Jordi Vitrià
João Miguel Sanches
Mario Hernández (Eds.)

LNCS 6669

Pattern Recognition and Image Analysis

5th Iberian Conference, IbPRIA 2011
Las Palmas de Gran Canaria, Spain, June 2011
Proceedings

Springer

Lecture Notes in Computer Science 6669

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

Jordi Vitrià João Miguel Sanches
Mario Hernández (Eds.)

Pattern Recognition and Image Analysis

5th Iberian Conference, IbPRIA 2011
Las Palmas de Gran Canaria, Spain, June 8-10, 2011
Proceedings

Springer

Volume Editors

Jordi Vitrià
Universitat de Barcelona, Facultat de Matemàtiques
Departament de Matemàtica Aplicada i Anàlisi
Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain
E-mail: jordi.vitria@ub.edu

João Miguel Sanches
Institute for Systems and Robotics (ISR)
Departamento de Bioengenharia (DBioEng) / Instituto Superior Técnico
Av. Rovisco Pais, 1, 1049-001, Lisbon, Portugal
E-mail: jmrs@ist.utl.pt

Mario Hernández
University of Las Palmas de Gran Canaria
Institute for Intelligent Systems and Numerical Application in Engineering (SIANI)
Edificio de Informática y Matemáticas
Campus Universitario de Tafira
35017 Las Palmas, Spain
E-mail: mhernandez@iusiani.ulpgc.es

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-21256-7 e-ISBN 978-3-642-21257-4
DOI 10.1007/978-3-642-21257-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): I.4, I.5, I.7, I.2.10, I.2.7

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

IbPRIA 2011 (Iberian Conference on Pattern Recognition and Image Analysis) was the fifth edition of a series of similar events co-organized every two years by AERFAI (Asociación Española de Reconocimiento de Formas y Análisis de Imágenes) and APRP (Associação Portuguesa de Reconhecimento de Padrões). Previous editions took place in Andraxt, Spain (2003), Estoril, Portugal (2005), Girona, Spain (2007) and Póvoa de Varzim, Portugal (2009). In 2011 the event was held in Las Palmas de Gran Canaria, Spain, hosted by the Universidad de Las Palmas de Gran Canaria (ULPGC) and with the support of the ULPGC and ULPGC's Institute SIANI (Institute for Intelligent Systems and Numerical Applications in Engineering).

IbPRIA is a single-track conference that provides an international forum for presentation of ongoing work and new-frontier research in pattern recognition, computer vision, image processing and analysis, speech recognition and applications. It acts as a meeting point for research groups, engineers and practitioners to present recent results, algorithmic improvements, experimental evaluations and promising future directions in related topics.

The response to the call for papers was positive. From 158 full papers submitted, 92 were accepted, 34 for oral presentation and 58 as posters. A high-level international Program Committee carried out the review stage, where each submission was reviewed in a double-blind process by at least two reviewers. We are especially grateful to the Program Committee and to the additional reviewers for the effort and high quality of the reviews, which have been instrumental in preparing this book. We also thank the very valuable contribution of the authors, in answering the call and sharing their work, hopes and enthusiasm to make IbPRIA2011 a successful event.

The conference benefited from the collaboration of three invited speakers: Marcello Pelillo from the Dipartimento di Informatica of Università Ca 'Foscari di Venezia, Irfan Essa from the School of Interactive Computing, Georgia Tech University, Atlanta, USA, and Sven Dickinson from the Department of Computer Science, University of Toronto, Canada. We would like to express our sincere gratitude for their participation.

The work of the eight Area Co-chairs, two for each of the four Conference Areas, was also very valuable. Furthermore, we are very grateful to all members of the Organizing Committee. Their work led to a successful conclusion of both the conference and these proceedings.

Finally, we hope that this book will provide a broad yet comprehensive overview of the research presented at the conference, both for attendees and readers, laying the groundwork for new challenges in our area.

June 2011

J. Vitrià
J. Sanches
M. Hernández

Organization

IbPRIA 2011 was jointly organized by AERFAI (Asociación Española de Reconocimiento de Formas y Análisis de Imágenes), APRP (Associação Portuguesa de Reconhecimento de Padrões) and was locally organized by Universidad de Las Palmas de Gran Canaria through the Institute for Intelligent System and Numerical Applications in Engineering (SIANI).

General Chairs

General Co-chair AERFAI:	Jordi Vitrià, University of Barcelona, Spain
General Co-chair APRP:	João M. Sanches, Institute for Systems and Robotics (Instituto Superior Técnico), Portugal
Local Chair:	Mario Hernández, Universidad de Las Palmas de Gran Canaria, Spain

Area Co-chairs

Area 1. Computer Vision:

María Vanrell	Centre de Visió por Computador, Spain
Theo Gevers	Universiteit van Amsterdam, The Netherlands

Area 2. Pattern Recognition and Machine Learning:

Roberto Paredes	Instituto Tecnológico de Informática, Spain
Mark Girolami	University of Glasgow, UK

Area 3. Image and Signal Processing:

Jorge Salvador Marques	Universidade Técnica de Lisboa, Portugal
Rebecca Willett	Duke University, USA

Area 4. Applications:

Luis Baumela	Universidad Politécnica de Madrid, Spain
Anton Van den Hengel	University of Adelaide, Australia

Tutorial Co-chairs

Joost van de Weijer	Centre de Visió por Computador, Universitat Autònoma de Barcelona, Spain
Jordi Gonzàlez	Centre de Visió por Computador, Universitat Autònoma de Barcelona, Spain

Tutorial Speakers

Tinne Tuytelaars	ESAT-PSI, Katholieke Universiteit Leuven, Belgium
Pushmeet Kohli	Machine Learning and Perception, Microsoft Research Cambridge, UK

Invited Speakers

Marcello Pelillo	Dipartimento di Informatica of Università Ca 'Foscari di Venezia, Italy
Irfan Essa	School of Interactive Computing, Georgia Tech University, Atlanta, USA
Sven Dickinson	Department of Computer Science, University of Toronto, Canada

Local Organizing Commitee

María Dolores Afonso	Universidad de Las Palmas de Gran Canaria, Spain
Luis Antón	Universidad de Las Palmas de Gran Canaria, Spain
Gloria Bueno	Universidad de Castilla-La Mancha, Spain
Jorge Cabrera	Universidad de Las Palmas de Gran Canaria, Spain
Óscar Déniz	Universidad de Castilla-La Mancha, Spain
Antonio C. Domínguez	Universidad de Las Palmas de Gran Canaria, Spain
Ibrahim Espino	Universidad de Las Palmas de Gran Canaria, Spain
José Évora	Universidad de Las Palmas de Gran Canaria, Spain
Antonio Falcón	Universidad de Las Palmas de Gran Canaria, Spain
Cayetano Guerra	Universidad de Las Palmas de Gran Canaria, Spain
José Juan Hernández	Universidad de Las Palmas de Gran Canaria, Spain
Daniel Hernández	Universidad de Las Palmas de Gran Canaria, Spain
Josep Isern	Universidad de Las Palmas de Gran Canaria, Spain
Javier Lorenzo	Universidad de Las Palmas de Gran Canaria, Spain

Juan Méndez	Universidad de Las Palmas de Gran Canaria, Spain
Ana Plácido	Universidad de Las Palmas de Gran Canaria, Spain
Elena Sánchez	Universidad de La Laguna, Spain

Program Committee

Lourdes Agapito	Queen Mary, University of London, UK
Narendra Ahuja	University of Illinois at Urbana-Champaign, USA
José Luis Alba	Universidad de Vigo, Spain
João Barreto	University of Coimbra, Portugal
Adrien Bartoli	Université d'Auvergne, France
Alexandre Bernardino	Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal
Manuele Bicego	Università degli Studi di Verona, Italy
Gunilla Borgefors	Swedish University of Agricultural Sciences, SLU, Sweden
Hervé Bourlard	Idiap Research Institute, Switzerland
Heinrich H. Bülthoff	Max Planck Institute for Biological Cybernetics, Germany
Horst Bunke	University of Bern, Switzerland
Modesto Castrillón	Universidad de Las Palmas de Gran Canaria, Spain
Andrea Cavallaro	Queen Mary, University of London, UK
Hervé Delinguette	Institut National de Recherche en Informatique et Automatique, INRIA, France
Ahmed Elgammal	Rutgers University, USA
Tapio Elomaa	Tampere University of Technology, Finland
Bob Fisher	University of Edinburgh, UK
Daniel Gatica-Perez	Idiap Research Institute, Switzerland
Shaogang Gong	University of London, UK
Nicolás Guil	Universidad de Málaga, Spain
Edwin R. Hancock	University of York, UK
Carlos Hitoshi Morimoto	Universidade de São Paulo, Brazil
Marc Hanheide	University of Birminghan, UK
Anton van den Hengel	The University of Adelaide, Australia
Marco La Cascia	Università degli Studi di Palermo, Italy
Christoph Lampert	Max Planck Institute for Biological Cybernetics, Germany
Christian Leistner	Graz University of Technology, Austria
Stan Z. Li	Institute of Automation, Chinese Academy of Science, China
Giosue Lo Bosco	Università degli Studi di Palermo, Italy

Simone Marinai	Università degli Studi di Firenze, Italy
Aleix Martinez	The Ohio State University, USA
Luisa Micó	Universidad de Alicante, Spain
Majid Mirmehdi	University of Bristol, UK
Thomas Moeslund	Aalborg University, Denmark
Fabien Moutarde	Ecole Nationale Supérieure des Mines de Paris, France
Vittorio Murino	Università degli Studi di Verona, Italy
Hermann Ney	University of Aachen, Germany
Carlos Orrite	Universidad de Zaragoza, Spain
Maria Petrou	Imperial College London, UK
Armando Pinho	Universidade de Aveiro, Portugal
Ioannis Pitas	Aristotle University of Thessaloniki, Greece
Pedro Quelhas	Faculdade de Engenharia da Universidade do Porto, Portugal
Peter Roth	Graz University of Technology, Austria
Gabriella Sanniti di Baja	Istituto di Cibernetica CNR, Italy
Bernt Schiele	Darmstadt University of Technology, Germany
Jasjit S. Suri	University of Idaho, USA
Karl Tombre	INRIA Nancy - Grand Est Research Centre, France
Antonio Torralba	Massachusetts Institute of Technology, USA
Fernando de la Torre	Carnegie Mellon University, USA
Marko Tscherepanow	Universität Bielefeld, Germany
John K. Tsotsos	York University, Canada
Raquel Urtasun	Toyota Technological Institute at Chicago, USA
Enrique Vidal	Universidad Politécnica de Valencia, Spain
David L. Wild	University of Warwick, UK
Reyer Zwiggelaar	University of Wales, UK

