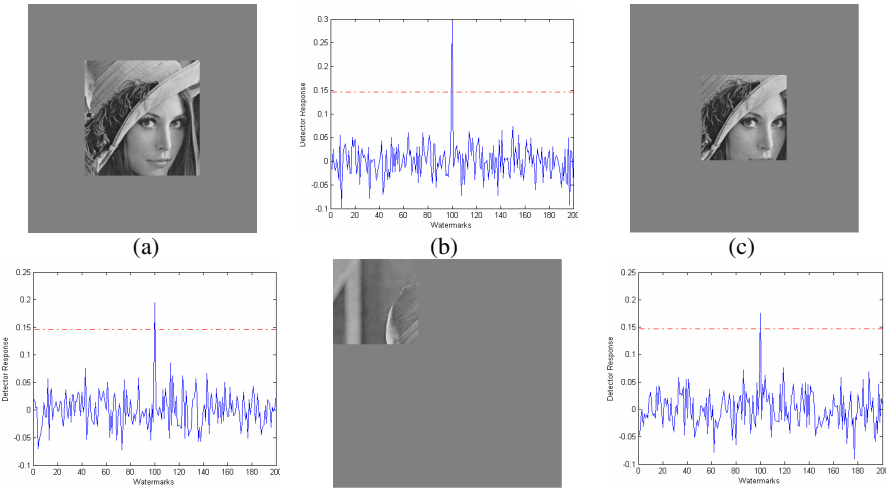


(a) (b)

Fig. 8. Watermarked object *Lena* after rescaling. (a) 0.5×0.5 rescaling. (b) Detector response of rescaling.

4.5 Cropping

Cropping is very common distort during the editing operation of digital images. In the experiment, Fig.9(a) illustrates the undisturbed watermarked image is cropped to 1/4 of the original size and 1/8 of the original size in different way [Fig.9(c) and Fig.9(e)]. The detector response for these cropping are 0.2949, 0.1954 and 0.1736 respectively as show in Fig.9(b), Fig.9(d) and Fig.9(f).



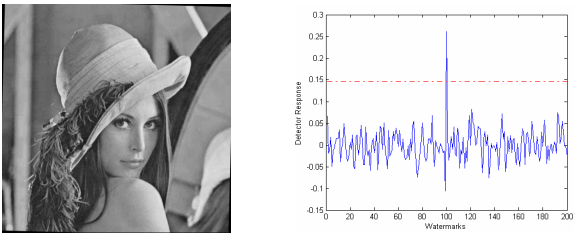
(a) (b) (c)

(d) (e) (f)

Fig. 9. Watermarked object *Lena* after cropping. (a) 1/4 of the original size. (c) and (e) 1/8 of the original size (b), (d) and (f) various detector response of cropping.

4.6 Rotation

Rotation is a widely used editing operation of digital images and the watermark should be resistant to this distortion. We test the scheme in the case of rotation the



(a) (b)

Fig. 10. Watermarked object *Lena* after rotation. (a) Rotate -3 degree. (b) Detector response of rotation.

watermarked object-3 degree. The experiment results show the watermark can still be retrieved as shown in Fig.10 with the detector response 0.2637.

4.7 Comparison to Other Schemes

The proposed scheme was compared to the Podilchuk's scheme[6], the results shown in Table 3. The proposed scheme was found to be superior to Podilchuk's scheme in PNSR.

Table 3. PSNR of the proposed scheme and Podilchulk's scheme

Image \ Algorithm	Proposed scheme	Podilchulk's scheme
Lena	42.4003	32.13
Baboon	41.6149	39.54
Goldhill	41.9125	37.51
Man	40.9172	36.96

5 Conclusions

We described a new blind digital watermarking scheme using DWT in this paper. To make the watermarking robust and imperceptibility, we embed it in the average of the low-pass wavelet blocks using visual mode. The proposed scheme has several advantages as follows:

(1) Given a set of wavelet coefficients, it has been observed that the average has a smaller change than that of individual coefficient. Thus, unlike most watermarking schemes, the watermarking is not embedded by adjust individual wavelet coefficient but by modulating the average of the wavelet blocks.

(2) Visual model is employed to achieve the best tradeoff between imperceptibility and robustness.

(3) Even though the attacker knows that DWT is used and understands the algorithm of watermarking embedding and detection, all data can be used as a private key, and encrypted with the user ID so that they are unknown to the public.

(4) Watermark detection is accomplished without the original.

Robustness of the watermarking is tested against image compression and common geometrical attacks. By the way, the proposed scheme can be easily extended to video watermarking using 3D wavelet transforms.

References

1. Cox, I.J., Kilian, J., Leighton, T., Shamoon, T.: Secure spread spectrum watermark for multimedia. NEC Research Institute Tech. Rep, pp. 95–100 (1995)
2. Dugad, R., Ratakonda, K., Ahuja, N.: A new wavelet-based scheme for watermark image. In: Proceedings ICIP, pp. 419–423 (1998)
3. Kong, X., Liu, Y., Liu, H., Yang, D.: Object watermarks for digital images and video. Image and Vision Computing 22, 583–595 (2004)
4. Yu, G., Lu, C., Liao, H., Sheu, J.: Average quantization blind watermark for image authentication. In: Proceedings of ICIP, Vancouver, Canada, vol. III, pp. 706–709 (2000)
5. Barni, M., Bartolini, F., Furon, T.: A general framework for robust watermarking security. Signal Processing 83(10), 2069–2084 (2003)
6. Podilchuk, C.I., Zeng, W.: Image-adaptive watermarking using visual models. IEEE Journal on Selected Areas in Communications 16(4), 525–539 (1998)

Fast Mode Decision by Exploiting Spatio-temporal Correlation in H.264*

Sung-Hoon Jeon, Tak-Gi Lee, Kwang-Mu Shin, Sung-Min Kim, and Ki-Dong Chung

Dept. of Computer Engineering, Pusan National Univ., Jangjeon-dong, Geumjeong-gu, Busan, 609-735, Korea
{cleanjun, zero}@melon.cs.pusan.ac.kr,
{shin, morethannow, kdchung}@pusan.ac.kr

Abstract. The H.264 video coding standard provides considerably higher coding efficiency than those of previous standards but its complexity is significantly increased. In this paper, we propose an efficient method of fast mode decision by exploiting spatio-temporal correlation in H.264. Firstly, we select skip mode or inter mode by considering the temporal correlation. Secondly, we select variable block size on inter mode by considering the spatial correlation. Simulations show that the proposed method reduces the encoding time by 71% on average without any significant PSNR losses.

Keyword: skip mode, inter mode, spatial and temporal correlation, H.264/AVC.

1 Introduction

The multimedia services have been evolved rapidly from the end of the 20th century. We can have various services including the broadcasting through the mobile communication. As a result, the need of a mobile video communication has been increased. Therefore, video compression standards which have high compression efficiency and reliability are required in multimedia service. The H.264/MPEG-4 AVC (H.264) was designed to meet all these needs. The latest video-coding standard, known as H.264, has been developed collaboratively by the Joint Video Team of ISO/IEC MPEG and ITU-T VCEG. At equivalent perceptual quality, AVC provides about 50% bit-rate savings relative to prior standards such as MPEG-4 [1]. It provides high compression efficiency compared to previous video coding standards, such as MPEG-4 and H.263, mainly due to variable block-size macroblock modes, weighted prediction and multiple reference frames motion compensation. However, its high complexity causes a large amount of encoding time [2]. Its complexity is too high to be widely applied in real-time applications. So, the goal of this paper is to reduce the complexity of H.264 encoder.

H.264 allows blocks of variable sizes and shapes. To be more specific, seven modes of different sizes and shapes, i.e. 16x16, 16x8, 8x16, 8x8, 8x4, 4x8 and 4x4,

* This work was supported by the Brain Korea 21 Project in 2006.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 464–473, 2007.
© Springer-Verlag Berlin Heidelberg 2007

are supported in H.264, as shown in Fig. 1 [3]. So, in P frame coding, a macroblock can be coded in the modes of SKIP, 16x16, 16x8, 8x16, 8x8, 8x4, 4x8 and 4x4. Among these modes, the SKIP mode represents the case where the block size is 16x16, and no information about motion and residual has to be coded. So it has the lowest computational complexity. In this paper, we propose an algorithm that can efficiently reduce the encoding time by selecting macroblock mode based on such characteristics mentioned above.

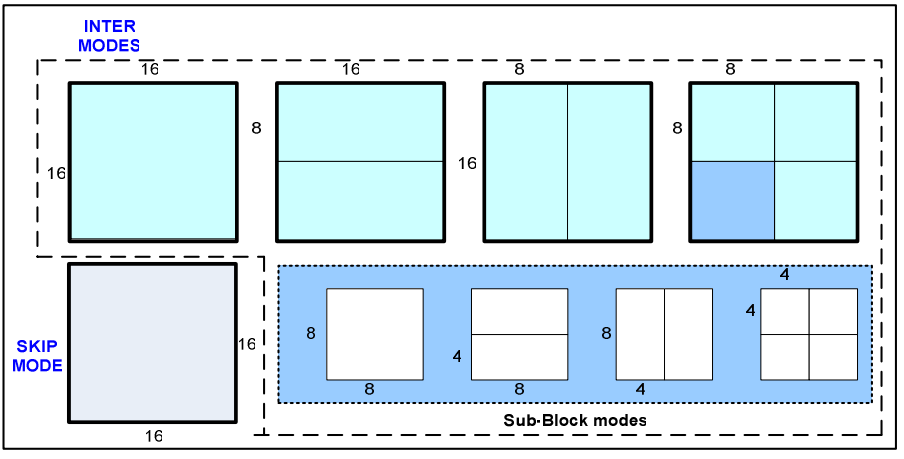


Fig. 1. Block modes for MB in P slice in H.264

Based on the previous studies, we propose a mode selection method considering temporal correlation and spatial correlation. The rest of this paper is organized as follows. Section 2 gives a brief overview of the related studies. Section 3 describes the proposed fast mode decision method. Section 4 shows the experimental results. The concluding remarks are given in section 5.

2 The Related Studies

H.264 is encoded in macroblock units. In P frame coding, a macroblock can be coded in eight modes. Among these modes, sophisticated computation process is necessary for selecting the best mode. Hence, this makes it difficult to be used in practical applications especially in real time environment [4].

Shen Gao proposed an algorithm using the frequency of use in each mode to reduce mode selection process, as shown in table 1 [5]. In column 8x8, the first number represents the percentage of all four sub-macroblock modes being used in sequence, and the number in bracket represents the percentage of the mode 8x8 being used in all sub-macroblock modes, in column INTRA, the number represents the percentage of all intra modes being used in P frames.

Table 1. Percentage of mode distribution (%)

	SKIP	16x16	16x8	8x16	8x8	INTRA
Container	82.7	8.3	3.3	2.7	3(63.5)	0.0
Claire	79.6	9.8	3.2	3.2	4.2(63.8)	0.0
Akiyo	83.0	6.2	3.0	3.6	4.2(62.3)	0.0
Highway	51.0	24.7	8.6	5.9	8.3(60.1)	1.5
News	76.3	7.4	3.5	4.5	8.2(58.4)	0.1
Stefan	25.5	32.1	10.9	9.8	19.5(51.4)	2.2
Salesman	79.0	5.1	3.2	3.6	9.1(55.2)	0.0
Silent	65.1	12.1	4.9	6.3	10.4(55.0)	1.2

The authors in [5], got some very useful information from table 1 : 1) about 70% of the macroblocks are encoded with SKIP mode, especially for sequences like video communication scenes. So this indicates that if the decision on SKIP mode can be made at the beginning, the encoding time can be saved dramatically; 2) in P frames, the probability of intra modes being used is very low. Skipping intra modes in P frames will not decrease the coding efficiency; 3) The probability of using small macroblock partitions of size 8x4, 4x8 and 4x4 is low. But for communication application, 8x8 block size is small enough to represent the details of the motion, in motion estimation/compensation.

In [6], due to the inherent spatial correlation within a single frame, the direction of the boundary in current MB can be predicted by those of the neighboring MBs.

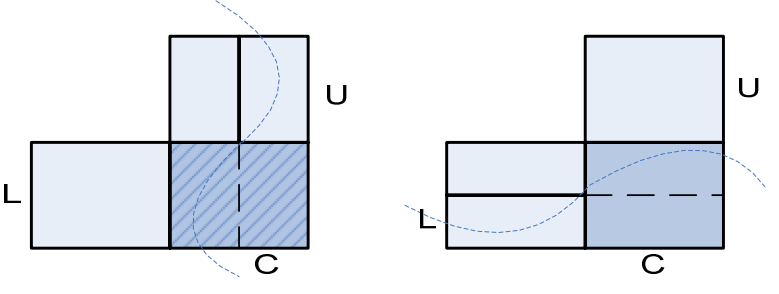


Fig. 2. The correlations of the boundary direction in the neighboring MBs

As illustrated in Fig. 2, if the boundary crosses through the upward MB in vertical direction, the direction of the boundary in the current MB is likely to be vertical too. For the same reason, if the optimum mode of the left MB is 16x8, which means that there is a horizontal boundary in the left MB, the optimum mode of the current MB may be 16x8 too with high possibility. The 4 cases and the corresponding Select modes are listed in the Table 3.

Table 2. Modes by block size

Mode	Block size
Mode 1	16x16
Mode 2	16x8
Mode 3	8x16
Mode 4	8x8

Table 3. Block size determination based on the mode of neighboring MBs

MB	flag			
Up	0	0	1	1
Left	0	1	0	1
Select mode	Mode 1	Mode 2	Mode 3	Mode 4

In [6], the coding modes are divided into four mode groups as shown in Table 2. From table 3, MB refers to the current MB. A flag is used to indicate whether the mode of a neighboring MB has influence on the mode of the current MB. If the mode of Up is 8x16 or 8x8, its flag is set to 1. Otherwise, it is set to 0. Obviously, there are 4 cases for the combinations of the up flag and left flag.

3 Mode Dicision Method Based on Spatial-temporal Correlation by Encoding Type

From table 1, encoding mode is selected only in the order of frequency in use of mode. So, the accuracy of the mode selection will be decreasing gradually. In this paper, mode is selected by using temporal correlation and spatial correlation in a frame.

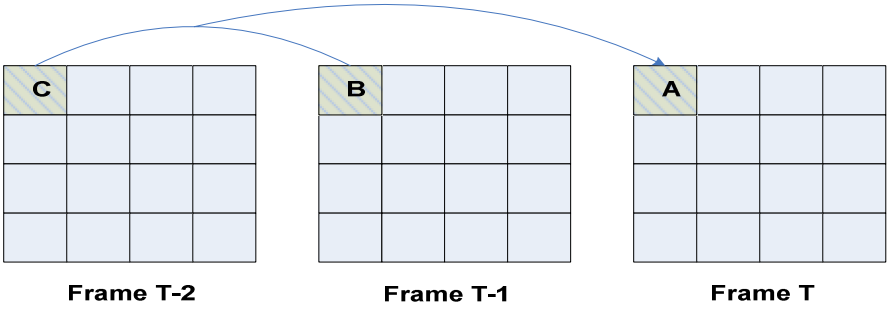


Fig. 3. Encoding mode correlation among frames

Table 4. The environment for temporal mode probability analysis

Reference software	JM 10.2
Frame	50
GOP structure	IPPP..P
Format	QCIF
sequences	Akiyo, carphone, clarie, football, foreman, grandma, miss_am, salesman

Table 5. Probilities of temporal average modes

T-2	T-1	T	Encoding probility(%)
SKIP MODE	SKIP MODE	SKIP MODE	48.7
SKIP MODE	SKIP MODE	INTER MODE	3.6
SKIP MODE	INTER MODE	SKIP MODE	4.1
SKIP MODE	INTER MODE	INTER MODE	3.7
INTER MODE	SKIP MODE	SKIP MODE	3.4
INTER MODE	SKIP MODE	INTER MODE	4.1
INTER MODE	INTER MODE	SKIP MODE	3.6
INTER MODE	INTER MODE	INTER MODE	28.8

As illustrated in Fig. 3, the current frame is labeled as Frame T and Frame T-1 is the encoded frame before Frame T, and it's also the case of Frame T-2. The modes of macroblock A, B, C have temporal correlations. And two previously encoded frames from the current frame have much influence on the current frame. So, we decide that two reference frames for current frame encoding are enough.