Additional Reviewers

Meng Ao	Volkmar Frinken
Niklas Beuter	Sebastian Gieselmann
Jose M. Buenaposada	Jordi Gonzàlez
Jaime Cardoso	Jose González
Lewis Chuang	Michael Holte
Costas Cotsaces	Laura Igual
Marco Crocco	Emanuel Indermühle
Cristobal Curio	Jose M. Iñesta
Bruno Damas	Xingguo Li
Laura Docío	Sifei Liu
Dario Figueira	Rafael Llobet
Preben Fihl	Liliana Lo Presti
Andreas Fischer	Miguel Lourenço

Samuele Martelli
Rui Melo
Plinio Moreno
Kamal Nasrollahi
Olivier Penacchio
João Pimentel
Raghavendra Ramachandra
Jonas Ruesch
Pietro Salvagnini
Denis Schulze
Frederic Siepmann
Sabato Marco Siniscalchi
Michele Stoppa
Matteo Taiana
Ekaterina Taralova
Diego Tosato
Laura Trutoiu
Ernest Valveny
Filippo Vella
Mauricio Villegas
Nicholas Vretos
Junjie Yan
Dong Yi
Yuanhao Yu
Zengyin Zhang
Zhiwei Zhang

Table of Contents

Oral Sessions

Computer Vision

Image Processing and Analysis

Medical Applications

Pattern Recognition

Image Processing and Analysis

Pattern Recognition

Applications

Deforming the Blurred Shape Model for Shape Description and Recognition

Jon Almazán, Ernest Valveny, and Alicia Fornés

Computer Vision Center – Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
{almazan,ernest,afornes}@cvc.uab.es
http://www.cvc.uab.es/

Abstract. This paper presents a new model for the description and recognition of distorted shapes, where the image is represented by a pixel density distribution based on the Blurred Shape Model combined with a non-linear image deformation model. This leads to an adaptive structure able to capture elastic deformations in shapes. This method has been evaluated using thee different datasets where deformations are present, showing the robustness and good performance of the new model. Moreover, we show that incorporating deformation and flexibility, the new model outperforms the BSM approach when classifying shapes with high variability of appearance.

Keywords: Shape Recognition, Deformation Model, Statistical Representation.

1 Introduction

Object recognition is one of the classic problems in Computer Vision. Different visual cues can be used to describe and identify objects. Color, texture or shape are some of them, being the last one probably the most widely considered. However, shape descriptors also have to face important challenges, for instance, elastic deformations. Therefore, they should be robust enough in order to guarantee intra-class compactness and inter-class separability in the presence of deformation.

In our case, we are interested in shape descriptors that could be applied to handwriting recognition. This is one of the applications where a large variability poses a big challenge to shape descriptors. Several descriptors have been applied to this field [11]. Two well-known examples are Shape context [1] or SIFT descriptor [8]. In the particular context of hand-drawn symbol recognition, the Blurred Shape Model (BSM) [4] has been introduced as a robust descriptor to classify deformed symbols. It is based on computing the spatial distribution of shape pixels in a set of pre-defined image sub-regions taking into account the influence of neighboring regions. The use of neighborhood information permits to handle a certain degree of deformation. However, due to the rigidity of the model, it has an open problem when large deformations may cause high differences in the

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 1–8, 2011.

spatial information encoded by the BSM. For this reason, a deformable model able to capture the deformations of the shape arises as an appealing alternative.

Several deformable models can be found in the literature [5] with different characteristics to manage deformations. Some models [2,3] are based on a trade-off between forces, which leads to an energy-minimization problem. It consists in a compromise between adjusting the model to the given image, and restoring the model to the original position. Another group of models approaches to deformations as a non-linear process. An example is the Image Distortion Model (IDM) by Keysers *et al.* [6] which is based on a local pixel matching. This is the model that we have selected to be combined with the BSM because their integration into a unified framework seems a straightforward process. BSM computes a local description based on the grid representation. Therefore, the local pixel matching of the IDM can be adapted to work with this local grid-based representation.

Considering this, the main contribution of this work is the combination of the BSM descriptor with the IDM in order to build a new descriptor capable to deal with deformations. For this purpose, first, we modify the BSM grid-based representation to provide more flexibility and an easily deformation, and then, we adapt the IDM matching process to this new representation. Finally, we evaluate the proposed method using three datasets where deformations are present, comparing the new model with the original BSM.

The rest of the paper is organized as follows: Section 2 is devoted to explain the background on which the work has been based, while Section 3 explains the proposed model. Experimental results, which include comparisons that demonstrate the improved performance of the proposed method over the original approach, are conducted in Section 4. Finally, Section 5 concludes the paper.

2 Background

The background of this work has two main components. First, we are going to introduce the main characteristics of the Blurred Shape Model. And then, we are going to describe the deformation model selected to be combined with the BSM: the Image Distortion Model.

2.1 Blurred Shape Model

The main idea of the BSM descriptor [4] is to describe a given shape by a probability density function encoding the probability of pixel densities of a certain number of image sub-regions. Given a set of points forming the shape of a particular symbol, each point contributes to compute the BSM descriptor. This is done by dividing the given image in a $n \times n$ grid with equal-sized sub-regions (cells). Then, each cell receives votes from the *shape pixels* located inside its corresponding cell, but also from those located in the adjacent cells. Thereby, every pixel contributes to the density measure of its sub-region cell, and its neighboring ones. This contribution is weighted according to the distance between the point and the centroid of the cell receiving the vote. In Fig. 1 an example of the

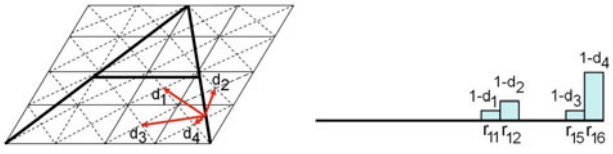


Fig. 1. BSM density estimation example. (a) Distances of a given shape pixel to the neighboring centroids. (b) Vector descriptor update using distances of (a).

contribution for a given pixel is shown. The output is a vector histogram, where each position contains the accumulated value of each sub-region, and contains the spatial distribution in the context of the sub-region and its neighbors.

2.2 Image Distortion Model

The Image Distortion Model (IDM) is introduced by Keysers *et al.* in [6]. It is a non-linear image deformation model for the task of image recognition. It consists in, given a test and a reference images, determining for each pixel in the test image the best matching pixel within a region of size $w \times w$ defined around the corresponding position in the reference image. Matching is based on the difference between a feature vector computed from the context (neighborhood) of both pixels. A final distance between two given images is simply computed summing the context differences between the mapped pixels.

Due to its simplicity and efficiency, this model has been described independently in the literature several times with different names. However, the novelty introduced by Keysers is the incorporation of pixel contexts to determine the best matching pixel. Among other possibilities, context that reported the lowest error rate in the data sets tested by Keysers is to compute the vertical and horizontal gradients in a sub-image of 3×3 around the concerned pixel. This leads to a vector of dimension $U = 18$ computed for each pixel. An example of the matching process applied to USPS digit images is shown in Fig. 2.

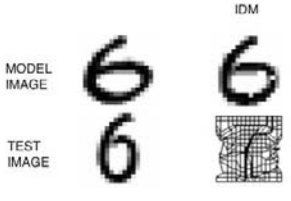


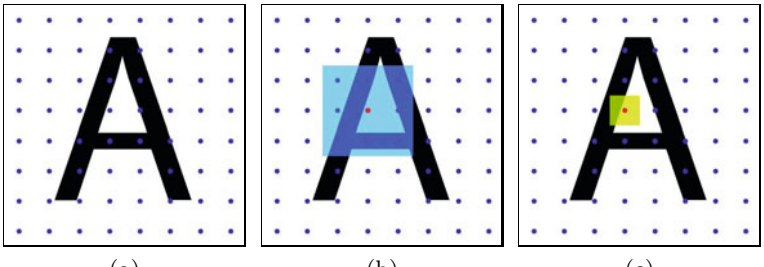
Fig. 2. Example of matching applied to the USPS digit images using the IDM. Image extracted from [6].

3 Deformable Blurred Shape Model (DBSM)

We now turn to the central problem addressed by this paper: the integration of the BSM descriptor and the IDM in a new deformable shape model. This new model will be based on deforming the grid structure of the BSM in order to adapt it to the given shape. Therefore, the first step is to modify the original grid-based representation in order to make it flexible and easily deformable (Sec. 3.1). Then, we will adapt the matching process of the IDM model using this new representation (Sec. 3.2).

3.1 DBSM Focus Representation

As it has been explained, BSM is based on placing a fixed regular grid over the image and accumulating votes of neighboring pixels on the centroid of each cell. In order to allow deformations of the grid we must adopt a slightly different representation. Instead of a regular grid of size $k \times k$ we will place over the image a set of $k \times k$ points, equidistantly distributed. These points, denoted as *focuses*, will correspond to the centroids of the original regular grid and, as in the original approach, will accumulate votes of the neighboring pixels weighted by their distance. Instead of defining the neighborhood as a set of fixed cells of the grid, it will be defined as an arbitrary *influence area* centered on the focus, in order to provide flexibility. The deformation of the grid will be obtained by moving independently each of the focuses along with their respective influence area. In order to limit the amount of deformation, each focus will be allowed to move only inside a pre-defined *deformation area*. In Fig. 3 we show an example of the focus representation and their influence and deformation areas. This resulting representation provides more flexibility and allows the focus deformation tracking.



(a) (b) (c)

Fig. 3. (a) Focuses representation. (b) Influence area. (c) Deformation area.

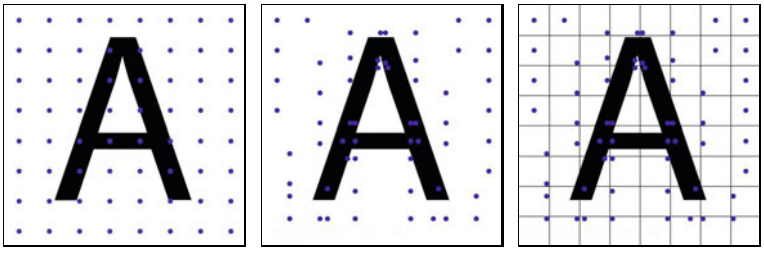
3.2 Deformation Process

Using this new representation the adaptation of the original IDM is relatively straightforward. In the IDM, every pixel in the test image is moved inside a certain region to find the best matching pixel in the reference image. In an

analog way, we will move every focus in the test image inside the deformation area to find the best matching focus in the reference image according to the BSM value obtained by accumulating votes from pixels in their respective influence areas.

The deformation model can be applied not only in the classification step to match two images, but also in the training step to adapt the initial representation of the focuses to the actual shape of the symbol. Thus, for every reference image in the training set, every focus will be moved independently inside the deformation area to maximize the accumulated BSM value. Therefore, the final position of each focus will be the local maximum of the density measure within its deformation area. Figure 4 shows an example of this process. As a result every image in the training set will be represented with two output descriptors:

- A vector histogram $\mathbf{v} \in \Re^{k^2}$ which contains the density measure of nearby pixels of each focus.
- A vector $\mathbf{p} \in \Re^{2k^2}$, which contains x and y coordinates of each focus.