The environment for temporal mode probability analysis is shown in table 4. From table 5, we can get useful information. When previous macroblock T-2 and T-1 are encoded as SKIP mode, the current macroblock has high probability to become the SKIP mode. As for the last line, the mode of the macroblock A in the current frame will be selected based on probability. However, if both frame T-2 and frame T-1 were selected within the same mode, the subsequent mode will always be selected as the same mode only by probability.

From table 6, we show that the mode is not selected only by probability, where ? is a mark which shows that the mode cannot be selected only by the probability. Accordingly, due to the defect in the algorithm, it needs a method for efficient mode

Table 6. The mode selection according to temporal correlation

T-2	T-2	T
SKIP MODE	SKIP MODE	?
SKIP MODE	SKIP MODE	?
SKIP MODE	INTER MODE	?
SKIP MODE	INTER MODE	?
INTER MODE	SKIP MODE	?
INTER MODE	SKIP MODE	?
INTER MODE	INTER MODE	?
INTER MODE	INTER MODE	?

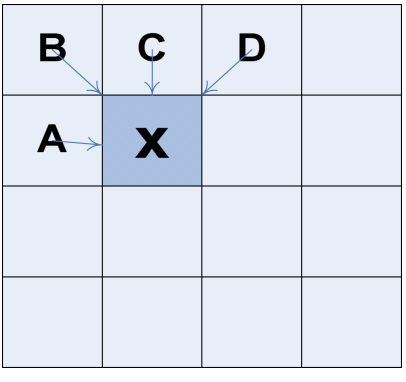


Fig. 4. The information of the neighboring MBs

selection. The method selects a mode according to temporal correlation and spatial correlation.

As illustrated in Fig. 4, when encoding the current macroblock, X selects the best mode after searching neighboring macroblocks (which are A, B, C, D) of current frame. The criterion of selection of the current macroblock X is shown in Table 7. $\alpha(i)$ presents distribution degree of i in neighboring macroblock for the criteria of judgment. It should be noted that $\alpha(0)$ here represents the distribution degree of 0 in neighboring macroblock. 0 is the SKIP MODE and 1 is the INTER MODE. The distribution degree presents the number of 0 or 1 around X.

We applied the probability between frame T-2 and frame T-1. The current frame mode is selected after looking for the number of neighboring 0(or 1) unit. In this

Table 7. The selection of mode according to neighboring macroblock information

T-2	T-1	Criteria of judgment	Selected mode
SKIP MODE	SKIP MODE	$\alpha(0) \geq 1$	SKIP MODE
		The others	INTER MODE
SKIP MODE	INTER MODE	$\alpha(0) \geq 2$	SKIP MODE
		The others	INTER MODE
INTER MODE	SKIP MODE	$\alpha(1) \geq 2$	INTER MODE
		The others	SKIP MODE
INTER MODE	INTER MODE	$\alpha(1) \geq 1$	INTER MODE
		The others	SKIP MODE

scheme, the reference modes of previous frames are used to determine the best mode of the current frame. For example, if the frame T-2 is the SKIP MODE and the frame T-1 is also the SKIP MODE, and if the number of surrounding 0s is bigger than 1 ($\alpha(0) \geq 1$), we select the SKIP MODE, else INTER MODE.

After the selection of mode according to neighboring macroblock information, if the INTER MODE is selected, the variable block size will be selected due to the method shown in paper [6].

4 Experimental Results

In the experiment, we have used eight video sequences. Each sequence has 30 frames and the Simulation environments are shown in table 8. The Fig. 5 shows the hit ratio of the proposed method and the reference software.

The sequence I in Fig. 5, presents the average hit ratio. The Hit ratio presents the level of equivalence between the mode from the proposed best mode selection algorithm and the mode selected by JM 10.2, as show in Fig. 5. Clarie and Football contain a little motion activities and drastic motion activities, respectively. Over 90% of selected modes were same between two sequences. The average hit ratio is a little over 86%.

The proposed algorithm was implemented by modifying the H.264 codec JM10.2 [7]. And the performances of the fast mode decision method are tested using the first 50 frames of 6 testing video sequences (Akiyo, Carphone, Clarie, Grandma, Miss_am, salesman). The frame rate was 30fps, and the frame coding structure was IPPP. Here the QP parameters in H.264/AVC are fixed as 28 and 36. The experiments was tested under 3.0GHz PC with 1GB memory.

Table 8. Simulation environments

Reference software	JM 10.2
CPU	P-IV 3.0
Memory	1GB
Frames	50
GOP structure	IPPP..P
Format	QCIF
sequences	Akiyo(A), carphone(B), clarie(C), football(D), foreman(E), grandma(F), miss_am(G), salesman(H)

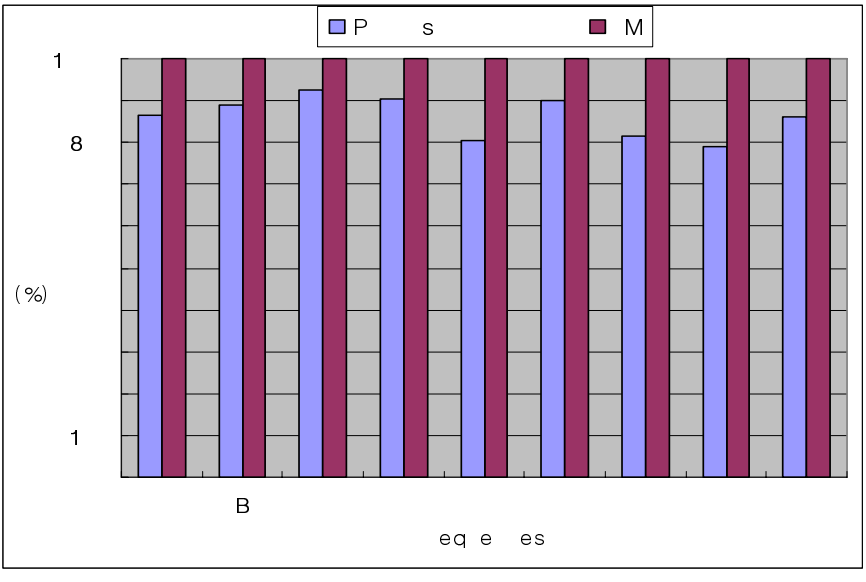


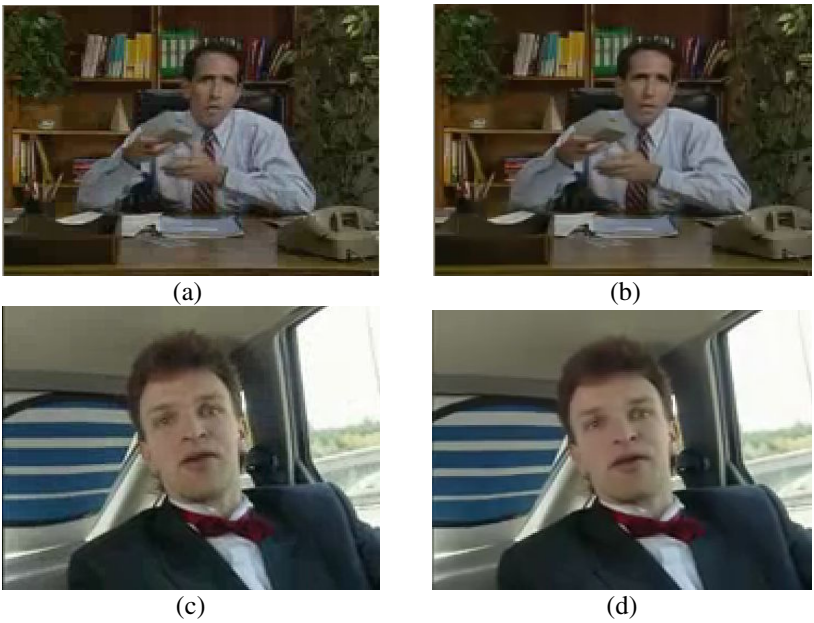
Fig. 5. The Hit Ratio comparison of JM 10.2 and proposed algorithm

From the table 9, it can be seen that the proposed algorithm can get up to 71% of encoding time reduction versus the reference software, meanwhile the degradation of the video quality is under a reasonable level.

Fig. 6 shows the visual quality comparisons of two sequences. One sequence "Salesman" has median motion activities and the other sequence "Carphone" has drastic motion activities. The pictures (a) and (c) in Fig.6 are the results by the

Table 9. Comparison of performance of the proposed algorithm and JM 10.2

sequences		Akiyo	Carphone	Clarie	Grandma	Miss_am	Salesman
QP=28	PSNR(dB)	-0.47	-0.69	-0.20	-0.12	-0.36	-0.36
	Bitrate(bps)	6.1	14.0	2.1	7.0	5.8	18.4
	Speedup(%)	73.3	71.3	69.8	71.4	69.2	70.2
QP=36	PSNR(dB)	**-0.53**	**-0.7**	**-0.5**	**-0.18**	**-0.3**	**-0.35**
	Bitrate(bps)	**1.33**	**3.6**	**1.5**	**1.46**	**2.2**	**3.7**
	Speedup(%)	**70.4**	**70.8**	**69.6**	**70.3**	**70**	**71**



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

Fig. 6. Visual quality comparisons of two sequences (QP=26)

proposed scheme whereas the pictures (b) and (d) are the results by JM 10.2. It is obvious that the proposed algorithm achieves a similar visual quality for human perceptual system.

5 Concluding Remarks

In this paper, we have proposed a fast mode decision algorithm using the early SKIP mode decision according to spatial-temporal correlation. The fast mode decision method has reduced the encoding time without any significant PSNR losses. Experimental results show that the proposed fast mode decision method reduces the encoding time by 71% on average with a slight PSNR drop. In this paper, we have only considered the early SKIP mode decision: however, we can also develop a variable block size mode decision of INTER mode in the future.

References

1. Wiegand, T., Sullivan, G.J., Bjntegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. IEEE Transaction on Circuits and Systems for Video Technology, 13(7) (July 2003)
2. Zhou, M.: Evaluation and Simplification of H.26L Baseline Coding Tools. ITU-T Q.6/16, Doc.#JVT-B030 (2002)
3. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264|ISO/IEC 14496-10 AVC) (March 2003)
4. Jeon, B., Lee, J.: Fast mode decision for H.264. ITU-T Q.6/16, Doc, #JVT-J033 (2003)
5. Gao, S., Lu, T.: An Improved Fast Mode Decision Algorithm in H.254 Video communications. In: ISSCAA 2006. 1st International Symposium (January 2006)
6. Feng, B., Zhu, G.-x., Liu, W.y.: Fast Adaptive Inter Mode Decision Method in H.264. In: Consumer Communications and Networking conference 2006, 3rd IEEE, January 2006. pp. 745–748 (2006)
7. http://iphome.hhi.de/suehring/tml/download/

SVM-Based Audio Classification for Content-Based Multimedia Retrieval

Yingying Zhu[1,2], Zhong Ming[1], and Qiang Huang[1]

[1] Faculty of Information Engineering, Shenzhen University, Shenzhen, 518060, P.R.China
[2] Software Engineering Ltd. of Harbin Institute of Technology, Haerbin, 150001, P.R. China
zyy211@gmail.com

Abstract. Audio classification is very important in multimedia retrieval such as audio indexing, analysis and content-based video retrieval. In this paper, we have proposed a clip-based support vector machine (SVM) approach to classify audio signals into six classes, which are pure speech, music, silence, environmental sound, speech with music and speech with environmental sound. The classification results are then used to partition a video into homogeneous audio segments, which is used to analyze and retrieve its higher-level content. The experimental results show that the proposed system not only improves classification accuracy, but also performs better than the other classification systems using the decision tree (DT), K Nearest Neighbor (K-NN) and Neural Network (NN).

Keywords: Audio Classification, Support Vector Machine, Multimedia Retrieval.

1 Introduction

Audio information often plays an important role in understanding the semantic content of multimedia. The audio signals can provide rich semantic cues that are very useful for analyzing multimedia content, so audio information has been recently used for content-based multimedia indexing and retrieval. Due the huge amount of audio data and the high expense of manual classification, an automatic audio classifier is need.

Many automatic audio classification schemes have been proposed in recent years. And audio classification has been an active research area. For example, Wold et al[1] presented an audio retrieval system named Music Fish based on audio classification . This work is a milestone about audio retrieval because of the content based analysis which distinguishes it from previous works. In this system, pitch, harmonicity, loudness, brightness and bandwidth were used as the audio features. The nearest neighbor (NN) rule was adopted to classify the query audio into one of the defined audio classes. Traditionally, the most common approach is speech/music classifications in which the highest accuracy has been achieved, especially when the segmentation information is known beforehand (i.e., manual segmentation). Saunders [2] developed a real-time speech/music classifier for audio in radio FM receivers

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 474–482, 2007.
© Springer-Verlag Berlin Heidelberg 2007

based on features such as zero crossing rate (ZCR) and short-time energy (STE). Zhang and Kuo [3] developed a content-based audio retrieval system, which performs audio classification into basic types such as speech, music and noise. In [4], Li concatenated the perceptual and cepstral feature sets for audio classification. A new classifier name nearest feature line (NFL) for audio classification was also presented and produced better results than the NN-based and other conventional methods. This study was improved by the author's later work [5]. By using SVM with the same feature set as [4] instead of the NFL method, Guo and Li managed to significantly improve the previous work on classification performance. Recently, Kiranyaz et al. [6] propose a generic frame work to classify audio into speech, music, fuzzy or silent. However, the fuzzy type can not tell whether the audio data is speech with music background or speech mixed with environmental sound.

There are simple audio types such as speech, music, silence and environmental sound, in addition, there are complex audio types such as speech with music, speech with environmental sound. The audio features that are used for differentiating simple audio types —such as zero crossing rate (ZCR)— do not work for complex audio types. From figure 1, we can see that it is difficult to distinguish three ZCR curves of speech, speech with music and speech with environmental sound. The challenge is how to classify complex audio type with the existing audio features.

In this paper, a clip-based SVM approach to classify audio signals into six classes. They are pure speech, music, silence, environmental sound, speech with music and speech with environmental sound. The clip-based SVM classifier is trained to recognize these audio types. Experiments carried out on the collected multimedia data have yielded an average 93.1% classification accuracy. Finally, a sophisticated comparison was conducted among the SVM-based, the DT (decision tree), K-NN (K nearest neighbor) and NN (neural network) classifiers. The results show that the proposed SVM-based scheme has outperformed the other three.

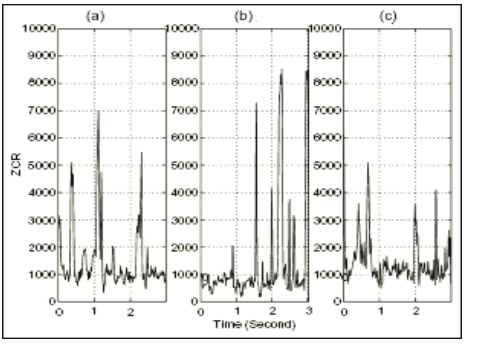


Fig. 1. The ZCR curves (a) speech with music (b) pure speech (c) speech with environmental sound

2 Overview of the Audio Classification System

Support Vector Machine (SVM) [7] has been successfully used in pattern recognition, such as speaker identification, face detection, and text recognition. A Support Vector Machine (SVM) is a supervised binary classifier which constructs a linear decision boundary or a hyperplane to optimally separate two classes. Since its inception, the SVM has gained wide attention due to its excellent performance on many real-world problems. Compared to other classifiers that separate the data in its original space, such as k Nearest Neighbor(k-NN), Neural Network (NN), and Naive Bayes (NB), SVM maps non-linear separable data to higher dimensional space and performs separation in that space. It is also reported that SVM can achieve a generalization performance that is greater than or equal to other classifiers, while requiring significantly less training data to achieve such an outcome [8]. Figure 2 shows the block diagram of our system. We present a clip-based SVM for audio classification. Experimental results show that the clip-based SVM is more effective than the file-based approaches by [5], [9]. In the feature selection, we adopted these audio features: Zero Crossing Rate (ZCR), Bandwidth (BW), Silence Ratio (SR), Harmonic Ratio (HR), Sub-band Energy (SBE1, SBE2, SBE3, SBE4), Spectrum Flux (SF), High ZCR Ratio (HZCRR), Low Short-Time Energy Ratio (LSTER).