(a) (b) (c)

Fig. 4. Example of the focuses deformation. (a) Initial position of the focuses. (b) Final position of the focuses after the maximization of their values. (c) Deformation area used.

Concerning the matching step, given a sample image I from the training set, and a test image J, we move the focuses in the test image inside the deformation area to optimize a certain matching criterion with the corresponding focuses in the reference image. For this purpose, we have defined two different criteria with different characteristics, whose performance will be compared experimentally.

- $\mathbf{DBSM_{min}}$: Every focus in J will be deformed in order to minimize the difference of its BSM with that of its correspondent focus of I.
- $\mathbf{DBSM_{max}}$: Analog to training, every focus in J will be moved to maximize its BSM value, independently of the value of its correspondent focus in I.

Both criterion result in two vectors ($\mathbf{v_I}$ and $\mathbf{v_J}$) containing the BSM value of the focuses of I and J, and two vectors ($\mathbf{p_I}$ and $\mathbf{p_J}$) containing the position coordinates of the focuses. Thus, after normalizing the vectors, the distance between two images is computed using the following equation:

$$distance(I, J) = d(\mathbf{v_I}, \mathbf{v_J}) \cdot \theta + d(\mathbf{p_I}, \mathbf{p_J}) \cdot (1 - \theta) \quad (1)$$

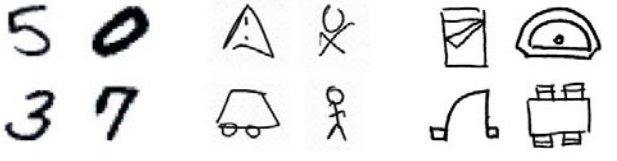
where d is the euclidean distance between two vectors, and θ is a factor for weighting the contribution.

4 Experiments

We now show experimentally the benefits of the model proposed in section 3. First, in section 4.1, we will describe the datasets that will be used in our experiments and then, in section 4.2, we will report and discuss the results obtained.

4.1 Datasets

We run our experiments in three different datasets (Fig. 5): MNIST, NicIcon and a database composed of handwritten symbols from floor plans. Following, we describe these datasets as well as the experimental protocol used in each one.



(a) (b) (c)

Fig. 5. Image samples of the three datasets. (a) MNIST. (b) NicIcon. (c) Floor plan symbol sketches.

MNIST. The MNIST [7] is a database of handwritten digits from different writers and it is divided in a training set of 60.000 examples, and a test set of 10.000 examples. The digit size is normalized and centered in a fixed-size image of 28×28 pixels. We have re-centered the digits by their bounding box, as it is reported in [7] to improve error rates.

NicIcon. NicIcon [10] is composed of 26163 handwritten symbols of 14 classes from 34 different writers. On-line and off-line data is available. The dataset is divided in three subsets (training, validation and test) for both *writer dependent* and *independent* settings. We have selected the *off-line writer dependent* configuration as benchmark to test our method. Every symbol in the off-line data has been cropped in an image of 240×240 pixels.

Floor plan symbols. This dataset [9] contains 7414 sketched symbols from architectural floor plans, divided in 50 classes, with an average of 150 symbols per class. Symbols have been centered in 128×128 fixed-size images. The experimental protocol selected is a 10-fold cross-validation.

4.2 Results and Analysis

Experiments have been done using a Nearest Neighbor classifier over the three datasets. We compare the three methods (the original BSM and the proposed DBSM using the two different criteria for matching) in terms of accuracy in classification. Before running the final experiments on the test set, we have optimized all the parameters of the different methods, which include the number of *regions/focuses*, the size of the *influence* and *deformation* areas, and the weight θ to compute the final distance, using a training set for each dataset. The final results of the experiments on the test set for each dataset are shown in Table 1.

Table 1. Accuracy rate of the compared methods in classification over the datasets selected

	DBSMmin	DBSMmax	BSM
MNIST	94'3	**94'4**	92'6
NicIcon	81'7	**82'3**	80'4
Floorplans	**99'2**	**99'2**	98'8

As we can see, DBSM outperforms the original approach when classifying in the three tested datasets. Although the results for the MNIST dataset are below the current state-of-the-art, this is due to the low accuracy of the BSM descriptor with this dataset. We can see that DBSM clearly improves the BSM approach. Furthermore, it is noteworthy that approaches with highest accuracy rate in MNIST use some pre-processing or classification methods which could considerably improve the performance of our approach. Thus, we can conclude that the integration of deformations to the fixed grid-based representation leads to a higher accuracy rate in all the databases.

Regarding both DBSM criteria, we notice that DBSMmax has a slightly higher accuracy rate, although the difference is not enough significant. However, the main advantage of the DBSMmax over DBSMmin is that the testing process is computationally faster because, given a test image, DBSMmin has to perform the matching process for every reference image in the training set. On the contrary, DBSMmax only has to run the deformation process once for every test image, obtaining a vector descriptor that is used to compare with all reference images. Moreover, this description could be used to train any classifier, and it is not only limited to the k-NN classifier as in the case of DBSMmin.

5 Conclusions

We have designed and developed a new model for shape recognition, integrating a deformable model with the Blurred Shape Model. We have shown, using three different datasets with deformations, that the resulting model is able to capture and deal with elastic distortions. Furthermore, the new model outperforms the original BSM approach in terms of accuracy rate in classification tasks.

As future work, we are going to work on improving one of the weaknesses of our model, its lack of rotation invariance. We have considered to use a different representation, instead of the rectangular grid based on the BSM. For example, a circular grid, will provide the model a way to deal with rotations. Furthermore, we have considered applying other deformation models with different characteristics or extending this work to other shape descriptors.

Acknowledgments. This work has been partially supported by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03 and CONSOLIDER-INGENIO 2010(CSD2007-00018) and by a research grant of the Universitat Autonòma de Barcelona (471-01-8/09).

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(24), 509–522 (2002)
2. Cheung, K., Yeung, D., Chin, R.: A bayesian framework for deformable pattern recognition with application to handwritten character recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(2), 1382–1387 (1998)
3. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
4. Escalera, S., Fornés, A., Pujol, O., Radeva, P., Lladós, J.: Blurred shape model for binary and grey-level symbol recognition. Pattern Recognition Letters 30, 1424–1433 (2009)
5. Jain, A.K., Zhong, Y., Dubuisson-Jolly, M.P.: Deformable template models: A review. Signal Processing 71(2), 109–129 (1998)
6. Keysers, D., Deselaers, T., Gollan, C.: Deformation models for image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(8), 1422–1435 (2007)
7. Lecun, Y., Cortes, C.: The mnist database of handwritten digits, `http://yann.lecun.com/exdb/mnist/`
8. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Sánchez, G., Valveny, E., Lladós, J., Romeu, J., Lozano, J., Mas, J., Lozano, N.: A platform to extract knowledge from graphic documents. Application to an architectural sketch understanding scenario. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 389–400. Springer, Heidelberg (2004)
10. Willems, D., Niels, R., van Gerven, M., Vuurpijl, L.: Iconic and multi-stroke gesture recognition. Pattern Recognition 42(12), 3303–3312 (2009), `http://unipen.nici.ru.nl/NicIcon/`
11. Zhang, D., Lu, G.: Review of shape representation and description techniques. Pattern recognition 37(1), 1–19 (2004)

A Shortest Path Approach for Vibrating Line Detection and Tracking

Pedro Carvalho, Miguel Pinheiro, Jaime S. Cardoso, and Luís Corte-Real

INESC Porto, Faculdade de Engenharia da Universidade do Porto,
Campus da FEUP, Rua Dr. Roberto Frias, n 378, 4200 - 465 Porto, Portugal
{pedro.carvalho,jaime.cardoso,lreal}@inescporto.pt,
miguel.amavel@gmail.com
http://www.inescporto.pt

Abstract. This paper describes an approach based on the shortest path method for the detection and tracking of vibrating lines. The detection and tracking of vibrating structures, such as lines and cables, is of great importance in areas such as civil engineering, but the specificities of these scenarios make it a hard problem to tackle. We propose a two-step approach consisting of line detection and subsequent tracking. The automatic detection of the lines avoids manual initialization - a typical problem of these scenarios - and favors tracking. The additional information provided by the line detection enables the improvement of existing algorithms and extends their application to a larger set of scenarios.

Keywords: Computer vision, vibrating lines, detection, tracking.

1 Introduction

The monitoring of vibrating structures is important in several areas including mechanics and civil engineering. The observation of structures like cable-stayed bridges, antennas or electricity distribution systems is vital for the extension of their life-span and for the reduction of system failures and consequent costs. Past approaches made use of devices, such as accelerometers or load cells (devices which translate force into an electrical signal), placed at the structures to be monitored [5,2,7]. Although the use of such techniques has produced accurate results, the installation of this type of equipment may easily become very expensive. Moreover, in many cases such as antennas or even some cable-stayed bridges it may not be possible to install such devices. Hence, there is a desire to use Computer Vision (CV) techniques to provide a non-invasive alternative. Despite the existence of past work, the natural difficulties of these scenarios pose serious problems. These include: cameras placed at large distances; oscillations in the camera; illumination changes; adverse weather conditions; reduced thickness of the structures and minimal displacements in the image. Line detection specifically targeted for video sequences is by itself a complex problem, also being tackled in other contexts [8,4,3].

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 9–16, 2011.

The characteristics of the lines or cables, namely their reduced thickness and lack of texture information, forced common CV approaches to this problem to introduce restrictions and assumptions. Difficulties in identifying the structures to be tracked typically require a user to mark a point or region of interest (ROI), for example on the line or cable themselves or on their image representation [6,10,9]. Displacement of the camera or occlusion situations may require a new marking of the points. Also, scenarios where the projection of the cables into the image plane causes the cables to be close together may originate problems as the vibrations may lead to occlusion between them.

We propose to use automatic line detection to augment the flexibility and robustness of vibrating lines tracking algorithms. We argue that the ability to automatically detect the objects (i.e., the lines) to be tracked may increase the reliability of tracking using state-of-the-art algorithms by providing more information. By enabling the automatic selection of an arbitrary number of points over the line to be tracked, it is possible to make a more complete characterization for any point and not only for those manually marked during the initialization. Following this reasoning we divided the description of the proposed method into line detection and tracking.

2 Line Detection and Tracking

The proposal described in this paper consists of two steps: detection of the lines; tracking of one or multiple points over the line. A vibrating line can be considered as a connected path between the two lateral margins of the image (for simplicity, the presentation will be oriented for horizontal lines only; the necessary adaptations for the detection of vertical lines should be clear at the end). As vibrating lines are almost the only extensive objects on the image, lines can then be found among the shortest paths between the two margins of the image if paths through high gradient positions are favored. Vibrating lines are then best modeled as paths between two regions Ω_1 and Ω_2, the left and right margins of the image. These same ideas have been successfully applied to other applications [4,3].

2.1 Line Detection Using a Shortest Path Approach

The method considers the image as a graph, where pixels are the nodes and a cost is associated with each edge connecting 8-neighbourhood pixels.

One may assume that vibrating lines do not zigzag back and forth, left and right. Therefore, one may restrict the search among connected paths containing one, and only one, pixel in each column of the image[1]. Formally, let I be an $N_1 \times N_2$ image and define an admissible line to be

$$\mathbf{s} = \{(x, y(x))\}_{x=1}^{N_1} \text{ , s.t. } \forall x \ |y(x) - y(x-1)| \leq 1,$$

[1] These assumptions, 8-connectivity and one pixel per column, impose a maximum detectable 45 rotation degree.

where y is a mapping $y : [1, \cdots, N_1] \rightarrow [1, \cdots, N_2]$. That is, a vibrating line is an 8-connected path of pixels in the image from left to right, containing one, and only one, pixel in each column of the image.

Given the weight function $w(p, q)$ defined over neigbouring pixels p and q, the cost of a line can be defined as $C(\mathbf{s}) = \sum_{i=2}^{N_1} w(v_{i-1}, v_i)$. The optimal vibrating line that minimizes this cost can be found using dynamic programming. The first step is to traverse the image from the second column to the last column and compute the cumulative minimum cost C for all possible connected vibrating lines for each entry (i, j):

$$C(i,j) = \min \begin{cases} C(i-1,j-1) + w(p_{i-1,j-1}; p_{i,j}) \\ \quad C(i-1,j) + w(p_{i-1,j}; p_{i,j}) \\ C(i-1,j+1) + w(p_{i-1,j+1}; p_{i,j}) \end{cases} . \tag{1}$$

where $w(p_{i,j}; p_{l,m})$ represents the weight of the edge incident with pixels at positions (i, j) and (l, m). At the end of this process,

$$\min_{j \in \{1, \cdots, N_2\}} C(N_1, j)$$

indicates the end of the minimal connected line. Hence, in the second step, one backtrack from this minimum entry on C to find the path of the optimal line.

Assume one wants to find all vibrating lines present in an image. This can be approached by successively finding and erasing the shortest path from the left to the right margin of the image. The erase operation is required to ensure that a line is not detected multiple times.

To stop the iterative vibrating lines search, the method receives from the user the number of lines to be detected and tracked. Since this can be done once for the entire sequence, it is a very reasonable solution.

Design of the Weight Function. The weight function on the edges was defined so that the shortest path corresponds to a path that maximises the amount of edge strength in the image along the contour. An immediate approach is to support the design of the weight function solely on the values of the incident nodes, fixing the weight of an edge as a monotonically increasing function of the average gradient value of the incident pixels.

Although a more general setting could have been adopted, the weight of the edge connecting 4-neighbour pixels was expressed as an exponential law

$$f(g) = \alpha \ \exp(\beta \ (255 - g)) + \gamma, \tag{2}$$

with $\alpha, \beta, \gamma \in$ and g is the average of the gradient computed on the two incident pixels. For 8-neighbour pixels the weight was set to $\sqrt{2}$ times that value.

Detecting Vibrating Lines in Different Positions. It should be obvious now how to adapt the detection technique for lines between the superior and inferior margins. It is also possible to apply the detection technique to adjacent

margins. Suppose that the endpoints of the vibrating lines are in the left and top margins. A difficulty with searching for the shortest paths (shortest in the sense on minimizing the cost of the path) between the top row and left column is that small paths, near the top-left corner are naturally favored. To overcome this challenge, we propose to pre-process the image, adopting the polar coordinates. The left column is mapped to the bottom row (corresponding to an angle of $\pi/2$ rads) and the top row stays in the same position. In this new coordinate-system, the path to search for is now between the top and bottom rows.

A different setting is when one of the endpoints of the vibrating lines is on the left OR the top margin and the other endpoint is in the bottom OR right margins. It is still possible to pre-process the image with an appropriate transform before the application of the proposed detection technique. The goal of the transformation is to place the endpoints in opposing margins. Although different solutions exist, we started by rotating the image followed by a linear scaling to transform the oblique lines into vertical ones.

Finally, it is worth mentioning that the vibrating line detection can be applied in a user defined region in the image (and in the video sequence). That can be important when other extensive objects are visible in the image.

2.2 Line Tracking

Many line tracking methods use optical flow to compute the displacement vector of one or more points over the line to be tracked. The typical lack of texture information over the line makes tracking methods that require this information unsuitable for these scenarios. Optical flow has limitations concerning occlusion situations and the deformation of objects, but given the characteristics of the scenarios one can assume that such restrictions are respected. Nevertheless, the aperture problem is present making it harder to reliably compute the tangential component of the displacement. We argue that using more points over the line and their corresponding displacements can minimize tracking errors, while contributing to a more complete characterization of the line behavior.

State-of-the-art optical flow methods were considered and experiments performed to assess the most adequate. Each method was applied to sequences containing vibrating lines and compared to the corresponding reference information. It was observed that the Pyramidal Lucas-Kanade [1] outperformed the others. Hence, this was the method used in the subsequent experiments.

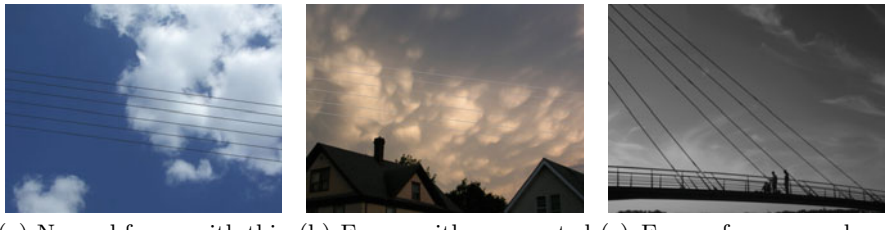
The proposed tracking approach consists of the following. On each frame, the lines are detected using the described shortest path method. For any point over the line a window of length L, also over the line, is considered and the median of the displacement vectors for every point in the window computed.

3 Validation of the Proposed Approach

The capture of sequences for the target scenarios is not a simple task due to aspects such as location or legal implications. Hence, there is a lack of appropriate sequences and of the corresponding reference information.

3.1 Dataset

To aid the validation of the proposal described in this paper a set of synthetic sequences with appropriate characteristics and the corresponding reference information was generated. The images consist of six lines, each vibrating with a given frequency and maximum amplitude. Different characteristics for the lines were also simulated, namely: thickness; frequency; amplitude; curvature radius. A background was added to approximate a real situation. The dataset also includes a real sequence of a typical scenario captured and provided by a civil engineering research institute. A set of frames containing reference information were manually generated for the real sequence for the purpose of assessing line detection. Sequence 1 and 2 feature curved lines and sequences 3 and 4 straight ones. The lines in sequence 1 and 3 are thinner. The background in sequence 5 is noisier to augment the difficulty in detection and tracking. Sequence 6 consists of images captured in a real scenario. Fig. 1 depicts an example of synthetic and natural images of the sequences.



(a) Normal frame with thin lines (b) Frame with augmented difficulty (c) Frame from a real sequence

Fig. 1. Examples of images from the dataset

3.2 Line Detection Assessment

As previously described, the shortest path method was applied over the gradient of the image. Experiments in the gradient computation over different color spaces were conducted and it was verified that the use of the grayscale space was not adequate causing large gaps in the edges corresponding to the lines. The best results were obtained using the HSV color space and in particular the Saturation channel provided robust information.

The evaluation of the line detection approach was accomplished by measuring the distance between reference and detected lines. For each possible pair, the distance is computed and correspondences determined using the Hungarian algorithm. In the distance calculation, for each point in the reference line, the distance to each point in the detected line is computed and two overall measures are taken: the average and the Hausdorff distance. For an image with multiple lines, the average of the line errors and the Hausdorff distance are taken. Similarly, for a sequence, the same process applies over the image values.

3.3 Tracking Assessment

For the evaluation of the tracking, the results of two tracking approaches, one based on measuring the displacement vectors in a point over the line and the one proposed, are compared to the reference information by measuring the euclidean distance. For each mark (reference point in one line), the root mean square error over the sequence is computed and the average over the number of lines to be tracked in the sequence is calculated. For the proposed approach, different window sizes were considered. Sequences 3 to 5 were use in the evaluation due to the possibility of using an arbitrary number of reference points.

4 Results

Table 1 presents the results for line detection, using shortest paths, for the dataset. As stated, the average and Hausdorff distance were calculated and normalized by the diagonal size of the image.

Table 1. Results for line detection normalized by the image diagonal

	Average Distance	**Hausdorff Distance**
Seq. 1	0.13%	0.13%
Seq. 2	0.13%	0.15%
Seq. 3	0.13%	0.15%
Seq. 4	0.14%	0.15%
Seq. 5	0.20%	1.21%
Seq. 6	0.04%	0.07%

One can observe that the errors obtained are small and for sequences 1 to 4 they are nearly the same. These sequences comprise the same background with changes only in the characteristics of the lines. The error is greater for sequence 5 since it presents a higher level of difficulty in both the noisy background and lack of distinctiveness of the lines. The best performance is achieved for the real sequence, as the distinctiveness of the lines is greater.

Table 2 presents the results for the evaluation of the tracking. For each line in a sequence, the root mean square error relatively to the reference at time instant t was calculated and the average over the lines in the sequence determined. Tracking using 1 point corresponds to using a state-of-the-art algorithm with the user marking the interest point in the image. The error gain in percentage relatively to the 1 pixel approach was calculated for the displacement vectors and its individual components (vertical and horizontal). Tracking results taking into account all the visible line are considered, but care must be taken in these situations since a high degree of curvature may induce errors.

The use of more points over the line enables a decrease in the RMSE. A considerable part of the improvement in the line tracking is due to a better

approximation of the horizontal component since, in the considered scenarios, it is the most affected by the aperture problem. For different line positions the changes are straightforward.

The frequencies of vibrations for each line were also computed and compared to the reference values. The results are presented in Table 3.