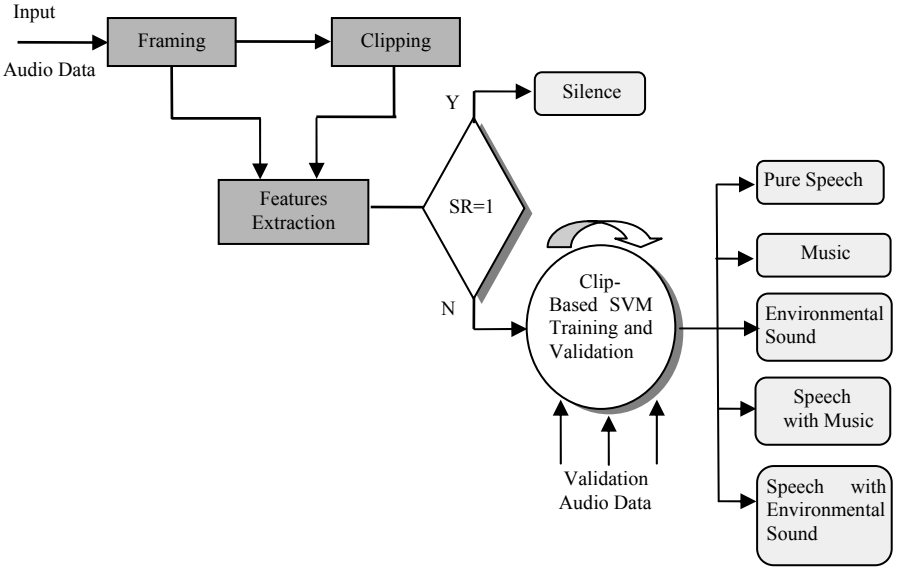


Fig.2. The block diagram of our system

3 Feature Extraction

To classify the audio track, all data are sampled at 11KHZ rate, with mono channel and 16 bits per sample. The extracted audio data are firstly segmented into non-overlapping one-second audio clips. This one-second audio clip is taken as the basic classification unit in our approach. It is further divided twenty 50ms non-overlapping audio frames, on which a 15HZ bandwidth expansion is applied. Then, various features are extracted from each frame or clip to represent it. Currently, these audio features, listed in section II, are considered in this work, which are chosen due to their effectiveness in capturing the temporal and spectral structures of different audio classes. A brief description of these features is given below.

3.1 Short-Time Zero Crossing Rate (ZCR)

Short-time Zero Crossing Rate (ZCR) is computed for every 50-ms frame which coarsely measures a signal's frequency content.

3.2 Bandwidth (BW)

Bandwidth is the square root of the power-weighted average of the squared difference between the spectral components and the frequency centroid. i.e.

$$BW = \sqrt{\int_0^{\overline{w}} (w - FC)^2 \left|F(w)\right|^2 dw / E} \quad (1)$$

3.3 Spectrum Flux (SF)

Mean of the spectrum flux (SF) is defined as the average variation of the spectrum between adjacent two frames in a one-second clip. The SF values of speech are higher than those of music, and those of environmental sound are the highest. Environmental sounds also change more dramatically than the other two signals. Figure 3 shows an example of spectrum flux of speech, music and environmental sound.

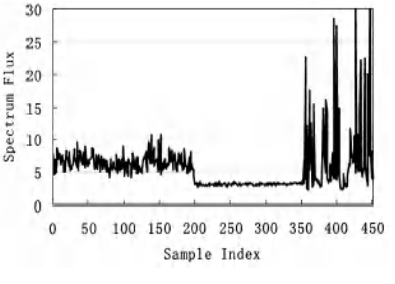


Fig. 3. The spectrum flux curve
(0-200s is speech; 201-350s is music and 351-450s is environmental sound)

3.4 Sub-band Energy (SBE)

The frequencies of audio segment are segmented into four ranges based on the relevant frequencies on speech and music. With 11kHZ sampling rate, we define the four frequency sub-bands to be [0, 700HZ], [700, 1400HZ], [1400, 2800HZ], and [2800, 5512HZ]. The sub-band energy for each audio segment is defined as the sum of the energy within each frequency range. We compute the variance of sub-band energy in SBE1 as the feature for each segment. Because of the bandwidth limitation that speech is within 8kHZ and music can span over 16kHZ, the variance of SBE1 of speech is usually higher than that of music.

3.5 Silence Ratio (SR)

Silence Ratio is defined as the ratio of the amount of silence frames in an audio clip to the total number of frames in the clip. SR is a useful statistical feature for audio classification, and it is usually used to differentiate music from speech [6]. Normally speech has higher SR than music. For each frame, the root mean square (RMS) is computed and compared to the RMS of the whole clip. If the frame RMS is less than 50% of clip RMS, we consider it as a silence frame.

3.6 Harmonic Ratio (HR)

Spectrum analysis shows that music is more harmonic than speech, since speech contains a sequence of tonal (vowels) and noise (consonants) [10]. Harmonic sound is defined as one that contains a series of frequencies which are derived from a fundamental or original frequency as a multiple of that. We compute the harmonic frequency of each frame using the algorithm in [11]. The harmonic ratio is defined as the ratio of the number of frames having a harmonic frequency to the total number of frames in the clip.

3.7 HZCRR

High ZCR Ratio (HZCRR) is defined as the ratio of the number of frames whose ZCR is above 1.5 fold average ZCR rate to the total number of frames in a one-second clip.

In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, its HZCRR will be greater than that of music. In order to illustrate the discriminability of HZCRR more clearly, we extracted HZCRR for each one-second audio clip in our training database. From these data, we obtained the probability distribution curves of HZCRR for speech and music, as shown in figure 4. Suppose we only use HZCRR to discriminate speech from music and use the cross-point of two curves as threshold, its error rate would be 19.36%.

3.8 LSTER

Low Shot-Time Energy Ratio (LSTER) is defined as the ratio of the number of frames whose short-time energy (STE) values are less than 0.5 times of the average STE to the total number of frames in a one-second clip.

In general, there are more silence frames in speech, so the LSTER measure will be much higher for speech than that for music. This can be seen clearly from the probability distribution curves of LSTER for speech and music, which are illustrated in the figure 5. Therefore, LSTER is good discriminator for speech and music.

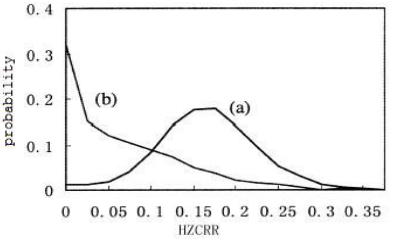


Fig .4. The probability distribution curves of HZCRR (a) speech (b) music

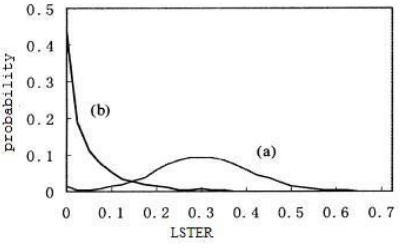


Fig.5. The probability distribution curves of LSTER (a) speech (b) music

4 Clip-Based SVM for Audio Classification

Except HZCRR, LSTER, SR and HR, the other features are extracted from each frame. The means and standard deviations in one audio clip are computed to get clip-based feature. Traditionally, each training and testing audio files merely generates one feature vector to train or test the SVM [5], [9]. The means and standard deviations of the feature trajectories over all the non-silent frames are computed, and these statistics are considered as feature sets to represent an audio file.

An audio file is segmented into separate audio clips. Passing through the procedure of feature extraction, each clip will be transformed into a feature vector. The steps for audio classification based on clip-based SVM follows.

(1) The input audio clips are first classified into silence and non-silence clip by SR. It will be marked as silence if SR=1.

(2) For non-silence audio clips :

a) Assume a N–clip audio file, $\mathbf{x}_i$, i = 1, …, N , is to be classified into type T_m, $m \in \{1,2,...,M\}$. For each type T_m, and for all the types T_n ($n \neq m$), we compute

$$f_H(T_{m,n} \mid \mathbf{x}_i) = \sum_{i=1}^{N} H(\mathbf{w}\mathbf{x}_i + b) - \sum_{i=1}^{N} H(-(\mathbf{w}\mathbf{x}_i + b)) \qquad (2)$$

by the $T_m - T_n$ 2-class SVM. For the optimal hyperlane , $\mathbf{w} \cdot \mathbf{x} + b = 0$, $\mathbf{w} \in R^N, b \in R$, $H(\cdot)$ is Heaviside step function.

b) The accumulated f for each type T_m is computed by

$$f(T_m \mid \mathbf{x}_i) = \sum\nolimits_n f_H(T_{m,n} \mid \mathbf{x}_i) \quad (3)$$

c) The most possible type T_{m^*} is chosen by

$$m^* = \arg\max_m f(T_m \mid \mathbf{x}_i) \quad (4)$$

(3) Finally, we group temporally adjoining one-second clips together if they share the same sound type. As a result, the entire audio stream will be partitioned into homogeneous segments with each having a distinct audio class label.

5 Experimental Results

In our experiments, the data used in our experiment are collected from real TV programs and movies, which is about 282 minutes in total. 92 minutes of data are used for training, and 190 minutes of data are used for testing. The training data test includes 22 minutes of speech, 18 minutes of environmental sound, 11 minutes of music, 16 minutes of speech with music and 25 minutes of speech with environmental sound, Music is mainly consisted of three components: pure instrumental music produced by different musical instruments, songs sung by male, female, or children, and some drama. Environmental sounds are composed of the sounds of applause, animal, footstep, explosion, vehicles, laugh, crowds, and so on. All data are converted into the uniform format of 11kHZ, 16-bit, mono-channel. In our experiments, we set one second as a test unit. If there are two audio types in a one-second audio clip, we will classify it as the time-dominant audio type.

Moreover, to find the optimal SVM parameters such as kernels, variance, margin and cost factor, we have also hand-labeled approximately 10 minutes of speech, 5 minutes of music, 2 minutes of environmental sound, 3 minutes of speech with music and 3 minutes of speech with environmental sound as validation data. Based on the validation results, we choose the radial basis function (RBF) as kernel and set parameters γ to 5 and C to 10.

Table 1 shows the classification results of the five audio types except silence. The classification accuracy is measured for each audio type which is defined as the ratio of correctly classified samples over the actual time of the type. In table 1, PT denotes predicted types, AT denotes actual types, PS denotes pure speech, M denotes music, E denotes environmental sound, S+M denotes speech with music, and S+E denotes speech with environmental sound.

Table 1. The classification results of clip-based SVM

PT(second) / AT(second)	PS	M	E	S+M	S+E
PS	7256	2	51	2	36
M	3	969	2	22	5
E	41	3	1056	0	71
S+M	4	38	0	603	0
S+E	48	28	112	8	1040
Total (second)	7352	1040	1221	635	1152
Accuracy	98.7%	93.2%	88.5%	95%	90.3%

In addition, to evaluate the effect of the presented SVM, we have also carried out the audio classification task using the DT, 5-NN and NN. The accuracies for DT, 5-NN and NN are shown in table 2. From these results, we find that environmental sound and speech with environmental sound have a lower accuracy than that of other types. This may be due to the fact that the definition of environmental sound is too board. Therefore, the characteristics displayed by different environmental sounds may be quite different, which results in lower accuracy compared with that of the other types without environmental sound. However, in all cases, SVM performs better than the other three classifiers. This is because SVM has a good performance on non-linear separable classes.

Table 2. The classification accuracies of four classifiers

Audio Types	DT	5-NN	NN	SVM
PS	90.3%	94.7%	98.2%	98.7%
M	84.5%	86%	91.7%	93.2%
E	78.6%	82.1%	83.6%	88.5%
S+M	77.5%	85.3%	94.5%	95%
S+E	68.9%	81.4%	87.2%	90.3%

6 Conclusions

In this paper, we have proposed a clip-based SVM approach to classify audio signals into six classes, which are pure speech, music, silence, environmental sound, speech with music and speech with environmental sound. Experiments have achieved the average 93.1%. Our comparison experiments show that SVM-based classifier outperforms other popular classifier such as DT, K-NN and NN. However, the best

choice of kernel for a given problem is still a research issue. On the other hand, the fuzzy definition of environmental sound has a negative effect on the accuracy of SVM (and other classifiers). Future works include the expansion of audio classes and the study for content-based audio segmentation.

References

1. Wold, E., Blum, T., Keislar, D., Wheaton, J.: Content-based Classification Search and Retrieval of Audio. IEEE Multimedia Magazine 3, 27–36 (1996)
2. Saunders, J.: Real Time Discrimination of Broadcast Speech/Music. In: Proc. ICASSP96. vol. 2, pp. 993–996, Atlanta, GA (1996)
3. Zhang, T., Kuo, C.-C.J.: Hierarchical Classification of Audio Data for Archiving and Retrieving, In: Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Proc., Phoenix, AZ, pp. 3001–3004 (1999)
4. Li, S.Z.: Content-based Audio Classification and Retrieval Using the Nearest Feature Line Method. IEEE Transactions on Speech and Audio Processing 8(5), 619–625 (2000)
5. Guo, G., Li, S.Z.: Content-based Audio Classification and Retrieval by Support Vector Machines. IEEE Transactions on Neural Networks 14(1), 209–215 (2003)
6. Qureshi, A.F., Kiranyaz, S., Gabbouj, M.: A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval. IEEE Trans. on Speech and Audio Processing 9(3), 517–523 (2006)
7. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 1–47 (1998)
8. Wan, V., Campbell, W.M.: Support Vector Machines for Speaker Verification and Identification. In: Proc. of the IEEE Signal Processing Society Workshop on Neural Networks. vol. 2, pp. 775–784 (2000)
9. Lin, C.C., Chen, S.H., Truong, T.K., Chang, Y.: Audio Classification and Categorization Based on Wavelets and Support Vector Machine. IEEE Transactions on Speech and Audio Processing 13(5), 644–651 (2005)
10. Srinivasan, S., Petkovic, D., Ponceleon, D.: Towards Robust Features for Classifying Audio in the Cue Video System. In: Proc. 7th ACM Int. Conf. on Multimedia, pp. 253–261 (1999)
11. Pfeiffer, S., Fischer, S., Effelsberg, W.: Automatic Audio Content Analysis. In: Proc. 4th ACM Int. Conf. on Multimedia, pp. 127–132 (1996)

Moving Object Tracking in H.264/AVC Bitstream

Wonsang You, M.S. Houari Sabirin, and Munchurl Kim

School of Engineering, Information and Communications University,
Munjiro 119, Daejeon, 305732, Republic of Korea
{wsyou, houari, mkim}@icu.ac.kr

Abstract. Data broadcasting services are required to provide user interactivity through connecting additional contents such as object information to audio-visual contents. H.264/AVC-based metadata authoring tools include functions which identify and track position and motion of objects. In this work, we propose a method for tracking the target object by using partially decoded texture data and motion vectors extracted directly from H.264/AVC bitstream. This method achieves low computational complexity and high performance through the dissimilarity energy minimization algorithm which tracks feature points adaptively according to these characteristics. The experiment has shown that the proposed method had high performance with fast processing time.

Keywords: Object Tracking, H.264/AVC, Dynamic Programming, Neural Network.

1 Introduction

H.264/AVC does not handle video objects directly while MPEG-4 contains object-based encoding scheme. However, the interactive broadcasting service should provide additional object information as the form of MPEG-7 metadata to support user interactivity. The metadata authoring tool includes object tracking function which generates the position information of the predefined target object in all frames.

Unlike pixel-based object tracking approaches, object tracking algorithms for H.264/AVC videos can achieve lower computational complexity by using block-based motion vectors or residual data extracted directly from encoded bitstream; these are called compressed domain approaches. One difficulty in these approaches is that motion vectors do not always coincide with the true motion or optical flow.

To overcome the above difficulty, many researchers have proposed a variety of object tracking algorithms for MPEG videos. These can be classified as two categories as the motion-based method and the residual-based method. The motion-based methods rely on the probabilistic properties of the motion vector field. Babu et al. [1] predicted the motion of objects corresponding to affine motion parameters which are computed by the expectation maximization (EM) algorithm. Treetasanatavorn et al. [2] applied the Bayesian method to separate the significant foreground object from the background given the motion vector field. Zeng et al. [3] assigned the object label to blocks with homogeneous motion through the Markovian labeling

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 483–492, 2007.
© Springer-Verlag Berlin Heidelberg 2007

procedure. On the other hand, the residual-based methods use the statistical properties of DCT coefficients such as histogram. Aggarwal et al. [4] found the target object by histogram matching and motion interpolation. However, these algorithms have great tracking error in long image sequences due to block-based information. Moreover, motion-based methods tend to have high computational complexity or low performance in cases of deformable objects. Also, the residual-based methods are not applicable for H.264/AVC videos since residual data of intra-coded blocks is transformed from spatially intra-predicted values instead of original pixel values.

In this paper, we propose the dissimilarity energy minimization algorithm which uses motion vectors and partially decoded luminance signals to perform tracking adaptively according to properties of the target object in H.264/AVC videos. It is one of the feature-based approaches that tracks some feature points selected by a user. First, it roughly predicts the position of each feature point using motion vectors extracted from H.264/AVC bitstream. Then, it finds out the best position inside the given search region by considering three clues such as texture, form, and motion dissimilarity energies. Since just neighborhood regions of feature points are partially decoded to compute this energy, the computational complexity is scarcely increased. The set of the best positions of feature points in each frame is selected to minimize the total dissimilarity energy by dynamic programming. Also, weight factors for dissimilarity energies are adaptively updated by the neural network.

This paper is organized as follows. First, we describe the proposed object tracking algorithm at Section 2 and 3. Then, experimental results are presented in Section 4 and finally the conclusion is drawn in Section 5.

2 Forward Mapping of Backward Motion Vectors

The motion vectors extracted directly from H.264/AVC bitstream can be used to predict roughly the motion of feature points. Since all motion vectors in P-frames have backward direction, it should be changed to have forward direction. Following Porikli and Sun [5], the forward motion field is built by the region-matching method. First, motion vectors of blocks with various sizes are dispersed to 4x4 unit blocks. After each block is projected to the previous frame, the set of overlapping blocks is extracted as shown at Fig. 1.

Forward motion vectors of overlapped blocks in the previous frame are updated with respect to the ratio of the overlapping area to the whole block area. Assuming

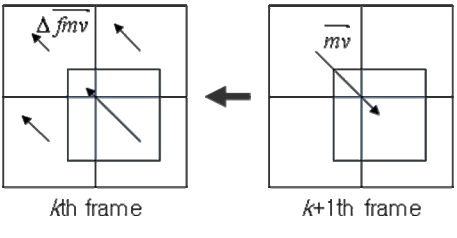


Fig. 1. The region-matching method for constructing the forward motion field

that the *j*th 4x4 block $b_{k,j}$ in the *k*th frame is overlapped with the *i*th 4x4 block $b_{k-1,i}$ in the *k-1*th frame, the forward motion vector $fmv_{k-1}(b_{k-1,i})$ is given by

$$\overrightarrow{fmv}_{k-1}\left(b_{k-1,i}\right) = -\sum_{j=1}^{N}\left(\frac{S_{k-1}(i,j)}{16}\cdot\overrightarrow{mv}_k\left(b_{k,j}\right)\right) \tag{1}$$

where $S_{k-1}(i,j)$ stands for the overlapping area between $b_{k,j}$ and $b_{k-1,i}$, and $mv_k(b_{k,j})$ denotes the backward motion vector of $b_{k,j}$ with $i,j=1,2,\ldots,N$. We assume that H.264/AVC videos are encoded in the baseline profile which each GOP contains just one I-frame and several P-frames. It should be noticed that the above region-matching method cannot be applied in the last P-frame in one GOP since the next I-frame does not have backward motion vectors. Assuming that the motion of each block is approximately constant within a small time interval, the forward motion vector of any block in the last P-frame can be assigned as a vector with the reverse direction of the backward motion vector as expressed by

$$\overrightarrow{fmv}_{k-1}\left(b_{k-1,i}\right) = -\overrightarrow{mv}_{k-1}\left(b_{k-1,i}\right). \tag{2}$$

Thereafter, positions of feature points in the next frame are predicted using forward motion vectors. If the *n*th feature point in the *k-1*th frame has the displacement vector $f_{k-1,n}=(fx_{k-1,n},fy_{k-1,n})$ and is included in the *i*th block $b_{k-1,i}$, the predicted displacement vector $p_{k,n}=(px_{k,n},py_{k,n})$ in the *k*th frame is defined as

$$\vec{p}_{k,n} = \vec{f}_{k-1,n} + \overrightarrow{fmv}_{k-1}\left(b_{k-1,i}\right). \tag{3}$$

3 Moving Object Tracking in H.264/AVC Bitstream

Since the predicted position of any feature point is not precise, we need the process of searching the best position of any feature point inside the search region centered at the predicted position $p_{k,n}= (px_{k,n},py_{k,n})$. It is checked whether each candidate point inside the search region is the best position using the dissimilarity energies related to texture, form, and motion. The set of candidate points with the minimum total dissimilarity energy is selected as the optimal configuration of feature points.

3.1 Texture Dissimilarity Energy

The similarity of texture means how the luminance property in neighborhood of a candidate point is similar with that in the previous frame. The set of candidate points inside the square search region is denoted as $C_{k,n}=\{c_{k,n}(1), c_{k,n}(2),\ldots, c_{k,n}(L)\}$ with $L=(2M+1)\times(2M+1)$ in the case of the *n*th feature point in the *k*th frame. Then, the texture dissimilarity energy E_C for the *i*th candidate point $c_{k,n}(i)=(cx_{k,n}(i),cy_{k,n}(i))$ is defined as

$$E_C\left(k;n,i\right) = \frac{1}{(2W+1)^2}\sum_{x=-W}^{W}\sum_{y=-W}^{W}\left|s_k\left(x+cx_{k,n}(i),y+cy_{k,n}(i)\right)-s_{k-1}\left(x+cx_{k,n}(i),y+cy_{k,n}(i)\right)\right| \tag{4}$$

where $s_k(x,y)$ stands for the luminance value in a pixel (x,y) of the *k*th frame, and *W* is the maximum half interval of neighborhood. The smaller E_C is, the more the texture of its neighborhood is similar with that of the corresponding feature point in the previous

frame. This energy forces the best point to be decided as the position with the most plausible neighbor texture as far as possible. Fig. 2 shows how the search region and the neighborhood of a candidate point are applied to calculate E_C.

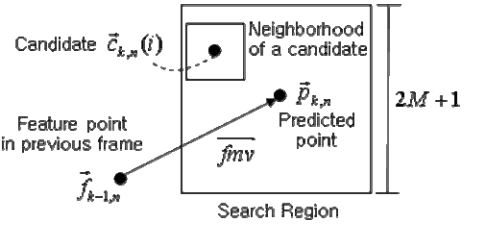


Fig. 2. The search region is centered at the predicted point located by a forward motion vector. A candidate point inside the search region has its neighborhood of square form to compute E_C.

Only necessary blocks can be partially decoded in P-frames to reduce the computational complexity. On the other hand, intra-coded blocks are impossible to be partially decoded since these are spatially intra-coded from these neighbor blocks.

General partial decoding takes long time since decoding particular blocks in P-frames requires many reference blocks to be decoded in the previous frames. We can predict decoded blocks to reduce the computation time. To predict decoded blocks in the *k*th P-frame, we assume that the velocity inside one GOP is as uniform as the forward motion vector of the *k-2*th frame. For the *i*th frame with $i=k,k+1,\ldots,K$, the predicted search region $P_{k,n}(i)$ is defined as the set of pixels which are necessary to calculate the texture dissimilarity energies of all possible candidate points for the *n*th feature point. Then, the half maximum interval $T_{k,i}$ of $P_{k,n}(i)$ is $T_{k,i}=(i-k+1)\times M+W+\gamma$ where γ denotes the prediction error. Then, $P_{k,n}(i)$ is given as follows:

$$P_{k,n}(i)=\left\{\vec{p}\,\middle|\,\vec{p}=(i-k+1)\overrightarrow{fmv}_{k-2}\left(b\left(\vec{f}_{k-2,n}\right)\right)+\vec{m}+\vec{f}_{k-1,n},\vec{m}=(x_m,y_m);x_m,y_m=-T_{k,i},\ldots,T_{k,i}\right\} \tag{5}$$

where $b(f_{k-2,n})$ stands for the block which includes the *n*th feature point $f_{k-2,n}$. The decoded block set $D_{k,n}(i)$ is defined as the set of blocks which should be decoded to reconstruct $P_{k,n}(i)$. Using the motion vector of the *k-1*th frame, $D_{k,n}(i)$ is given by

$$D_{k,n}(i)=\left\{b\left(\vec{d}\right)\middle|\vec{d}=(i-k)\overrightarrow{mv}_{k-1}\left(b\left(\vec{f}_{k-1,n}\right)\right)+\vec{p},\vec{p}\in P_{k,n}(i)\right\}. \tag{6}$$

Assuming that there exist *F* feature points, the total decoded block set D_k in the *k*th frame can be finally computed as

$$D_k=\bigcup_{n=1}^{F}\bigcup_{i=k}^{K}D_{k,n}(i). \tag{7}$$

Fig. 3 shows how partial decoding is performed in the first P-frame of one GOP which contains one I-frame and three P-frames. It should be noticed that the time for calculating the total decoded block set is proportional to the GOP size.

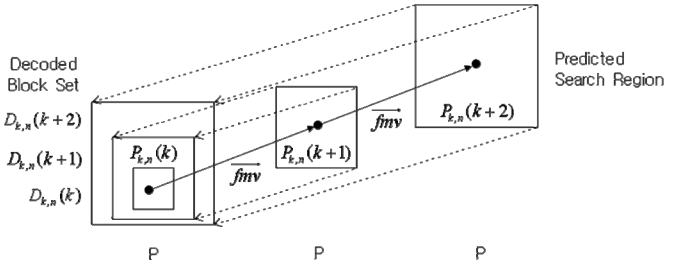


Fig. 3. The structure of partial decoding in the first P-frame of a GOP which contains one I-frame and three P-frames. Two decoded block sets $D_{k,n}(k+1)$ and $D_{k,n}(k+2)$ in the first P-frame are projected from two predicted search regions $P_{k,n}(k+1)$ and $P_{k,n}(k+2)$.

3.2 Form Dissimilarity Energy

The similarity of form means how the network of candidate points is similar with the network of feature points in the previous frame. Each feature point is jointly linked by a straight line like Fig. 4. After a feature point is initially selected, it is connected to the closest one among non-linked feature points. In this way, the feature network in the first frame is built by connecting all feature points successively.

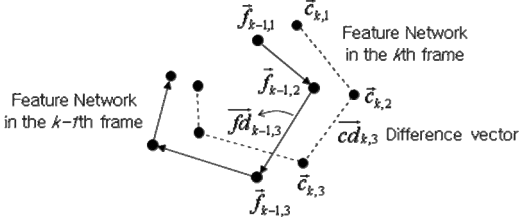


Fig. 4. The network of feature points in the previous frame and the network of candidate points in the current frame

To calculate the form dissimilarity energy of each candidate point, we assume that each feature point is arranged in the order named at the first frame. The feature point $f_{k-1,n}$ in the k-1th frame has its difference vector $fd_{k-1,n}(i)=f_{k-1,n}(i)-f_{k-1,n-1}(i)$ as shown at Fig. 4. Likewise, the ith candidate point of the nth feature point in the kth frame has its difference vector $cd_{k,n}(i)=c_{k,n}(i)-c_{k,n-1}(j)$. Then, the form dissimilarity energy E_F for the ith candidate point of the nth feature point (n>0) is defined as follows:

$$E_F(k;n,i)=\left\|\overrightarrow{cd}_{k,n}(i)-\overrightarrow{fd}_{k-1,n}\right\|^{1/2}. \tag{8}$$

All candidate points of the first feature point (n=0) have zero form dissimilarity energy $E_F(k;0,i)$=0. The smaller E_F is, the less the form of the feature network will be transformed. The form dissimilarity energy forces the best position of a candidate point to be decided as the position where the form of the feature network is less changed as far as possible.

3.3 Motion Dissimilarity Energy

The reliability of a forward motion vector means how it is similar with true motion enough to get a predicted point as exactly as possible. Following Fu et al. [6], if the predicted point $p_{k,n}$ which has located by the forward motion vector fmv_{k-1} returns to its original location in the previous frame by the backward motion vector mv_k, fmv_{k-1} is highly reliable. Assuming that $p_{k,n}$ is included to the jth block $b_{k,j}$, the reliability R can be given as follows:

$$R\left(\vec{p}_{k,n}\right)=\exp\left(-\frac{\left\|\overrightarrow{fmv}_{k-1}\left(b_{k-1,i}\right)+\overrightarrow{mv}k\left(b_{k,j}\right)\right\|^2}{2\sigma^2}\right) \tag{9}$$

where σis the variance of reliability. Fig. 5 shows forward motion vectors with high and low reliability. In a similar way of Fu's definition [6], the motion dissimilarity energy E_M for the ith candidate point is defined as follows:

$$E_M(k;n,i)=R\left(\vec{p}_{k,n}\right)\left\|\vec{c}k,n(i)-\vec{p}_{k,n}\right\|. \tag{10}$$

With high reliability R, E_M has greater effect on finding the best point than E_C or E_F since it is sharply varying according to the distance between a predicted point and a candidate point.

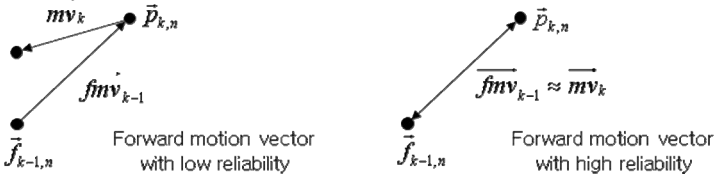


Fig. 5. The reliability of forward motion vectors. The great gap between a forward motion vector and a backward motion vector results in low reliability.

3.4 Energy Minimization

The dissimilarity energy $E_{k,n}(i)$ for the ith candidate point of the nth feature point is defined as follows:

$$E_{k,n}(i)=\omega_C(k)E_C(k;n,i)+\omega_F(k)E_F(k;n,i)+\omega_M(k)E_M(k;n,i) \tag{11}$$

where $w_C(k)$, $w_F(k)$, and w_M(k) are weight factors for texture, form, and motion dissimilarity energy. If the configuration of candidate points is denoted as $I=\{c_{k,1}(i_1), c_{k,2}(i_2),\ldots,c_{k,F}(i_F)\}$, the optimal configuration $I_{opt}(k)$ in the kth frame is selected as what minimizes the total dissimilarity energy $E_k(I)$ expressed by

$$E_k(I)=\sum_{n=1}^{F} E_{k,n}(i_n). \tag{12}$$

When all possible configurations of candidate points are considered, it takes so much time $((2M+1)^{2F})$ that causes high computation complexity especially in cases of large search region or many feature points. We can reduce the amount of computations by (F) using the discrete multistage decision process called the dynamic programming which corresponds to two steps [7]:

1) The accumulated dissimilarity energy (ADE) $E_{local}(n,i)$ for the ith candidate point of the nth feature point ($n>0$) is calculated as follows:

$$E_{local}(n,i)=\min_{j}\left[E_{k,n}(i,j)+E_{local}(n-1,j)\right]. \tag{13}$$

The ADE for the first feature point is $E_{local}(0,i)=E_{k,0}(i)$. Then, the point which minimizes the ADE is selected among candidate points of the n-1th feature point; the index of this point is saved as

$$s(n,i)=\arg\min_{j}\left[E_{k,n}(i,j)+E_{local}(n-1,j)\right]. \tag{14}$$

2) In the last feature point, the candidate point with the smallest ADE is selected as the best point o_F. Then, the best point o_n for the nth feature point is heuristically decided as follows:

$$o_F=\arg\min_{i}\left[E_{local}(F,i)\right] \text{ and } o_n=s(n+1,o_{n+1}). \tag{15}$$

The best position for nth feature point $f_{k,n}$ is $f_{k,n}=c_{k,n}(o_n)$.