Table 2. Results for line tracking

		Window Size					
		1 points	3 points	11 points	101 points	301 points	all line
	Direction	**RMSE**	**Error gain relatively to the 1 point approach (%)**				
Seq. 3	horizontal	1,22	-2,45	-13,72	-54,52	-87,16	-90,72
	vertical	0,79	0,03	4,32	-7,10	-39,01	-60,26
	total	1,73	-1,39	-7,46	-35,74	-68,98	-80,48
Seq. 4	horizontal	1,85	-6,34	1,05	-44,75	-69,94	-91,23
	vertical	0,46	-2,65	-12,98	-28,38	-42,18	-29,43
	total	1,94	-5,97	-0,42	-42,59	-66,79	-80,22
Seq. 5	horizontal	0,89	-0,60	-4,45	-39,32	-61,24	-81,86
	vertical	1,16	-0,13	0,93	-15,76	-10,78	-7,43
	total	1,80	0,22	-0,79	-25,97	-31,18	-38,90

Table 3. Frequency outputs (in rad/s) from sequences 3, 4 and 5

Seq. 3		Seq. 4		Seq. 5	
Reference	Obtained	Reference	Obtained	Reference	Obtained
0,800	0,800	0,800	0,801	0,800	0,800
0,850	0,848	0,850	0,848	0,300	0,300
0,900	0,901	0,900	0,902	0,200	0,200
0,950	0,951	0,950	0,950	0,950	0,950
1,000	1,000	0,050	0,050	0,100	0,100
1,050	1,051	1,500	1,502	1,500	1,501

5 Conclusions

The use of the shortest path method enables the automatic detection of the vibrating lines, straight or curved, to be tracked with very small errors. Such automatic detection avoids the need for manual initialization of the points to be tracked. Moreover, obtained results show that using more points over the line enables a reduction of the tracking errors due to the optical flow computation.

The application of the proposed method to the monitoring of civil engineering structures can take advantage of the knowledge of the structure dimensions to perform camera calibration and obtain 3D measures of the displacement.

Future work will include the validation of the method with data captured from other devices such as accelerometers. Other forms of using the additional information provided by the lines detected are also a research topic of interest.

Acknowledgments. The authors would like to thank the Fundação para a Ciência e a Tecnologia (FCT) - Portugal - and the European Commission for financing this work through the grant SFRH/BD/31259/2006 and Fundo Social Europeu (FSE) respectively.

References

1. Bouguet, J.: Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. Microprocessor Research Labs, Intel Corporation (2000)
2. Calçada, R., Cunha, A., Delgado, R.: Analysis of traffic induced vibrations in a cable-stayed bridge. Journal of Bridge Engineering, ASCE 10(4), 370–385 (2005)
3. Cardoso, J.S., Capela, A., Rebelo, A., Guedes, C., da Costa, J.F.P.: Staff detection with stable paths. IEEE Transactions Pattern Analysis Machine Intelligence 31(6), 1134–1139 (2009)
4. Cardoso, J., Capela, A., Rebelo, A., Guedes, C.: A connected path approach for staff detection on a music score. In: Proceedings of the International Conference on Image Processing (ICIP 2008), pp. 1005–1008 (2008)
5. Magalhães, F., Caetano, E., Cunha, A.: Operational model analysis and finite element correlation of the braga stadium suspended roof. Engineering Structures 30, 1688–1698 (2008)
6. Olaszek, P.: Investigation of the dynamic characteristic of bridge structures using a computer vision method. Measurement 25, 227–236 (1999)
7. Roberts, G., Meng, X., Meo, M., Dodson, A., Cosser, E., Iuliano, E., Morris, A.: A remote bridge health monitoring system using computational simulation and GPS sensor data. In: Proceedings, 11th FIG Symposium on Deformation Measurements (2003)
8. Rodrigo, R., Shi, W., Samarabandu, J.: Energy based line detection. In: Canadian Conference on Electrical and Computer Engineering, pp. 2061–2064 (2006)
9. Silva, S., Bateira, J., Caetano, E.: Development of a vision system for vibration analysis. In: 2nd Int. Conf. on Experimental Vibration Analysis for Civil Engineering Structures (2007)
10. Wahbeh, A., Caffrey, J., Masri, S.: A vision-based approach for the direct measurement of displacements in vibrating systems. Smart Materials and Structures 12(5), 785–794 (2003)

And-Or Graph Grammar for Architectural Floor Plan Representation, Learning and Recognition. A Semantic, Structural and Hierarchical Model

Lluís-Pere de las Heras and Gemma Sánchez

Computer Vision Center, Campus UAB, 08193 Bellatera, Barcelona, Spain
{lpheras,gemma}@cvc.uab.es
http://www.cvc.uab.es/

Abstract. This paper presents a syntactic model for architectural floor plan interpretation. A stochastic image grammar over an And-Or graph is inferred to represent the hierarchical, structural and semantic relations between elements of all possible floor plans. This grammar is augmented with three different probabilistic models, learnt from a training set, to account the frequency of that relations. Then, a *Bottom-Up/Top-Down* parser with a pruning strategy has been used for floor plan recognition. For a given input, the parser generates the most probable parse graph for that document. This graph not only contains the structural and semantic relations of its elements, but also its hierarchical composition, that allows to interpret the floor plan at different levels of abstraction.

Keywords: And-Or Graph, Stochastic Grammar, Grammar Inference, Conditional Random Fields, Architectural Floor Plan Interpretation.

1 Introduction

Architectural floor plan images, see figure 1, are documents used to model the structure of the buildings and its components (windows, doors, walls, etc). In the last 15 years, floor plans interpretation has been studied for different final applications [4–7]. However, at the present time, its interpretation is a non-solved problem, essentially because there is no standard notation defined; same building components are differently modelled in distinct floor plans; and the variability in their structure with squared and rounded shaped or even "non-existent" walls.

The contribution of this paper is the design of a *generic* means to represent, recognize and validate correct floor plan documents. To do so, a grammatical formalism over an And-Or graph augmented with three probabilistic models is inferred from the dataset to represent the structural, hierarchical and semantic relations of the plan and its elements. Then, a parser has been implemented for plan recognition. The parser analyses the input plan model and classifies it as valid or not depending whether it satisfies the grammar productions. The result obtained from a valid plan is an And-graph representation of the structure, the hierarchy and the semantic composition of the document.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 17–24, 2011.

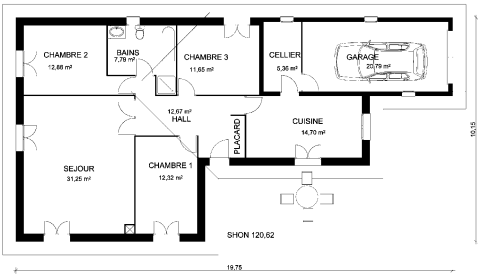


Fig. 1. Architectural floor plan document

This paper is structured as follows. In section 2 the syntactic model, a grammar over an And-Or graph, is presented together with its inference and parsing processes. Section 3 explains the three different component extraction approaches used to proof the usability of this model. Section 4 presents quantitative and qualitative results. Finally, in section 5 we conclude the overall work.

2 Syntactic Model

The contribution of this paper is the syntactic model created to model, learn and recognize floor plan documents. We divide our model in three main steps: model definition, model learning and model recognition, see figure 2. Firstly, we define how we represent the knowledge in the domain using a grammar. Secondly, the grammar inference process given a training set is explained. Finally, the plan recognition process for an input document is explicated.

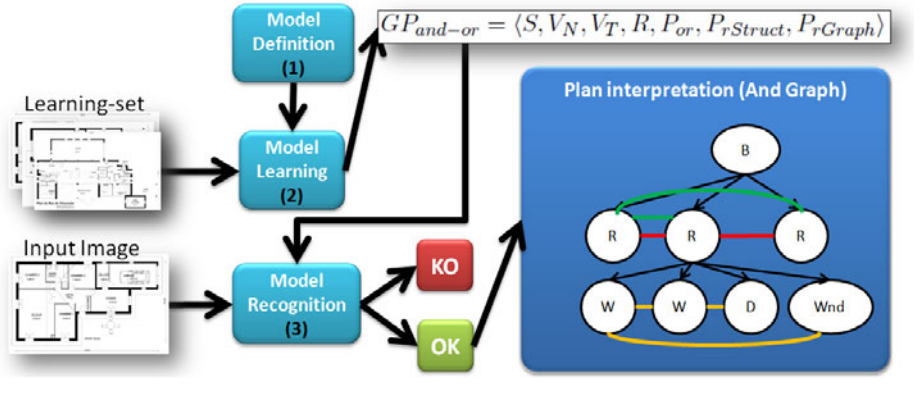


Fig. 2. Syntactic floor plan document representation and recognition

2.1 Model Definition

The hierarchical composition of the elements in a floor plan can be represented in terms of attributed tree structure, see figure 3a. In this model, a *building* is

composed by a set of *rooms*, and a *room* by a set of *walls, doors* and *windows*. To add structural and semantic information to this model, the attributed tree is augmented to an attributed graph by adding horizontal attributed edges between elements of the same abstraction level, see figure 3b. These horizontal edges represent three kinds of relations: neighbouring, accessibility and incidence. Neighbouring and incidence are structural relations while accessibility is a semantic one. Nevertheless, since this graph only represents a single plan, to represent all possible instances, the And-Or graph structure, which has been proposed by Zhu et al. [8] for natural images framework, is used. In our And-Or graph model, see figure 3c, And nodes represent the elements: *Building, Room, Wall, Door* and *Window*; while Or nodes represent the possible configurations between these elements; the unknown numbers *m,n,i* and *j*. Therefore, our And-Or graph grammar G_{and-or} describing all possible floor plans is the $4-tuple$:

$$G_{and-or} = \langle S, V_N, V_T, R \rangle, \tag{1}$$

where S is the *Building* element. V_N is the set of non-terminal symbols {*Building, Room*}, V_T is the set of terminal symbols {*wall, door, window*}. Finally, R is the set of rules that enables to expand elements into its certain components and define the hierarchical, structural and semantic relations between elements.

2.2 Model Learning

The model is learnt from a training-set composed by different architectural plans. For that purpose, one And-graph is constructed for each example. With the resulting set of And-graphs, the possibilities of the Or nodes are computed using the Maximum Likelihood Estimation algorithm (MLE). Then, the grammar G_{and-or} is augmented to an stochastic grammar G^P_{and-or} defined by the $7-tuple$:

$$G^P_{and-or} = \langle S, V_N, V_T, R, P_{or}, P_{rStruct}, P_{rGraph}, \rangle, \tag{2}$$

where S, V_N, V_T and R are the same of G_{and-or}. P_{or}, $P_{rStruct}$ and P_{rGraph} are three probabilistic models learned from the training-set. P_{or} is defined over the Or nodes of the And-Or graph to account the relative frequency of appearance of the elements normalized by the *Building* and *Room* area. In this way, P_{or} allows to model, for instance, the relative number of rooms and walls for a building of a certain size (area). $P_{rStruct}$ defines the rooms structure. Rooms are considered to be composed by *walls*, *doors* and *windows*; thus, this stochastic model considers how rooms are likely to be constructed. Finally, P_{rGraph} is a probabilistic model to add more information to room formation. It accounts the relative probability between room area and perimeter and plan area and perimeter.