3.5 Adaptive Weight Factors

The arbitrarily assigned weight factors for texture, form, and motion dissimilarity energy can give rise to tracking error since the target object can have various properties. In this reason, weight factors need to be decided adaptively according to properties of the target object. For instance, for an object which texture is scarcely changing, the weight factor w_C should be automatically set up as high value.

Weight factors can be automatically updated in each frame by using the neural network as shown in Fig. 6. The dissimilarity energy E_k is transformed to its output value E_k by the nonlinear activation function ξ. The update of weight factors is performed by the backpropagation algorithm which minimizes the square output error ε_k defined as follows:

$$\varepsilon_k = \frac{1}{2}\left(E_d-\dot{E}_k\right)^2. \tag{16}$$

where E_d denotes the ideal output value. If the activation function ξ is the unipolar sigmoidal function ($\xi(x)=1/(1+e^{-x})$), the gradient of a weight factor is calculated as

$$\Delta\omega_x(k) = \eta\left(E_d-\dot{E}_k\right)\dot{E}_k\left(1-\dot{E}_k\right)E_x(k) \tag{17}$$

where x can be T (texture), F (form), or M (motion), and η is the learning constant [8].

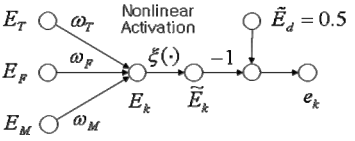


Fig. 6. The neural network for updating weight factors

4 Experimental Results

To demonstrate the performance of the proposed method, the tracking results of various objects have extracted from videos such as "Stefan", "Coastguard" and "Lovers" with CIF size. Each video was encoded as the GOP structure of 'IPPP' in the baseline profile, and included P-frames whose previous frame only can be a reference frame. Fig. 7(a) shows the tracking results of a rigid object with slow motion in "Coastguard". Four feature points were well tracked in the uniform form of feature network. Fig. 9(b) also shows the tracking result of a deformable object with fast motion in "Stefan". We can observe that tracking is successful even though the form of feature network is greatly changing due to fast three-dimensional motion.

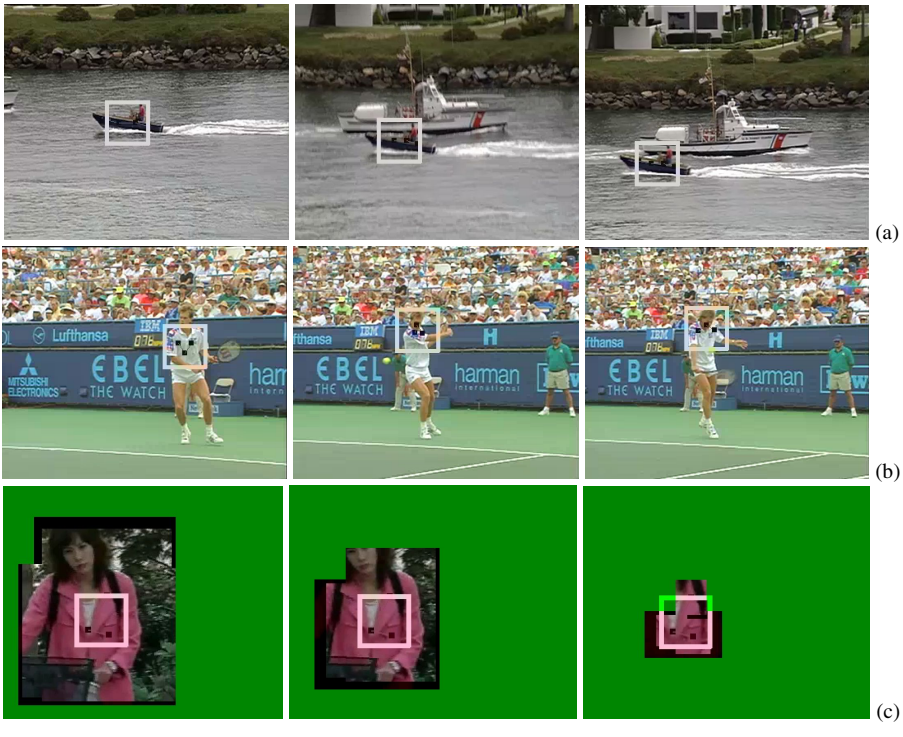


Fig. 7. The object tracking in (a) "Coastguard", (b) "Stefan" with 100 frames, and (c) "Lovers" with 300 frames. Partially decoded regions are shown in "Lovers".

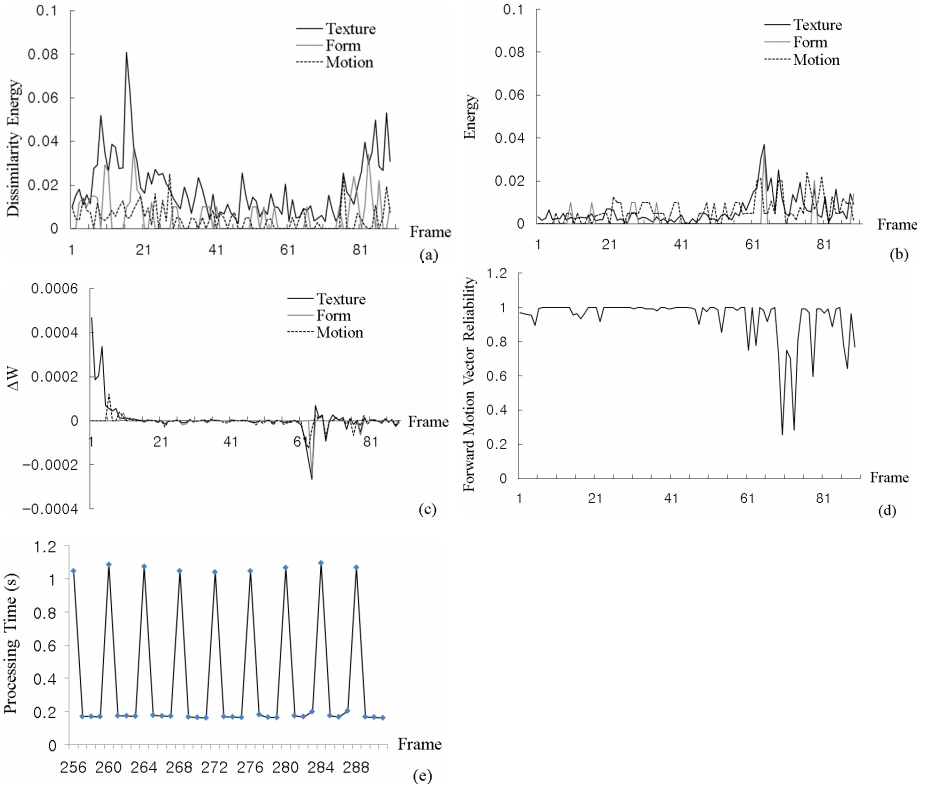


Fig. 8. (a) Dissimilarity energies in "Stefan" and (b) "Coastguard", (c) the variation of weight factors, (d) the average reliability of *fmv* in "Coastguard", (e) the processing time in "Lovers"

Fig. 7(c) represents the visual results of partial decoding in P-frames of "Lovers" when the search half interval M and the neighborhood half interval W are assigned as 10 and 5. Only the neighborhood region of three feature points was partially decoded. Even in a sequence "Lovers" with 300 frames, no tracking errors were found.

Numerical data of tracking from two video samples is shown at Fig. 8. In Fig. 8(a) and (b), dissimilarity energies in "Coastguard" are lower than those in "Stefan". We can see from this result that the variation of texture, form, and motion in "Coastguard" is smaller than "Stefan". Fig. 8(d) shows a plot for the average reliability of forward motion vectors in "Coastguard". The average percentage of reliabilities in "Coastguard" is 93.9% higher than 12.2% in "Stefan"; it indicates that the motion dissimilarity energy is the good measure for the motion property of the target object.

Through the neural network, the square error of dissimilarity energy is minimized. When the learning constant was equal to 5, this error had approximately zero value after the 15th frame. Moreover, weight factors converge on optimal values as shown at Fig. 8(c). We can observe that weight factor variations and dissimilarity energies increase greatly from the 61th frame to the 66th frame in "Coastguard"; it illustrates that weight factors are adaptively controlled when another ship is approaching.

When the JM reference software was used to read H.264/AVC bitstream, the computation time was about 430ms/frame (Intel Pentium 4 CPU 3.2GHz 1GB RAM). As shown at Fig. 8(e), most computations are performed at I-frames which are fully decoded. In this reason, if faster decoder is used, we can reduce the computation time by 230ms/frame maximally. The computation time is nearly similar with Zeng's algorithm [3] which computation time is roughly 450ms/frame in "PIE" sequence. However, Zeng's algorithm cannot track a target object identified by a user since it extracts all moving objects. In order to select the target object from all segmented objects, it requires more computation. Therefore, the proposed algorithm can track the target object in real-time applications with faster speed than Zeng's algorithm.

5 Conclusion

We have proposed a novel object tracking algorithm with low computational complexity and high performance. It finds the best positions of feature points which have high similarity in texture, form, and motion. Moreover, the computational complexity can be reduced by the partial decoding and the dynamic programming for optimal energy minimization. Also, main parameters are adaptively optimized according to properties of the target object. We demonstrated that the proposed algorithm can track precisely deformable objects or fast-moving objects in small computation time. It can be applied to the metadata authoring tool which generates the position information of the target object. In future work, we will study the automatic extraction of feature points using motion vectors in H.264/AVC bitstream.

References

1. Babu, R.V., Ramakrishnan, K.R.: Video Object Segmentation: A Compressed Domain Approach. IEEE Trans. Circuits Syst. Video Technol. 14, 462–474 (2004)
2. Treetasanatavorn, S., Rauschenbach, U., Heuer, J., Kaup, A.: Bayesian Method for Motion Segmentation and Tracking in Compressed Videos. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 277–284. Springer, Heidelberg (2005)
3. Zeng, W., Du, J., Gao, W., Huang, Q.: Robust moving object segmentation on H.264/AVC compressed video using the block-based MRF model. Real-Time Imaging 11(4), 290–299 (2005)
4. Aggarwal, A., Biswas, S., Singh, S., Sural, S., Majumdar, A.K.: Object Tracking Using Background Subtraction and Motion Estimation in MPEG Videos. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3852, pp. 121–130. Springer, Heidelberg (2006)
5. Porikli, F., Sun, H.: Compressed Domain Video Object Segmentation. Technical Report TR2005-040 of Mitsubishi Electric Research Lab (2005)
6. Fu, Y., Erdem, T., Tekalp, A.M.: Tracking Visible Boundary of Objects Using Occlusion Adaptive Motion Snake. IEEE Trans. Image Processing 9, 2051–2060 (2000)
7. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge, MA (2001)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, New York (2001)

A Prediction Error Compression Method with Tensor-PCA in Video Coding

Jian Liu, Fei Wu, Lei Yao, and Yueting Zhuang

College of Computer Science, Zhejiang University,
Hangzhou, 310027, China
{ken3300,yzhuang,wufei}@zju.edu.cn, leiyzju@yahoo.com.cn

Abstract. Discrete Cosine Transform (DCT), which is employed by block-based hybrid video coding to encode motion prediction errors, has dominated practical video coding standards for several decades. However, DCT is only a good approximation to Principle Component Analysis (PCA, also called KLT), which is optimal among all unitary transformations. PCA is rejected by coding standards due to its complexity. This paper tries to use a matrix form of PCA (which we call tensor-PCA) to encode prediction errors in video coding. This method retains the performance of traditional PCA, but can be computed with much less time and space complexity. We compared tensor-PCA with DCT and GPCA in motion prediction error coding, which shows that it is a good trade-off between compression efficiency and computational cost.

Keywords: Tensor-PCA, video coding, prediction errors.

1 Introduction and Related Work

Transform coding is the most common technique used in lossy image and video compression [7]. The motivation behind transform coding is that the succeeding job of quantization and entropy coding can be easier in the transform domain. Natural images are information sources with memory, that is, strong dependencies exist among pixels. If the value of one signal is known, the value of the next one can be predicted to some extent. Thus, the amount of information carried by a single signal is much less than that in memoryless sources. Therefore, one cannot get satisfying coding efficiency by directly applying entropy coding to individual signals. To solve the problem, we transform the signal sequence into a low dimensional vector space, where the transform coefficients are more independent. People have found a lot of transforms that can achieve this goal while retaining the energy of the original signal, e.g. DFT, DCT, Hadamard transformation, Harr transformation, etc. [5]. Among these transforms, Principle Component Analysis (PCA, also called Karhunen-Loeve transform, KLT) is the optimal for coding of Gaussian source [4]. But PCA has a major shortcoming: it has to estimate a linear model for every image. The computational cost made it unsuitable for practical use. So DCT is widely used as an approximation.

The wisdom of PCA, however, is based on the assumption that the source signal is Gaussian distributed. This assumption is not strictly true for natural images.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 493–500, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Experiments show that the image signals tend to have a multi-linear structure, which means that the signals tend to reside in several sub-spaces of possibly different dimensions or from a mixture of multiple Gaussian distributions. So an approach to the problem of estimating a mixture of linear subspaces from sample data points is proposed by René Vidal, Yi Ma et al. [6]. It is called Generalized Principle Component Analysis (GPCA). GPCA has shown good performance on image and video compression [8]. But this method is even more complex than PCA coding, which makes it impractical for implementation.

Both traditional PCA and GPCA use vectors to represent images or image blocks. But we should be aware that images are two-dimensional matrices in nature. Turning matrices into vectors may cause loss of spatial information and on the other hand results in large matrices for computation, which leads to high space and time cost. Another scheme proposed by Jieping Ye et al. [3] works directly with matrices. Instead of projecting images into a traditional vector space, it treats images as matrices and projects them into a space which is the tensor product of two vector spaces. This scheme is also called generalized PCA, but we name it as tensor-PCA to differentiate it from the GPCA in [6]. Theory and experiments prove that this method has a lower computational cost as well as retains better image quality than PCA.

In the field of video coding, motion compensation is an essential step for removing temporal dependencies among neighboring frames. Transform coding is then used to encode the motion prediction errors. DCT is prevalently used for its efficiency. GPCA has also been applied on residual data coding [2], but it is way too complex. This paper proposes to apply tensor-PCA on prediction error coding. Given a group of prediction error images, a low dimensional tenor space is approximated and the images are projected into this space, thus a compressed representation can be achieved. Our experiments show that tensor-PCA achieves better PSNR than DCT with an acceptable computational cost.

The rest of the paper is organized as follows. Section 2 introduces the algorithm of tensor-PCA. Section 3 presents how tensor-PCA is used for prediction error coding. Section 4 gives experiment results. Conclusion is given in Section 5.

2 From PCA to Tensor-PCA

Here we give a brief introduction to PCA and tensor-PCA, which is introduced by Jieping Ye et al [3].

Principle Component Analysis (PCA) is a powerful tool in dimension reduction, which can be applied to data compression. It condenses most of the information into a small number of dimensions by projecting the data into a low dimensional vector space. PCA can only cope with data in a vector form. Taking an image of $M \times N$ pixels for example, we first divide it into $m \times n$ blocks. The pixels in each block are stacked into a $m \times n$ vector v_i. There are c vectors altogether which forms a matrix U, where $c = (M \times N)/(m \times n)$. Then PCA computes the covariance matrix Σ of v_i and diagonalizes it as $\Sigma = \Phi \Lambda \Phi^T$. The diagonal matrix Λ contains the eigenvalues and the matrix Φ contains the eigenvectors $\{\phi_i\}_{i=1}^{m \times n}$. Then the transform

$V = \Phi^T U$ projects all vectors v_i into the new space spanned by ϕ_i. The energy of the data points in the transformed space is concentrated in a small number of dimensions. By keeping the p eigenvectors corresponding to the largest p eigenvalues, the dimension of the transformed space is reduced to p. So Φ and V form a compressed representation of the image. The original data can be recovered with minor loss of information.