2.3 Model Recognition

A *Bottom-Up/Top-Down* parser to validate and interpret floor plans has been implemented. The parser builds an And-graph representation for an input plan instance by means of a *Bottom-Up* strategy. Then, the And-graph is parsed using a *Top-Down* approach to verify whether the plan is consistent with the stochastic grammar. In that case, the correct interpretation of the plan is extracted.

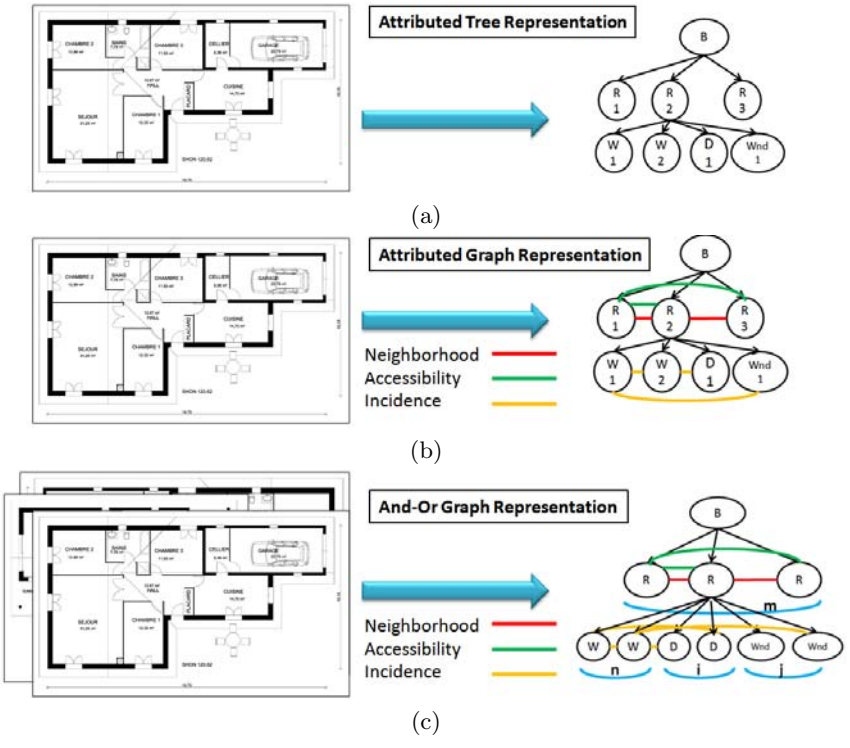


Fig. 3. Model for floor plans. B: *Building*, R: *Room*, W: *Wall*, D: *Door*, Wnd: Window. (a) Hierarchy of a single plan. (b) Hierarchy, structure and semantics of a single plan. (c) Hierarchy, structure and semantics of all possible plans.

Bottom-Up Parser Methodology

Given a floor plan, the parser builds an And-graph that possibly represents the plan following a *Bottom-Up* strategy. First, the terminal symbols $V_T = \{walls, doors, windows\}$ are extracted using the segmentation techniques presented in section 3. Then, *possible_rooms* are extracted analysing the relations between terminal symbols, explained in section 3.2. Finally, the starting symbol $S = \{Building\}$ is synthesized. At each step, the structural and semantic relations between elements are evaluated to build the higher hierarchic level. The resulting And-graph of this process is a first representation approach of the plan.

Top-Down Parser Methodology

The And-graph created is parsed using a *Top-Down* strategy for analysing whether the plan is consistent with G^P_{and-or}, and for extracting its interpretation when the plan is considered valid. By means of two room-pruning strategies (probabilistic pruning and semantic pruning), the parser generates possible parse And-graphs by pruning those *possible_room* element derivations that are less probable or are not consistent with the grammar.

In the room probabilistic pruning step, for each *possible_room* from the And-graph, its probability of being a room $P(room)$ is calculated as:

$$P(room_i) = \frac{P_{or}(room_i) + P_{rStruc}(room_i) + P_{rGraph}(room_i)}{3} \tag{3}$$

where $P_{or}(room_i)$ is the normalized probability of the number of walls, doors, and windows of $room_i$ relative to its area and perimeter. $P_{rStruc}(room_i)$ is the probability of the spatial composition of the elements of $room_i$. And $P_{rGraph}(room_i)$ is the relative probability of the area and perimeter of $room_i$ over the building area and perimeter. When $P(room_i)$ is very low, the parser generates a new sub-parse graph instance by pruning this room and all its children relations from the first parse graph. Both graphs will be taken into account by the parser to decide which is the most probable parse graph that describes the input floor plan.

In the room semantic pruning, for each room, the *accessibility* and *neighbourhood* restrictions imposed by the grammar productions are studied by the parser. If a room does not fulfils any of these restrictions, the room and its derivation is pruned from the And-graph. Only the new pruned graph is taken into account for most probable parser selection.

Finally, given a floor plan *FP*, N multiple possible parse graphs can be generated. The parse graph that better describes an input floor plan instance according to the grammar, is that one that maximize the posterior probability for the set of all possible parse graphs:

$$PG_{FP} = \max_i(p(PG_i|FP)), \forall i \in N \tag{4}$$

$$p(PG|FP) = \frac{(mean(P_{or}(PG)) + mean(P_{rStruct}(PG)) + mean(P_{rGraph}(PG))}{3}. \tag{5}$$

Then, if $p(PG_{FP})$ is over an appropriate threshold, the plan would be classified as valid. Since the parse graph is an And-graph, it contains the hierarchy, the structure and the semantic composition of the plan at different levels of abstraction.

3 Components Extraction

Even though the definition of new component extraction approaches is out of the scope of this work, here we present the methodologies used to probe the usability of the syntactic and recognition model presented in this paper. Since a hierarchical representation is used to model the floor plans, we have used three different extraction approaches at different level of abstraction: a patch level for windows detection, a line level for doors and walls detection, and a region level for rooms detection. For wall and door detection, at line level, the graphical convention oriented approach presented by Macé et al. in [6] is used. That means that for a different convention, other component extraction approaches would be used, but maintaining the same syntactic and recognition model.

3.1 Patch Level Extraction Approach for Windows Detection

A bag of patches approach has been implemented to detect windows. Due to the great intra-variability of this class, a CRF based on [2], has been used to detect windows by computing the spatial relations between them and walls.

First, all the binarized images of the learning set are divided into non-overlapping squared patches (regular grid). To reduce the feature space dimensionality, PCA is applied to the patches. Then, to create the codebook, a fast K-means clustering based on [1] is computed. After that, a probability of belonging to each one of the entity classes $\{wall,\ window,\ other\_\ element\}$ is assigned to each codeword, using the learning set already labelled. Then, by means of nearest neighbour (1-NN), each patch is assigned to one codeword. The probability of a codeword $w_j \in W = \{w_1, ..., w_j, ...w_N\}$ to belong to a class $c_i, i = \{wall, window, other\_element\}$ is:

$$p(w_j, c_i) = \frac{\#(pt_{w_j}, c_i)}{\#pt_{w_j}}, \tag{6}$$

where $\#(pt_{w_j}, c_i)$ is the number of patches assigned to the codeword w_j that has label c_i, and $\#pt_{w_j}$ is the number of patches assigned to the codeword w_j. Finally, for each patch in an input image, the euclidean distance is calculated over all the codewords in the dictionary. The closest codeword is assigned to this patch toguether with its probability of pertaining to each one of the three classes computed by 6.

The output probability for each patch obtained in that way are then the unary potentials ψ of the CRF graphical model, on which $P(\mathbf{c}|G;k)$ is the conditional probability of the set of class label assignments $\mathbf{c}$ given an adjacency graph $G(S,E)$ and a weight k:

$$-log(P(\mathbf{c}|G;k)) = \sum_{s_i \in S} \psi(c_i|s_i) + k \sum_{(s_i,s_j)\in E} \phi(c_i, c_j|s_i, sj), \tag{7}$$

where the adjacency graph G is the graph topology in the model and ϕ are the pairwaise edge potentials.

In this model, two graph topologies have been defined: horizontal for horizontal windows detection, and vertical for vertical ones. Then, two CRF have been applied over the patch level classifier (one for each topology).

3.2 Region Level Extraction Approach for Rooms Detection

Once all the low level components are extracted from the floor plan, a connected planar graph is created by joining each one of the elements with its incident neighbours. Moreover, a new element called *abstraction* is added between those walls that are relatively closer and well-oriented to split rooms that are not physically separated, e.g. open-plan kitchens, where the kitchen joins the living room. Then, regions are found in this graph by means of [3]. The regions found in the graph are considered as *possible_rooms* in the model.

4 Results

To test our model, we have used 25 real architectural floor plan drawings with the same graphical conventions, which contain a total number of 220 rooms, 911 walls, 243 windows and 203 doors, that have been manually labelled by a sequence of clicks. Using Cross-validation, all plans have been used for testing after a learning step with the 24 remaining documents.

Table 1. Quantitative performance obtained on the overall dataset

Quantitative Results	CE	PCE
Classification rate of well-interpreted FP	83%	100%
Classification rate without pruning	72%	100%
Room neighbouring rate	92%	100%
Room accessibility rate	88%	100%
Room detection rate	84%	-
Closed room detection rate	94%	-
Windows detection rate	92%	-
Rooms pruned rate	88%	-

Table 1 shows the results obtained in terms of plan validation rate with and without using the semantic and probabilistic pruning strategies explained in section 2.3 *Top-Down parser methodology*. Notice that the validation rate using the component extraction techniques *CE* explained in section 3 is 72%. The pruning strategies presented in this paper increase the validation rate up to 83% with same extraction means. But, with perfect components extraction approach *PCE* assumed, our model represents and recognize perfectly all the plans of the corpus. Table 1 also shows the room detection rate over those rooms that are described by closed environment formed by walls, doors and windows; and the performance of the room pruning strategy. The final results for floor plan interpretation have been manually evaluated due to the lack of an appropriate automatic evaluation strategy. Figure 4 shows a graphical interpretation of a valid plan.

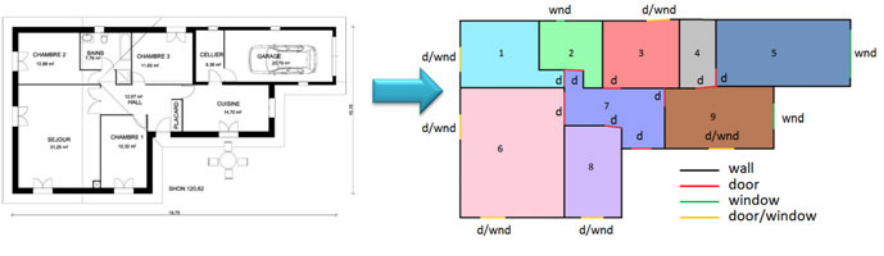


Fig. 4. Interpretation of a floor plan documents in terms of its elements

5 Conclusions

We have presented a syntactic model for document representation and recognition to interpret and validate architectural floor plans. The stochastic image grammar over an And-Or graph structure is learned from a set of examples using MLE, and permits to model floor plan documents hierarchy, structure and semantic composition at different level of abstraction. Our parser generates multiple parse graphs for an input and selects that one that better represents the instance. Moreover, the room-pruning strategies allow to discard those regions that are not rooms accordingly to the stochastic grammar and thus, improve the interpretation rate. Furthermore, the probabilistic models defined over the grammar increase the room detection rate by ruling out most of the *possible_room* false positive examples. In addition to that, the misclassification rate in plan recognition is caused due to a loss of components in the component extraction step. Our model is able to represent and recognize all the examples of the corpus assuming idealistic component extraction techniques.

Acknowledgements. This work has been partially supported by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03 and CONSOLIDER-INGENIO 2010(CSD2007-00018) and by a research grant of the Universitat Autonòma de Barcelona (471-02-1/2010).

References

1. Elkan, C.: Using the triangle inequality to accelerate k-means. In: Proceedings of the Twentieth International Conference on Machine Learning, pp. 147–153 (2003)
2. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: IEEE 12th International Conference on Computer Vision, pp. 670–677 (2009)
3. Jiang, X.Y., Bunke, H.: An optimal algorithm for extracting the regions of a plane graph. Pattern Recogn. Lett. 14(7), 553–558 (1993)
4. Llados, J., Lopez-Krahe, J., Marti, E.: A system to understand hand-drawn floor plans using subgraph isomorphism and hough transform. Machine Vision and Applications 10, 150–158 (1997)
5. Lu, T., Tai, C., Su, F.: A new recognition model for electronic architectural drawings. Computer-Aided Design 37(10), 1053–1069 (2005)
6. Mace, S., Locteau, H., Valveny, E., Tabbone, S.: A system to detect rooms in architectural floor plan images. In: DAS 2010: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 167–174. ACM, New York (2010)
7. Ryall, K., Shieber, S., Marks, J., Mazer, M.: Semi-automatic delineation of regions in floor plans. In: International Conference on Document Analysis and Recognition, vol. 2, p. 964 (1995)
8. Zhu, S.C., Mumford, D.: A stochastic grammar of images, Found. Trends. Comput. Graph. Vis. 2(4), 259–362 (2006)

Linear Prediction Based Mixture Models for Event Detection in Video Sequences

Dierck Matern, Alexandru Paul Condurache, and Alfred Mertins

Institute for Signal Processing, University of Luebeck
{matern,condura,mertins}@isip.uni-luebeck.de

Abstract. In this paper, we propose a method for the detection of irregularities in time series, based on linear prediction. We demonstrate how we can estimate the linear predictor by solving the Yule Walker equations, and how we can combine several predictors in a simple mixture model. In several tests, we compare our model to a Gaussian mixture and a hidden Markov model approach. We successfully apply our method to event detection in a video sequence.

1 Introduction

Event detection [2,4,11] is one of the basic tasks for automatic surveillance. Suppose we observe a complex machinery using several sensors, and we want to distinguish between normal activities and a malfunction (event). As failures are rare occurrences and their properties are commonly unknown, it is difficult to model the malfunctions in a direct manner. To circumvent this problem, we can create a model that describes the machinery when it works normal, and define the "events" as the absence of the "normal case". Our goal is to determine a simple but effective method for this distinction.

An usual approach is to predict the next observation using the knowledge of several previous ones, measure the true observation afterwards, and compare the prediction with the true observation [3,1]. If the difference is higher than a given threshold, we decide "event". Using linear functions for the prediction [8,5] provides several benefits, in particular the ease with which the parameters are estimated.

Two of the most commonly used models for event detection are Gaussian Mixture Models (GMMs) [11] and Hidden Markov Models (HMMs) [4]. While GMMs ignore any temporal connection between samples of the observed stochastic process, HMMs include some temporal coherence. Our approach has several similarities with the GMMs and HMMs. While GMMs use the location in a feature space to distinguish the "normal case" from the "events", our method uses a multiple filter approach [3], respecting a temporal connection between measurements. In comparison to HMMs, our method is simpler to implement, and we use the temporal connection directly, not over the abstract concept of hidden states.

Assuming the input signal is not stationary, adaptive filters like Kalman filters [3] are needed. However, linear adaptive filters include several strong assumptions with respect to the observed data, like Gaussianity and linearity. As a bridge gap solution, linearity is assumed over short intervals. This leads to methods like the extended Kalman filter [3]. We propose here an alternative, in the form of a mixture of linear one step predictors.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 25–32, 2011.

Our method has several advantages, for example less training vectors are needed to achieve comparable results. Furthermore, for event detection, should what we define as the normal case change with time, our method can be easily adapted.

The rest of this paper is structured as follows. In Section 2, we estimate the parameters of a linear prediction, and demonstrate how we can apply a mixture of such predictors to event detection. In Section 3, we demonstrate the effectiveness of the model we propose here in several experiments. In Section 4 we present our conclusions.

2 Linear Predictor Mixtures

The parameters of one step linear predictors (see Section 2.1) are computed from the Yule-Walker-equations [9,10]. In Section 2.2, we show how we can build a mixture of several predictors to describe more complex data.

2.1 Linear Predictors and Linear Prediction Error Filters

There is a strong relationship between Linear Predictors (LPs) and Autoregressive (AR) models [1]. Let $\mathbf{x}$ be a sequence of observations, $\mathbf{x}(t) \in \mathbb{R}^N$, we assume that $\mathbf{x}(t)$ is a linear combination of its p predecessors $\mathbf{x}(t-p),\ldots,\mathbf{x}(t-1)$, a constant term and an error term

$$\mathbf{x}(t) = \sum_{i=1}^{p} a(i)\mathbf{x}(t-i) + a(0)\mathbf{e}_N + \mathbf{v}(t), \tag{1}$$

where $\mathbf{a} := [a(0), a(1), \ldots, a(p)]^\top$ is the *(linear) predictor* and $\mathbf{e}_N := [1,1,\ldots,1]^\top$, $\mathbf{v}(t) \sim N(\mathbf{0},\Sigma)$. (1) is an AR model. From $E(\mathbf{v}(t)) = \mathbf{0}$ follows

$$E[\mathbf{x}(t)] = \hat{\mathbf{x}}(t) := \sum_{i=1}^{p} a(i)\mathbf{x}(t-i) + a(0)\mathbf{e}_N. \tag{2}$$

$\hat{\mathbf{x}}(t)$ is called the *linear prediction* of $\mathbf{x}(t)$.

With $\mathbf{X}(t) := [\mathbf{e}_N, \mathbf{x}(t-1), \mathbf{x}(t-2), \ldots, \mathbf{x}(t-p)]$, we write (2) in matrix notation as $\hat{\mathbf{x}}(t) = \mathbf{X}(t) \cdot \mathbf{a}$. With a combination of $\mathbf{x}$-vectors $\mathbf{y}(t) := \left[\mathbf{x}(t)^\top, \mathbf{x}(t-1)^\top, \ldots, \mathbf{x}(t-n)^\top\right]^\top$ and $\mathbf{X}$-matrices $\mathbf{Y}(t) := \left[\mathbf{X}(t)^\top, \ldots, \mathbf{X}(t-n)^\top\right]^\top$ respectively, $\hat{\mathbf{y}}(t) = \mathbf{Y}(t) \cdot \mathbf{a}$. Using the assumption that the errors $\mathbf{v}$ are mutually independent Gaussian distributed, we can estimate the linear predictor [1,9,10] by

$$\hat{\mathbf{a}}(t) := \left[\mathbf{Y}(t)^\top \cdot \mathbf{Y}(t)\right]^{-1} \mathbf{Y}(t)^\top \mathbf{y}(t). \tag{3}$$

The quadratic prediction error at time step s is $\varepsilon(s)^2 := (\mathbf{x}(s) - \hat{\mathbf{x}}(s))^\top (\mathbf{x}(s) - \hat{\mathbf{x}}(s))$. If we use the estimated predictor $\hat{\mathbf{a}}(t)$, we obtain in (2) the estimation of $\hat{\mathbf{x}}(s)$, that is $\hat{\hat{\mathbf{x}}}(s) := \mathbf{X}(s) \cdot \hat{\mathbf{a}}(t)$, and we estimate the prediction error by

$$\hat{\varepsilon}(s)^2 := (\mathbf{x}(s) - \hat{\hat{\mathbf{x}}}(s))^\top (\mathbf{x}(s) - \hat{\hat{\mathbf{x}}}(s)). \tag{4}$$

This error is most important for the event detection, because if the prediction error is high, we have observed an event.

Using a matrix representation of the linear predictor

$$\mathbf{A}(t) := [\mathbf{I}, -\mathbf{I}\cdot\hat{a}(0), -\mathbf{I}\cdot\hat{a}(1), \ldots, -\mathbf{I}\cdot\hat{a}(p)] \tag{5}$$

and for a shorter notation $\eta(s) := [\mathbf{x}(s)^\top, \mathbf{e}_N^\top, \mathbf{y}(s-1)]^\top$, (4) reads

$$\hat{\varepsilon}(s)^2 = (\mathbf{A}(t)\eta(s))^\top(\mathbf{A}(t)\eta(s)) = \eta(s)^\top\mathbf{H}(t)\eta(s), \tag{6}$$

with $\mathbf{H}(t) := \mathbf{A}(t)^\top\mathbf{A}(t)$. This becomes useful for the Linear Predictor Mixture model (LPM) we describe in the next section.

2.2 Mixture Model and Detection of Events

In order to create a LPM, we use the *exponential representation* of the error

$$f_t(\eta(s)) := \exp\left(-\eta(s)^\top\mathbf{H}(t)\eta(s)\right) = \exp(-\hat{\varepsilon}(s)^2), \tag{7}$$

$0 < f_t(\eta(s)) \leq 1$.

The LPM has similarities to Gaussian Mixture Models (GMMs) [11]. Let $g_i(\mathbf{x})$ be an Gaussian distribution, than the GMM $p(\mathbf{x}) = \sum_{i\in I} w(i)g_i(\mathbf{x})$ is an approximation to a more complex distribution; I is a set of indices, and $w(i)$ are weights with $\sum_{i\in I} w(i) = 1$, $w(i) \geq 0$. In the same manner, the LPM is a mixture of several linear prediction error filters (see Equation (5)), and therefore an approximation to complex time series that are not stationary. We use the exponential representation (7) in a weighted sum, similar to GMMs:

$$F(\eta(s)) := \sum_{t\in T} w(t)f_t(\eta(s)), \tag{8}$$

T is a set of time indices that refers to a training dataset. Note that F is not a probability function, because we use no normalization of the exponential functions. Hence, we refer to F by *score* in the following. Similar to GMMs, an event is detected if the score is below a threshold θ with $0 \leq \theta \leq 1$.