Tensor-PCA does not need to transform an image or image-block into a vector. It directly takes a matrix as input. A matrix representation can retain the spatial information of pixels, as well as reduce the size of the matrix for computing the eigenvectors.

We still take the $M \times N$ image divided into $m \times n$ blocks: $\{X_i\}_{i=1}^k$. First we compute the mean of X_i as $mean = \frac{1}{k}\Sigma_{i=1}^k X_i$, and let $\tilde{X}_i = X_i - mean$. Now we treat each $m \times n$ matrix $\tilde{X}_i$ as a data point in a matrix space (or a 2-rank tensor space). To project the matrix into a lower dimensional space, we use a $l_1 \times l_2$ dimensional tensor axis system: $u_i \otimes v_j$, where $\otimes$ denotes the tensor product, $u_i \in \mathbb{R}^{m\times 1}$, $v_i \in \mathbb{R}^{n\times 1}$, $i = 1,2,...l_1$, $j = 1,2,...,l_2$. Given a matrix X, its projection on the (i,j)-th coordinate $u_i \otimes v_j$ is $u_i \cdot X \cdot v_j$. In fact, if we form two matrices $U = (u_1, u_2, ..., u_i)$ and $V = (v_1, v_2, ..., v_j)$, the computation of the projection can be denoted as $Y_i = U^T \tilde{X}_i V$, where Y_i is a $l_1 \times l_2$ matrix. Because the width and height of a natural image is usually comparable, l_1 and l_2 are typically set equal, e.g. $l_1 = l_2 = d$. The idea behind this method is that there should be an optimal $u_i \otimes v_j$ axis system that maximizes the variance of the data points in the projected space, so that the transformation results in a sparse representation of the original image. This maximization problem can be solved by an iterative process. The algorithm is based on the following theorem.

Theorem 2.1. Let U, V be the matrices maximizing the variance $\text{var}(U,V) = \frac{1}{k-1}\Sigma_{i=1}^k \| U^T \tilde{X}_i V \|^2$. Then

1. For a given V, matrix U consists of the l_1 eigenvectors of the matrix $M_U = \Sigma_{i=1}^k \tilde{X}_i V V^T \tilde{X}_i^T$ corresponding to the largest l_1 eigenvalues.
2. For a given U, matrix V consists of the l_2 eigenvectors of the matrix $M_V = \Sigma_{i=1}^k \tilde{X}_i^T U U^T \tilde{X}_i$ corresponding to the largest l_2 eigenvalues.

The proof of the theorem can be found in [3].

With the above theorem, an iterative procedure can be developed to compute U and V. Given an initial matrix U, the corresponding V can be found by computing the eigenvectors of M_U. Then the matrix U can be updated by computing the eigenvectors of M_V. This procedure is repeated until the result converges. For the proof of convergence of the algorithm please refer to [3]. Experiments show that by choosing the initial matrix $U = (I_d, 0)^T$, where I_d is the identity matrix, the result can be excellent. And surprisingly, the result usually converges within two iterations!

When U and V are computed, the transformation of $\tilde{X}_i$ can be computed as $Y_i = U^T \tilde{X}_i V$. Conversely, with U, V and Y_i, we can reconstruct $\tilde{X}_i$ as $\tilde{X}_i \approx U Y_i V^T$. The reconstruction error for $\tilde{X}_i$ is $e_i = \left\| \tilde{X}_i - U Y_i V^T \right\| = \left\| \tilde{X}_i - U U^T \tilde{X}_i V V^T \right\|$, where $\|\cdot\|$ means the norm of a matrix. To measure the convergence of the algorithm, a root mean square error (RMSE) is defined as:

$$RMSE = \sqrt{\frac{1}{k} \Sigma_{i=1}^{k} \left\| \tilde{X}_i - U U^T \tilde{X}_i V^T V \right\|^2} \tag{1}$$

It has been shown in [3] that the iterative algorithm of computing U and V converges, which can be measured by RMSE. The pseudo code of the algorithm is shown below.

```
Tensor-PCA ( X_1, X_2, ..., X_k, d )
Input:  X_1, X_2, ..., X_k, d
Output: U, V, Y_1, Y_2, ..., Y_k
Begin
0.  mean = (1/k) Σ_{i=1}^{k} X_i
1.  For i from 1 to k do begin
2.          X̃_i = X_i − mean
3.  End
4.  U_0 = (I_d, 0)^T ,  j = 0 ,  RMSE_i = ∞ .
5.  Do
6.          Form the matrix M_V = Σ_{i=1}^{k} X̃_i^T U_j U_j^T X̃_i
7.          Compute the d eigenvectors {φ_i^V}_{i=1}^{d} of M_V
                corresponding to the largest d eigenvalues
8.          j = j + 1
9.          V_j = [φ_1^V, φ_2^V, ..., φ_d^V]
10.         Form the matrix M_U = Σ_{i=1}^{k} X̃_i V_j V_j^T X̃_i^T
11.         Compute the d eigenvectors {φ_i^U}_{i=1}^{d} of M_U
                corresponding to the largest d eigenvalues
12.         V_j = [φ_1^U, φ_2^U, ..., φ_d^U]
13.         RMSE(j) = sqrt( (1/k) Σ_{i=1}^{k} || X̃_i − U_j U_j^T X̃_i V_j^T V_j ||^2 )
14. Until ( RMSE(j−1) − RMSE(j) ≤ η )
15. U = U_j, V = V_j
16. For i from 1 to k do begin
17.         Y_i = U^T X̃_i V
18. End
19. Return ( U, V, Y_1, Y_2, ..., Y_N )
20. End
```

Since the size of the transform results depends on d, by carefully selecting the value of d the compressed data rate can be controlled. We tested the algorithm to compress a video frame and found that this kind of transformation has a good energy compaction effect. As shown in Figure 1, the energy in the transformed matrix centralizes in the upper-left corner.

(a) (b)

Fig. 1. Transform a video frame using tensor-PCA. (a) is the original image, (b) is the matrix resulting from the transform. It can be easily seen that energy concentrate on the upper-left corner.

At the same time, tensor-PCA dramatically reduces the computational cost of traditional PCA. The time complexity of the algorithm is *O(InNd)* where I is the number of iterations, n is the number of images in a group to be coded, N is the number of pixels in a single image (rows * columns), d is the rank of the transform matrix. For PCA, the time complexity for diagonalizing the large matrix itself is $O(n^3N^3)$, where n is the number of images in a group to be coded, N is the number of pixels in a single image. In terms of spatial complexity, tensor-GPCA is also superior to PCA, since it does not need to construct a large matrix of size $n \times N$. For n images of size 800×600, the matrix size for PCA computing would be $480{,}000 \times n$. That certainly is unacceptable for today's ordinary computational power.

3 Incorporation with Video Coding

With the time and spatial efficiency of tensor-PCA, it has the potential to be used in video coding. As shown in Figure 2, Traditional hybrid video coding has a common structure of motion compensation, transform coding, quantization and entropy coding []. In this framework, transform coding is applied to motion prediction errors, which are matrices with small values. Because of its low cost of time and space in computation, tensor-PCA algorithm, as a transform coding method, can be incorporated into video coding by replacing traditional DCT.

The algorithm described in the last section projects a group of matrices into a low dimensional tensor space. Thus we need to represent the prediction error with a group of equal-sized matrices. Fortunately, the prediction error is already matrices after motion compensation which is of the same size with the video frames. We just need to take the error images in a GOP (Group of Picture) as the input to our transform

algorithm. But we have other options: 1) dividing each prediction error image into sub-blocks, and encoding each error image separately with all the sub-blocks in the image as input matrices; 2) dividing each prediction error image into sub-blocks, and organizing the sub-blocks in the same position of an image into a group to form the input of tensor-PCA transform. The three matrix representing methods are shown in Figure 3. The sub-block dividing operation reduces the size of the matrices, which can further reduce the computational cost.

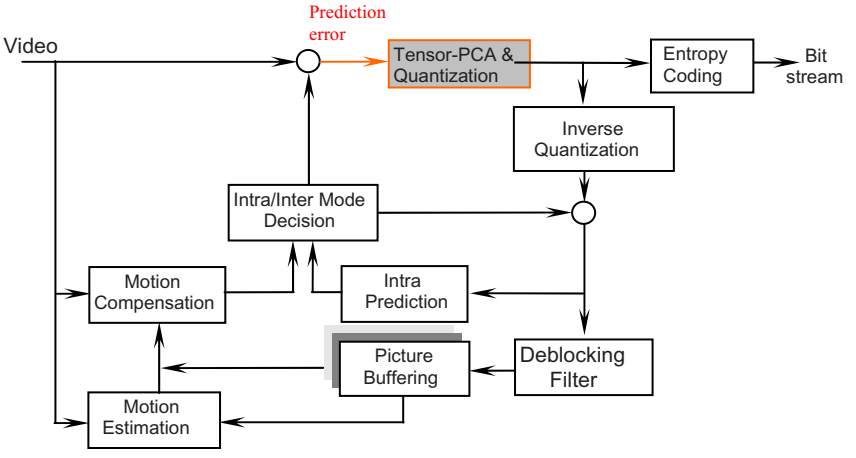


Fig. 2. The block diagram of hybrid video coding process, where traditional DCT transform can be replaced by tensor-PCA, as shown by the grey block

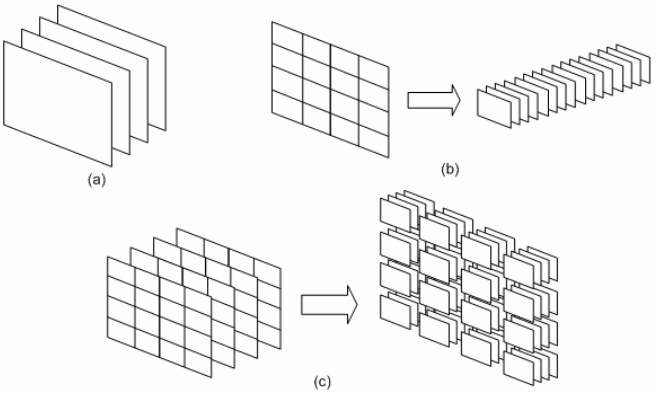


Fig. 3. Three different ways of representing prediction error images with groups of matrices. (a) Use the original error images in a GOP. (b) Divide each error image into sub-blocks and form a group of matrices. (c) Divide each error image into sub-blocks and organize the sub-blocks in the same position of an image into a group (16 groups for a GOP in the example).

Experiments show that the method shown in Figure 3 (c) gets the best performance.

4 Experiments

We experimented the tensor-PCA algorithm on encoding the video sequence "foreman" in comparison with DCT and the GPCA method in [8]. We use JSVM 2.0 H.264/AVC compatible coding mode for motion compensation, because it generates a whole frame of prediction error image instead of image blocks. Then we try DCT, GPCA and tensor-PCA on the prediction error images. The resulting data is then quantized and entropy coded into a data stream. By keeping different amount of transform coefficients, we get video streams with different bit rate. Figure 3 compares the Y-PSNR (PSNR of luma color channel) of the three coding schemes at different bit rate.

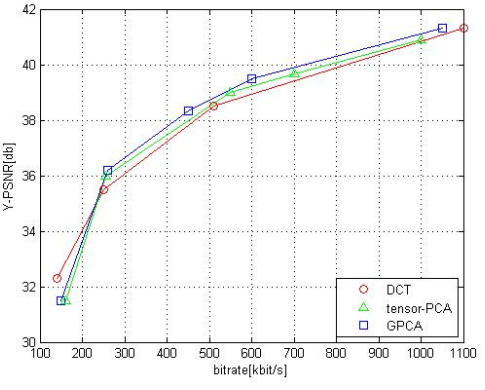


Fig. 4. PSNR of coded video sequence "foreman" by DCT, GPCA and tensor-PCA, respectively

The three different ways of matrix representation of prediction error images mentioned in Section 3 are also tested in the experiment, which shows the approach shown in Figure.3(c) is superior to the others.

5 Conclusion

In this paper, we propose to incorporate tensor-PCA into the video coding framework. Tensor-PCA is a new dimension reduction technology which projects matrices in to a space spanned by the tensor product of two vector spaces. This technology has been proved to be superior to traditional PCA in terms of computational complexity and compression quality. We incorporated the approach into hybrid video coding framework, and showed that it can achieve better compression efficiency than DCT while can be acceptable in computational cost.

Acknowledgments. This work is supported by National Natural Science Foundation of China (No.60533090, No.60525108), Science and Technology Project of Zhejiang Province (2006C13097 2005C11001-05) and China-US Million Book Digital Library Project (www.cadal.zju.edu.cn).

References

[1] Sadka, A.: Compressed Video Communications, John Wiley & Sons Ltd ISBNs: 0-470-84312-8 (Hardback); 0-470-84671-2 (Electronic)
[2] Liu, J., Zhuang, Y., Yao, L., Wu, F.: A Novel Scalable Texture Video Coding Scheme with GPCA. In: ICASSP (2007)
[3] Ye, J., Janardan, R., Li, Q.: GPCA: An Efficient Dimension Reduction Scheme for Image Compression and Retrieval. In: KDD'04, August 22–25, Seattle, Washington, USA (2004)
[4] Bursalioglu, O., Tuncel, E.: On Optimal Transforms and Quantization Schemes in Gaussian Distributed Source Coding. In: Proceedings. International Symposium on Information Theory, 10.1109/ISIT.2005.1523575 (2005)
[5] Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley, Longman Publishing Co., Inc, Reading ISBN:0201508036
[6] Vidal, R., Ma, Y., Sastry, S.: Generalized Principal Component Analysis (GPCA). IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12) (December 2005)
[7] Goyal, V.K.: Theoretical Foundations of Transform Coding, IEEE Signal Processing Magazine,1053-5888/01
[8] Hong, W., Wright, J., Huang, K., Ma, Y.: A Multi-Scale Hybrid Linear Model for Lossy Image Representation. In: ICCV (2005)

On the Robustness of Parametric Watermarking of Speech*

Aparna Gurijala and J.R. Deller Jr.

Michigan State University, East Lansing, MI 48824, USA
{gurijala,deller}@egr.msu.edu
http://www.egr.msu.edu/∼deller

Abstract. Parameter-embedded watermarking is effected through slight perturbations of parametric models of some deeply-integrated dynamics of a signal. This paper is concerned with particular model form, linear prediction (LP), which is naturally suited to the application of interest, speech watermarking. The focus of this paper is on the robustness performance of LP-embedded speech watermarking. It is shown that the technique is quite robust to a wide array of attacks including noise addition, cropping, compression, filtering, and others. In the LP formulation, a set-theoretic adjunct to the parameter embedding can be used to identify a watermark that is optimally robust against certain attacks, within a quantified fidelity constraint.

1 Parametric Speech Watermarking

General Parametric Embedding. The general parametric watermarking algorithm is formulated in the following way. Let $\{y_n\}$ denote the coversignal, and let $\{\tilde{y}_n\}$ be the ultimate stegosignal. Each of these is assumed to be a real scalar sequence over discrete-time n. It is assumed that the signals are generated according to operations of the form [2] $y_n = \phi_\pi(\xi_n, x_n, n)$ and $\tilde{y}_n = \Phi_{\tilde{\pi}}(\tilde{\xi}_n, \tilde{x}_n, n)$, in which $\{\xi_n\}$, $\{\tilde{\xi}_n\}$, $\{x_n\}$, and $\{\tilde{x}_n\}$ are measurable vector-valued random sequences. The operator ϕ is parameterized by a set π, the alteration of which (to create parameter set $\tilde{\pi}$) is responsible for changing the operator ϕ to Φ and the sequences $\{\xi_n\}$ and $\{x_n\}$ into their "tilded" counterparts.

LP Embedding. In the present study, the coversignal is assumed to be generated by a linear prediction (LP) model,

$$y_n = \sum_{i=1}^{M} a_i y_{n-i} + \xi_n, \tag{1}$$

a special case of the general formulation above. LP analysis of a (long) frame selected for watermarking [3]. The sequence $\{\xi_n\}$ is the prediction residual associated with the estimated model. The duration of the FIR linear predictor is

* This work was supported by the National Science Foundation of the United States under Cooperative Agreement No. IIS-9817485.

N. Sebe, Y. Liu, and Y. Zhuang (Eds.): MCAM 2007, LNCS 4577, pp. 501–510, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Table 1. Watermark embedding and recovery algorithms

Watermark embedding Let $\{y_n\}_{n=-\infty}^{\infty}$ denote a coversignal, and let $\{y_n\}_{n=n_k}^{n_k'}$ be the k^{th} of K speech frames to be watermarked. Then, for $k = 1, 2, \ldots, K$:

1. Using the "autocorrelation method" (e.g., [4, Ch. 5]), derive a set of LP coefficients of order M, say $\{a_i\}_{i=1}^{M}$, for the given frame.
2. Use the LP parameters in an *inverse filter* configuration to obtain the prediction residual on the frame, $\left\{\xi_n = y_n - \sum_{i=1}^{M} a_i y_{n-i}\right\}_{n=n_k}^{n_k'}$.
3. Modify the LP parameters in some predetermined way to produce a new set, say $\{\tilde{a}_i\}_{i=1}^{M}$. The modifications to the LP parameters (or, equivalently, to the autocorrelation sequence or line spectrum pairs, etc.) comprise the watermark.
4. Use the modified LP parameters as a (suboptimal) predictor of the original sequence, adding the residual obtained in Step 2 above at each n, to resynthesize the speech over the frame, $\left\{\tilde{y}_n = \sum_{i=1}^{M} \tilde{a}_i y_{n-i} + \xi_n\right\}_{n=n_k}^{n_k'}$.
5. The sequence $\{\tilde{y}_n\}_{n_k}^{n_k'}$ is the k^{th} frame of the watermarked speech (stegosignal).

Next k.

Watermark recovery For $k = 1, 2, \ldots, K$:

1. Subtract residual frame $\{\xi_n\}_{n_k}^{n_k'}$ from the stegosignal frame $\{\tilde{y}_n\}_{n_k}^{n_k'}$. This results in an estimate of the modified predicted speech, $\{d_n = \tilde{y_n} - \xi_n\}_{n_k}^{n_k'}$.
2. Estimate the *modified* LP coefficients $\{\tilde{a}_i\}_1^M$ by computing the least-square-error solution, say $\{\hat{\tilde{a}}_i\}_1^M$, to the overdetermined system of equations: $d_n \approx \sum_{i=1}^{M} \alpha_i y_{n-i}, \quad n = n_k, \ldots, n_k'$ [requires solution of *normal equations*]
3. Use the parameter estimates from Step 2 to derive the watermark values.

Next k

naturally based on the assumed order of the LP model, M, used to initially parameterize the speech. The stegosignal is constructed using the FIR filter model

$$\tilde{y}_n = \sum_{i=1}^{M} \tilde{a}_i y_{n-i} + \xi_n, \tag{2}$$

where $\{\tilde{a}_i\}$ represents a deliberately perturbed version of the "true" set $\{a_i\}$. The algorithmic steps of the LP parameter-embedded watermarking procedure appear in Table 1. There are numerous ways in which parametric modification can be effected [1]. *For simplicity, we shall refer to "LP watermarking" throughout the paper to indicate the parameter-embedded method.*

Equation (2) can be rewritten as

$$\tilde{y}_n = \sum_{i=1}^{M} a_i y_{n-i} + \sum_{i=1}^{M} \omega_i y_{n-i} + \xi_n, \tag{3}$$

wherein the watermark *sequence* (or set of watermark *coefficients*) $\{\omega_i\}_{i=1}^{M}$ is interpretable as the impulse response of an M^{th}-order non-recursive filter. The

filtered version of original speech incorporates the watermark information. The watermark *signal*, $w_n = \sum_{i=1}^{M} \omega_i y_{n-i} = \tilde{y}_n - y_n$ has a spectrum determined by the watermark coefficients and the coversignal. For example, the watermark can be designed to have predominantly lowpass, highpass or mid-band energy.

Dissimilarly to the way in which the LP model is employed in conventional speech coding and recognition, here it is used as a device to parameterize long intervals of generally nonstationary speech *without the intention of properly modeling stationary dynamics in the waveform.* Rather, the parameters are derived according to the usual optimization criterion – to minimize the total energy in the residual [4, Ch. 5], but the aggregate time-varying dynamics will be distributed between the long-term parametric code and the residual sequence.

The algorithm for recovering the watermark from the stegosignal also appears in Table 1. The LSE solution method is based on time averages, and its performance depends on the frame length used in the estimation [5].

2 Assessment of Watermarking Strategies

The quality of watermarking strategies is assessed with respect to several interrelated properties and measures. The *security* of a watermark refers to its ability to withstand operations designed to remove or detect authentic watermarks, or to embed bogus marks – in spite of the public availability of the watermarking algorithm. LP watermarks have been argued to be highly secure due to the deep integration of the information, and the diversity of placement, size, energy, and type of information [3].

The *data payload* of a watermarking algorithm is a measure of the number of bits of watermark data that can be embedded per second of stegosignal in the context of certain fidelity constraints. Payload is application dependent, but LP-embedding has shown the potential to deliver 10 - 30% more payload than state-of-the-art algorithms, with the added ability to tradeoff fidelity and robustness at high payload levels [2].

The existence of a watermark is ascertained by either simple *detection*, or *complete recovery* of the embedded information. Very successful detection strategies for LP watermarks have been reported for cases of white and colored noise corruption of the stegosignal [6]. The performance gains result from the inherent noise suppression that occurs between the time and parameter-domain manifestations. Both detection and recovery depend on a degree of watermark robustness, and the remainder of the work is devoted to this issue.

3 Robustness Studies

3.1 Introduction

The *robustness* of a watermark refers to the degree to which it remains viable for its intended purpose in spite of deliberate or systemic distortions of the stegosignal. Deliberate alteration of a stegosignal with the explicit purpose of

removing or destroying the watermark(s) is called an *attack*. We shall also use this term to refer to systemic alterations of the signal (e.g., compression) which do not have malicious intent. A watermark must be robust to distortion that does not compromise the fidelity of the content beyond an application-dependent standard. Some of the factors affecting the robustness of the present technique include the length of the speech frame watermarked, the choice of watermark sequence, and the relative energies, temporal locations and durations of the watermarks in the stegosignal.

3.2 Analytical Tools and Methods

Studying robustness requires a means for assessing fidelity. Generally speaking, a measure of *fidelity* quantifies the degree of *perceptual* similarity between the coversignal and the stegosignal. Without human listening tests (which themselves can be quite subjective), perceptual quality must be inferred from some measure(s) of signal integrity. The *segmental signal-to-noise ratio*, which in the present application is more appropriately called the *segmental coversignal-to-watermark ratio* ($\mathrm{CWR_{seg}}$), has been shown to be a reasonably effective measure of speech fidelity [4]. The $\mathrm{CWR_{seg}}$is defined as

$$\mathrm{CWR_{seg}} = \frac{1}{K}\sum_{j=1}^{K} 10\log_{10}\left[\sum_{l=k_j-L+1}^{k_j}\frac{y_l^2}{w_l^2}\right], \tag{4}$$

where, $k_1, k_2, ..., k_K$ are the end-times for K frames over which the measure is averaged, each of which is length L. w_n is the watermark sample at time n. The computation over multiple frames tends to regularize the loud and soft portions of speech. 15 ms frames are used to compute $\mathrm{CWR_{seg}}$ in the present study.

The baseline method against which LP methods are compared is the standard *spread spectrum* (SS) approach of Cox *et al.* [7]. Speech signals were watermarked using both LP and SS algorithms under noise conditions described below. The coversignal consists of a 1 s utterance "She had your dark suit in greasy wash water all year" by a female talker, from the TIMIT database [8], sampled at 16 kHz. The sample correlation coefficient is used as the similarity measure between original and recovered watermark vectors for both techniques.

For LP watermarking, the speech was divided into eight frames of 2000 samples (0.125 s) each. The watermarks were comprised of (eight) non-binary orthogonal vectors of length eight. In each of the speech frames, a watermark vector was embedded into to the coefficients of an eighth-order LP model, resulting in a bit rate of 24 bits per second. For the LP watermarking experiments, the watermark embedding and recovery involved selective normalization [6]. The resulting stegosignal was subjected to various attacks discussed below.

For the SS algorithm, the stegosignal $\{\breve{y}_j\}_{j=1}^{N}$ was obtained by adding the watermark sequence $\{g_i\}_{i=1}^{1000}$ to the 1000 largest DCT coefficients of the coversignal of 1 s duration, $\breve{Y}_i = Y_i(1 + \lambda g_i)$, where every g_i is independently drawn from $\mathcal{N}(0, 1)$, and Y_i and $\breve{Y}_i$ are the i^{th} largest DCT coefficients of the coversignal and stegosignal, respectively. Parameter λ is adjusted to obtain a desired $\mathrm{CWR_{seg}}$.

3.3 Subjective Listening Tests

In the robustness experiments to be presented, two renditions of LP watermarking were used – one at $\text{CWR}_{\text{seg}} = 7.715$ dB, and one at 10.68 dB. LP watermarks were deemed inaudible by human listeners at these or higher CWR_{seg} [9]. For performance comparison, implementations of SS watermarking at 7.715 dB and 10.68 dB were also used. Additionally, listening tests were performed to identify CWR_{seg}'s of SS-watermarked stegosignals, whose fidelity was comparable to the 7.715 dB and 10.68 dB versions of LP watermarking. Five subjects were asked to select the SS-watermarked utterance, from a set of stegosignals with CWR_{seg}'s ranging from 1 dB to 40 dB, that sounded most similar to the 7.715 dB and the 10.68 dB LP cases. The sounds files used in the listening tests are available at [9]. Based on the subjective tests it was concluded that the 7.715 dB LP watermarked sample was perceptually similar to the 27 dB of SS sample, and the fidelity of the 10.68 dB LP signal was comparable to the 30 dB SS case. This, in itself, is significant because it demonstrates the fidelity benefits that can be achieved through LP watermarking.

3.4 Watermark Robustness

In this section, we analyze the robustness of LP watermarks to common attacks. For meaningful analysis of detection performance, it is necessary to consider stationary signal segments. That is, segments of y_n, w_n, and $\tilde{y}_n$, are assumed to be partial realizations of wide-sense stationary, ergodic random processes. Generally, speech sequences can be considered stationary across frames of much shorter

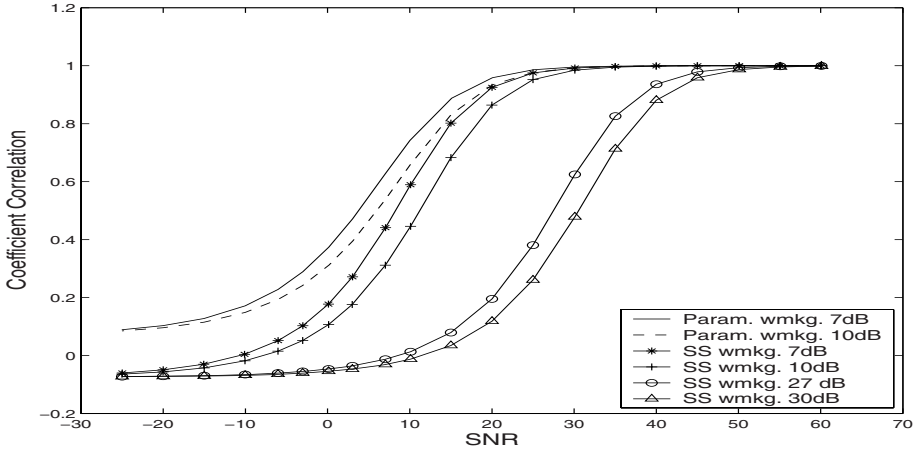


Fig. 1. Watermark robustness to white noise attack. Performance of LP watermarking at CWR_{seg}'s of 7.715 dB and 10.68 dB is compared with that of SS watermarking at 7.715 dB, 10.68 dB, 27 dB and 30 dB CWR_{seg}.

duration (typically 20 ms) than those used in the experiments (typically 125 ms) which are necessary for proper estimation. Hence, one important experimental observation is the effect of non-stationarity on watermark robustness.

Robustness to Additive Noise Attacks. Initially, we consider a simple attack in which i.i.d. white noise is added to the stegosignal. It has been shown [10] that LP watermarking is "asymptotically immune" to such attacks because the normal equations used to recover the LP watermarked parameters converge asymptotically to the "noise free" solution. In fact, for any finite frame length, the parameter estimate comprises an unbiased and consistent estimator of the true watermarked parameters.

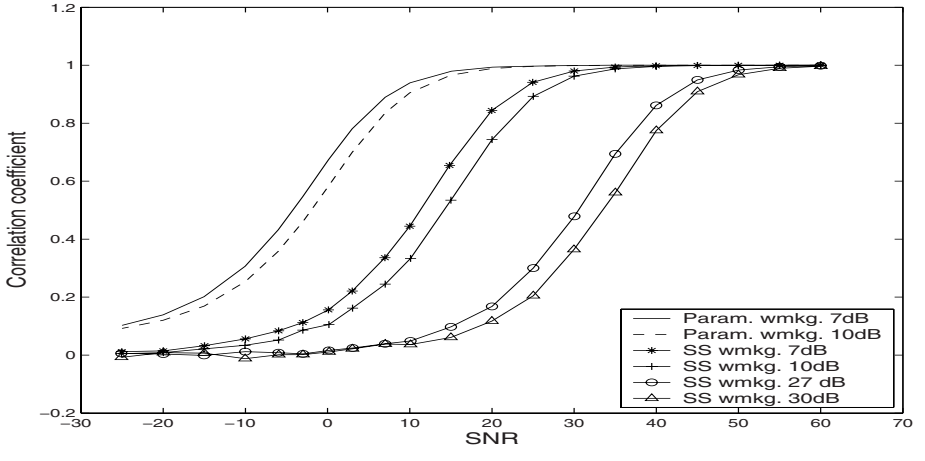


Fig. 2. Watermark robustness to colored noise attack. No prewhitening was used in obtaining these results.

To verify the analysis, experiments were performed in which white Gaussian noise was added to speech watermarked by both LP and SS algorithms. The correlation coefficients between the original and recovered watermarks from eight stegosignal frames were determined and averaged. It is seen in Fig. 1 that, at any noise level, LP watermarking at 7.715 dB and 10.68 dB CWR_{seg} results in higher correlation between original and recovered watermarks than that of SS at CWR_{seg}'s of 7.715 dB, 10.68 dB, 27 dB or 30 dB. This improvement in the correlation coefficient values, and hence robustness, is mainly due to the LSE-based recovery algorithm. This level of robustness to white noise attack is sufficient for a wide-range of applications, as the stegosignal is highly noisy below an CWR_{seg}'s of 15 dB (for details see [9]). The non-stationarity of the 2000-sample watermarked speech frame can be ignored for practical applications of LP watermarking. As expected, watermark robustness increases as the CWR_{seg} is decreased, since there is greater watermark energy in the same coversignal.

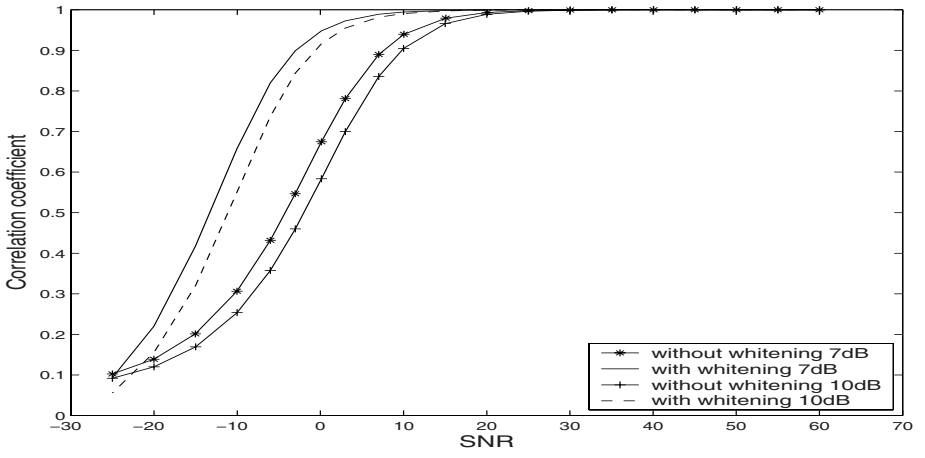


Fig. 3. Improvement in watermark robustness to colored noise attack due to whitening transformation (*c.f.* Fig. 2).

When the attack noise is *colored*, direct solution of the normal equations will produce a biased solution. In this case a prewhitening procedure can be applied to the signal using noise correlation properties which are readily determined from the data. In the "prewhitened" signal space, the solution has the same immunity to the added noise (for large N) as in the case of a white noise perturbation. In the next set of experiments, the stegosignal was distorted by colored noise which was generated by filtering a white noise process using a 11^{th}order FIR lowpass filter with a cut-off frequency of 0.4 (normalized) or 6400 Hz.

For information purposes, the colored-noise-corrupted signals were analyzed with and without the prewhitening step. Figure 2 illustrates results without prewhitening. It is observed that LP parametric watermarking is fairly robust to colored noise, even in the *absence* of prewhitening during recovery. In Fig. 2, the differences in performance between LP and SS watermarking at 7.715 and 10.68 dB is even greater than in case of white noise (Fig. 1). An improvement in watermark robustness to colored noise attack is observed in Fig. 3, where the watermark recovery process involves prewhitening. In fact, LP watermarking with prewhitening at 10.68 dB results in better robustness than at 7.715 dB without whitening, even though the latter outperforms SS at 7.715 dB.

Robustness to Filtering. Linear filtering of the stegosignal is another common, and potentially very destructive, attack on watermarks. Since filtering has direct spectral interpretations, this type of attack has many implications for fidelity, perception, and watermark frequency allocation. Generally speaking, watermark robustness to filtering can be improved through spectral diversity. For example, a watermark with a highpass spectrum can be transformed to one with lowpass energy by the simple transformation of the watermark coefficients $\omega_{\text{hp},i} = (-1)^i \omega_{\text{lp},i}$. Robustness is improved by embedding the "same" watermark in both the original and frequency-translated forms.

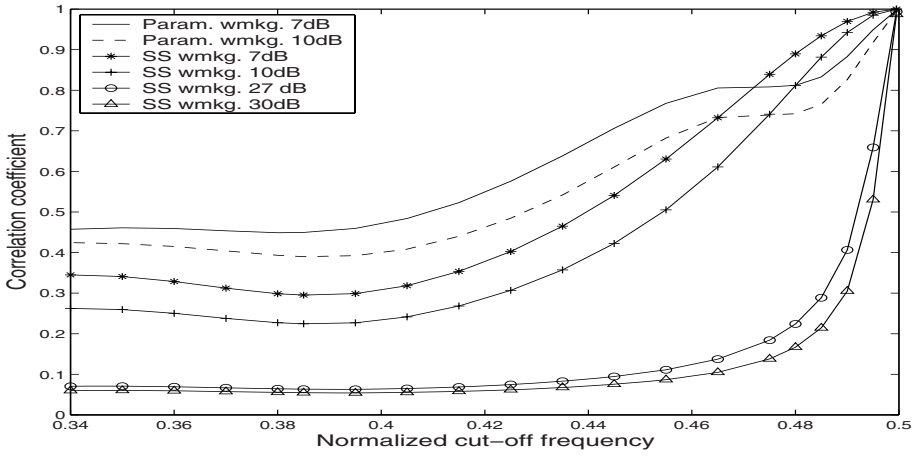


Fig. 4. Robustness to lowpass filtering

Robustness to Cropping. In a *cropping* attack, arbitrary samples of the stegosignal are removed. Since LP watermarking involves an additive operation during the embedding and recovery processes, cropping results in desynchronization of the cover- and the stegosignals. However, as the present method is an informed watermarking technique, the algorithm described in [11] can be used for resynchronization.

In the present experiment, the stegosignal was subjected to a modified version of cropping, sometimes called the *jitter attack* [12]. In this modified implementation, random samples of the stegosignal were replaced by zeros. A specified percentage of samples from each frame of 2000 samples was randomly replaced by zeros. The fact that watermark information is spread-out in the stegosignal, while it is concentrated during the recovery process, contributes to increased robustness of LP watermarking to cropping.

Robustness to Speech Compression and Coding. Experiments were performed to study the effects of low-bit-rate speech compression on LP watermarking. The stegosignal was compressed, then decompressed, and the watermarks were recovered from the decompressed signal. The correlation coefficients between the original and recovered watermarks for different speech coders are tabulated in Table 2. The attacked stegosignals are available at the website [9]. The G.711 μ law, G.726 ADPCM, GSM (13.2 k bits/s), LPC10, and CELP (4.5 k bits/s) coders were obtained from the website [13]. The G.711 speech codec software uses PCM and operates at 8 KHz, with 8-bits per sample. G.726 uses adaptive differential PCM and is widely used in VoIP application. CELP and LPC10 coders are based on parametric models of speech. At 4.5k bits per second or less, the attacked stegosignals are intelligible, but are of low fidelity [9].

Table 2 indicates that LP watermarking outperforms SS watermarking for all bit rates tested. The performance of both SS and LP watermarking is degraded significantly due to low-bit-rate CELP and GSM. Although LP watermarking

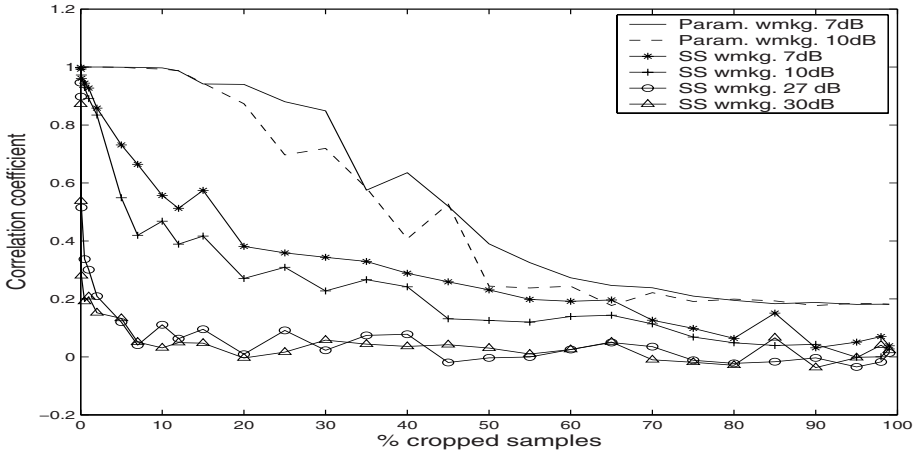


Fig. 5. Robustness to cropping. Samples of the stegosignal were randomly cropped.

Table 2. Robustness to speech coding

Speech codec	**Bit rate**	**Para wmkg 7 dB**	**Para wmkg 10 dB**	**SS wmkg 7 dB**	**SS wmkg 10 dB**
	k bits/s	**corr coef**	**corr coef**	**corr coef**	**corr coef**
G.711	64	0.9990	0.9998	0.9985	0.9966
ADPCM	32	0.9889	0.9682	0.8207	0.7658
GSM	13.2	0.6584	0.5545	0.4095	0.3140
CELP	4.5	0.1488	0.1490	−0.0225	−0.0290
CELP	2.3	0.1269	0.1464	0.0209	0.0472
LPC10	2.4	0.1762	0.1762	−0.0184	−0.0470

involves perturbation of parameters of significance to speech coding, it performs better than SS watermarking in the presence of a CELP coder. This is because LP analysis is performed over nonstationary segments of speech and synthesis is not involved in stegosignal reconstruction.

4 Summary and Conclusions

Parametric watermarking offers an alternative to methods that are strictly time or frequency based. Indeed, parametric methods are akin to more recent developments in time-frequency watermarking in that the embedding has both temporal and spectral aspects. LP watermarking is a particular form of parametric embedding which is well-suited to speech processing. LP watermarking has been shown to be quite robust to a number of common attacks, with superior performance to the benchmark SS algorithm in every case. Another interesting capability of LP watermarking is the ability to add a set-theoretic overlay which allows the

selection of optimally robust watermarks within a set meeting a ℓ_∞ signal fidelity criterion. This latter feature is described in [2,10].

References

1. A. Gurijala, *Speech Watermarking Through Parametric Modeling* (Ph.D. dissertation), Michigan State Univ., East Lansing, 2006.
2. A. Gurijala and J.R. Deller, Jr., "Speech watermarking by parametric embedding with an ℓ_∞ fidelity criterion," *Proc. EUROSPEECH* (CD-ROM), Geneva, Sep. 2003.
3. A. Gurijala, J.R. Deller, Jr., M.S. Seadle and J.H.L. Hansen, "Speech watermarking through parametric modeling," *Proc. Int'l Conf. Spoken Language Proc'g (ICSLP)* (CD-ROM), Denver, Sep. 2002.
4. J.R. Deller, Jr., J.H.L. Hansen and J.G. Proakis, *Discrete-Time Processing of Speech Signals* (2d ed.), New York: Wiley–IEEE Press, 2000.
5. S. Haykin, *Adaptive Filter Theory* (3d ed.), New York: Prentice-Hall, 1996.
6. A. Gurijala and J.R. Deller, Jr., "Detector design for parametric speech watermarking," *IEEE Int'l Conf. Multimedia & Expo (ICME)*, Amsterdam, July 2005.
7. I.J. Cox, J. Kilian, T. Leighton and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Proc'g*, vol. 6, pp. 1673-1687, Dec. 1997.
8. P.J. Price, "A database for continuous speech recognition in a 1000-word domain," *Proc. IEEE Int'l Conf. Acoustics, Speech & Signal Proc'g (ICASSP)*, New York, pp. 651-654, 1988.
9. Speech files, http://www.egr.msu.edu/~deller/ParaWmkg/WAVFILES
10. A. Gurijala and J.R. Deller, Jr., "Speech watermarking with objective fidelity and robustness criteria," *Proc. Asilomar Conf. Signals, Systems & Computers*, Pacific Grove, Calif., Nov. 2003.
11. A. Gurijala and J.R. Deller, Jr., "Robust algorithm for watermark recovery from cropped speech," *Proc. IEEE Int'l Conf. Acoustics, Speech & Signal Proc'g (ICASSP)* (CD-ROM), Salt Lake City, 2001.
12. F. Petitcolas, R.J. Anderson and M.G. Kuhn, "Attacks on copyright marking systems," *Proc. 2nd Workshop on Info. Hiding*, Portand, pp. 218-238, Apr. 1998.
13. Hawkvoice from Hawk Software, http://www.hawksoft.com/hawksoft

Author Index

Aizawa, Kiyoharu 14
Anjulan, Arasanathan 172

Bailer, Werner 305
Bak, YuHyeon 140
Berretti, S. 444
Bertini, Marco 18
Bhagavathy, Sitaram 416
Bosman, Remi 391
Byun, Hae Won 295

Canagarajah, Nishan 172
Cao, Zisheng 232
Chan H.W. 203
Chang, Chin-Chen 130, 351
Chang, Edward Y. 24
Chang, Shih-Fu 57
Chen, Chang Wen 42
Chen, Feng 232
Chen, Homer H. 52
Chen, Qiang 242
Chen, Shuangjia 325, 334, 398
Chen, Shudong 391
Chen, Yi-Hui 130
Cho, Miyoung 426
Choi, Chang 426
Choi, Deok Jae 408
Chong C.F. 203
Chou, Yung-Chen 130
Chow, K.P. 203
Chung, Ki-Dong 464
Chung, S.C. 151

de With, Peter H.N. 434
Del Bimbo, Alberto 18, 444
Deller Jr., J.R. 501
Du, Jun 223

Fu, H.C. 151
Fu, Qiang 325, 334, 398
Fu, Yun 343

Gao, Wen 37, 325, 334, 398
Guangnan, Ni 112
Guo, Rui 325, 334, 398
Gurijala, Aparna 501

Han, Jungong 434
Hou, Shuhui 252
Houari Sabirin, M.S. 483
Hu, Chengyu 262
Hu, Yuxiao 343
Huan, Thomas S. 315
Huang, Qiang 474
Huang, Qingming 223
Huang, Thomas S. 343
Huang, Tiejun 65
Huang, Yu 416
Hui, Lucas C.K. 203
Huo, Longshe 37, 325, 334, 398

Jan, Jinn-Ke 351
Jeon, Sung-Hoon 464
Jeong, Jin-Woo 184
Jiali, Zheng 112
Jiang, Shuqiang 223
Jianjun, Guo 267
Jin, Cong 454

Kim, HagYoung 140
Kim, Hanil 426
Kim, Munchurl 483
Kim, Pankoo 426
Kim, Sung-Min 464
Kui, Dai 267

Lee, CheolHoon 140
Lee, Dong-Ho 184
Lee, Guee Sang 408
Lee, Oukseh 184
Lee, Tak-Gi 464
Lee, YongJu 140
Li, Daxing 262
Li, Houqiang 42
Li, Lin 203
Li, Qiming 60
Li, Wenlong 161
Li, Zhiwei 33
Lim, Won Taek 408
Liu, Jian 493
Liu, Ju 213
Liu, Peiya 361

Liu, Pengtao 262
Liu, Ying 381
Liu, Yuncai 85, 102, 242, 371
Liu, Zhisheng 95
Llach, Joan 416
Lu, Guojun 381
Lu, Yijuan 315
Lukkien, Johan 391

Ma, Wei-Ying 33
Memon, Nasir 60
Min, OkGee 140
Ming, Zhong 474
Mingche, Lai 267
Minoh, Michihiko 252
Morimura, Yoshitaka 252

Nguyen, Toan 408

Pala, P. 444
Palanivelu, Sridharan 361
Pan, Liang-Gang 454
Pao, H.T. 151
Park, Kyung-Wook 184
Park, Mi Seon 408
Peng, Yuhua 95
Pun, K.H. 203

Qamra, Arun 24
Qi, Guo-Jun 5
Qiao, Jianping 213

Radovanovic, Igor 391
Rui, Yong 5

Schügerl, Philipp 305
Shi, Fanhuai 85
Shin, Jungpil 426
Shin, Kwang-Mu 464
Silva Mata, F.J. 444
Smeulders, Arnold W.M. 1
Smith, John R. 16
Sorschag, Robert 305
Su, Limin 195
Su, Ting 454
Sun, Junding 76
Sun, Lifeng 161

Thallinger, Georg 305
Tian, Qi 315

Tjiong, Melissa 391
Tong, Minglei 102
Tong, Xiaofeng 161
Tsang, W.W. 203

Uehara, Tetsutaro 252

Verhoeven, Richard 391

Wang, Bang-ju 277
Wang, Chung-Chuan 351
Wang, Fei 161
Wang, Fusheng 361
Wang, Jianhua 85, 371
Wang, Meiqin 203
Wang, Tao 161
Wang, Yaowei 195
Wang, Yonggang 286
Wang, Yu-hua 277
Wang, Zhang-yi 277
Wu, Fei 493
Wu, Xiaosheng 76

Xia, Tian 361
Xie, Xing 33
Xu, Wenli 232
Xu, Xun 343
Xu, Y.Y. 151

Yan, Longchuan 223
Yanfei, Shen 112
Yang, Bo 161
Yang, Jie 286
Yang, Shiqiang 161
Yao, Jason J. 52
Yao, Lei 493
Yasuai, Lv 267
Ye, Qixiang 195
Yiu, S.M. 203
Yongdong, Zhang 112
Yoo, Jae Myung 408
You, Wonsang 483
Yu, Shengyang 286

Zhang, Dengsheng 381
Zhang, Huan-guo 277
Zhang, Jing 85
Zhang, Lei 33
Zhang, Weidong 232
Zhang, Xinpeng 351

Zhang, Yan 286
Zhang, Yimin 161
Zhang, Yingxiang 242
Zhang, Zhenqiu 343
Zhiying, Wang 267
ZhongHua, Fu 122
Zhu, Yingying 474
Zhuang, Yueting 493