2.3 Parameter Estimation

The parameter estimation for our model proceeds in two steps: first, we estimate a set of several linear predictors $\hat{\mathbf{a}}(t)$, second, we estimate the weights $w(t)$.

Let $\mathbf{x}_0$ be a training set of observations. We initialize the index set T with one time index $t^{(1)}$: $T \leftarrow \{t^{(1)}\}$. Note that with $t^{(1)}$ and (3), a unique estimated linear predictor $\hat{\mathbf{a}}(t^{(1)})$ is defined.

At iteration $\tau > 1$, we want to add an estimated linear predictor $\hat{\mathbf{a}}(t^{(\tau)})$ that reduces the highest prediction error of the training dataset. Hence, we set

$$t^{(\tau)} := \arg\min_{\tilde{t}\notin T} \sum_{i=1}^{\tau-1} f_{t^{(i)}}(\eta_0(\tilde{t})) \tag{9}$$

and $T \leftarrow T \cup \{t^{(\tau)}\}$; η_0 is the combination of several observation vectors similar to η in Section 2.1 and 2.2, with respect to $\mathbf{x}_0$. We terminate this parameter estimation stage if we have a fixed number τ^{max} of predictors.

The weights in Equation (8) are computed by

$$w(t^{(i)}) := \frac{\sum_s f_{t^{(i)}}(\eta_0(s))}{\sum_{t^{(j)} \in T} \sum_s f_{t^{(j)}}(\eta_0(s))} \tag{10}$$

for each $t^i \in T$. The estimated linear predictors and the weights define the LPM, see Equation (8). In the next section, we will test this model in comparison to GMMs and HMMs.

3 Experiments and Discussion

Our experiments consist of two parts: in the first part, we use simulated data and compare the LPM with a GMM and an HMM. In the second part, we use a LPM to detect events in a video stream.

3.1 Comparison to GMMs and HMMs on Synthetic Data

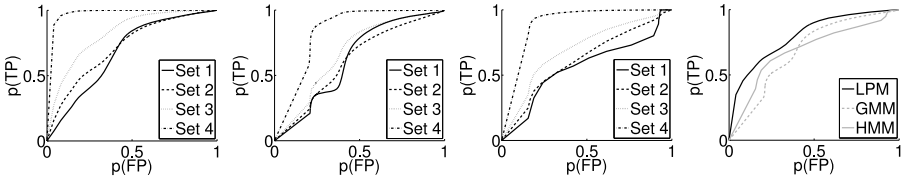
In this test, we compare the LPM with two of the most commonly used models for event detection, the GMM and the HMM. This test is designed as a proof of concept; we concentrate on a real problem in the next section, in this one, we use synthetic data, which has the benefit that we have enough data for each model. A problem with insufficient data especially arises with the HMM, because to estimate the transition probabilities, we need to observe enough transitions.

3.1.a Synthetic data. Similar to HMMs, the model we use as a generator for the synthetic data consists of five states, and it switches between the states at random. As a difference to HMMs as they are discussed in [7], each state in our model is associated with a linear filter, not with a distribution, in order to create a more sophisticated temporal connection. In detail, the synthetic 5D-"observations" are recursively defined by $\mathbf{x}(t) := \mathbf{m}(s) + \overline{\mathbf{x}}_s(t) + \mathbf{v}(t)$ where $\overline{\mathbf{x}}_s(t) := \sum_{i=1}^{3} a_s(i)\,(\mathbf{x}(t-i) - \hat{\mu}(t))$ and $\hat{\mu}(t) := \frac{1}{3}\sum_{i=1}^{3} \mathbf{x}(t-i)$, $\mathbf{v}(t) \sim N(\mathbf{0}, \Sigma)$. The filters $[a_s(i)]_{i=1}^{3}$ and offsets $\mathbf{m}(s)$ are preset values. $s = s(t)$ represents the state of the model, with $P(s(t)|[s(\tau)]_{\tau=1}^{t-1}) = P(s(t)|s(t-1))$.

The event data is generated with a similar model. We use five states with the same transition probabilities. To simulate events, we changed the offsets $\mathbf{m}(s)$ (Set 1), we used noisy filters $a_s(i) + r(t,i)$, $r(t,i) \sim N(0,1)$ (Set 2) and $r(t,i) \sim N(0,2)$ (Set 3) respectively, and we used Gaussian noise only (Set 4). We generated each 50000 normal case and event observations.

3.1.b Tested models. We build a LPM with $\tau^{max} = 50$ predictors. We use $p = 10$ previous observations to predict the following one, and in the training, we use 15 observations to estimate one predictor ($n = 14$, see Section 2.1). For the GMM, we tried several

numbers of Gaussians from five to one hundred. We decided to use ten Gaussians for the comparison, because with this, we have obtained the best results in our tests. The HMM consists of ten states. Each state is associated with a Gaussian distribution [7] (Gaussian Hidden Markov Model, GHMM).



(a) LPM (b) GMM (c) GHMM (d) LPM, GMM and GHMM

Fig. 1. ROCs of (a) the LPM, (b) the GMM, (c) the GHMM and (d) overview of (a), (b) and (c), computed using all test sets at once

3.1.c Results. In Figure 1, we can see different Receiver Operating Characteristics (ROCs) of the three models. The ROC is the curve that results if we plot the probability to correctly detect an event ($p(TP)$) against the probability that an normal observation is falsely classified as an event ($p(FP)$), that is $p(TP) = \frac{\#(\text{Detected, simulated events})}{\#(\text{Simulated events})}$ and $p(FP) = \frac{\#(\text{Falsely detected events})}{\#(\text{Normal observations})}$. A method is superior to another one at a fixed false positive rate if its ROC curve is above another ROC. In order to reduce false alerts and respect that events are rare occurrences, we are particularly interested in parameters with low false positive rates.

In Figure 1(a), 1(b) and 1(c), we can see the performance of the different models separately, one ROC for each test. In Figure 1(d), we can see the overall performance (results using all event datasets as one set) of each model.

Comparing the GHMM and the GMM, the GHMM performs better. But as we can see in Figure 1(a) and 1(d), the LPM outperforms both methods. Hence, there are event detection problems where the LPM can be successfully applied, and they perform better than GMMs or GHMMs. In the next section, we apply the LPM on real data.

3.2 Car Tracking and Event Detection

The setup of this test is as follows. A web cam is positioned to monitor a fixed area. We drive an RC car in the visible area and perform several actions. The “normal case” consists of any combination of normal movements (driving straight, turning left or right), an “event” is an arbitrary action that differs from these possible actions (for example, if the car hits another object).

To adapt the LPM to tracking and motion detection, every time window of p observations is rotated so that the difference vector of the first two is orientated in one particular direction. This simplifies the prediction, and reduces the number of predictors.

3.2.d Tracking. We use a background subtraction algorithm [6] for the motion detection. This algorithm estimates for every image of a video sequence the foreground and updates a background model. It uses one threshold for each pixel. It is similar to many other background subtraction algorithms, but we have adapted it to our problem, especially to color images.

In detail, let $\mathbf{B}_t(i,j) \in [0,1]^3$ be the (normalized color) pixel (i,j) of the tth background image, $\mathbf{C}_t(i,j) \in [0,1]^3$ the corresponding pixel in the tth measured image, $T_t(i,j) \in \mathbb{R}_+$ the tth threshold for pixel (i,j). We say, $\mathbf{C}_t(i,j)$ belongs to the foreground if $(\mathbf{B}_t(i,j) - \mathbf{C}_t(i,j))^\top (\mathbf{B}_t(i,j) - \mathbf{C}_t(i,j)) > T_t(i,j)$. Let $G_t(i,j) = 1$ if $\mathbf{C}_t(i,j)$ is foreground, and 0 otherwise, than

$$\mathbf{B}_{t+1}(i,j) := (1 - G_t(i,j)) \cdot (\alpha_B \mathbf{B}_t(i,j) + (1-\alpha_B)\mathbf{C}_t(i,j)) + G_t(i,j) \cdot \mathbf{B}_t(i,j),$$
$$T_{t+1}(i,j) := (1 - G_t(i,j)) \cdot (\alpha_B (T_t(i,j) + 0.01) + (1-\alpha_B) D_t(i,j)) + G_t(i,j) \cdot T_t(i,j),$$

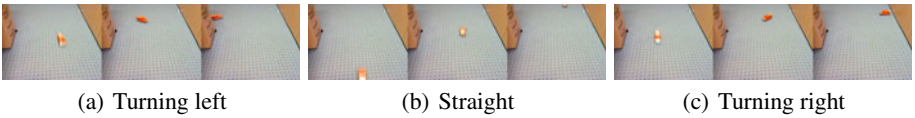
where $D_t(i,j) := (\mathbf{B}_t(i,j) - \mathbf{C}_t(i,j))^\top (\mathbf{B}_t(i,j) - \mathbf{C}_t(i,j))$. The constant 0.01 is used for noise suppression, $\alpha_B \in (0,1)$ controls the adaption to the background of $\mathbf{C}$. The resulting blob (all foreground pixels) for several frames can be seen in Figures 3(a) and 3(b).

3.2.e Extended model. We model a special case of an AR model in this test (see Equation (1)),

$$\mathbf{x}(t) = \sum_{i=1}^{p} a(i)\mathbf{x}(t-i) + \mathbf{a}_0 + \mathbf{v}(t), \tag{11}$$

$\mathbf{a}_0 \in \mathbb{R}^2$, $\mathbf{x}(t) \in [0,1]^2$, and we assume $p = 3$. In this test, $\mathbf{x}(t)$ is the position of the car at frame t, one dimension of $\mathbf{a}_0$ represents the forward movement, the other one the drift. We use this adaption because we assume these two values to be very different. This adaption implies that $\mathbf{X}^{(a)}(t) := [\mathbf{I}_N, \mathbf{x}(t-1), \mathbf{x}(t-2), \ldots, \mathbf{x}(t-n)]$ is used for the Yule-Walker equations instead of the $\mathbf{X}(t)$ assumed in Section 2.1. We use for Equation (7) $f_t(\eta(t_1)) = \exp(-10 \cdot \hat{\varepsilon}(t_1))$. This scaling is used for visualization only.

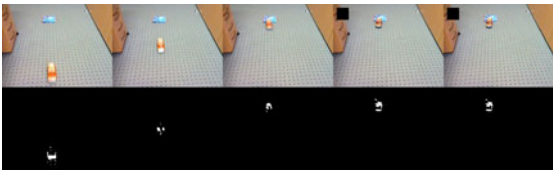
We use three predictors, one for straight parts, one for turning left and one for turning right. As weights, we set $w(1) = w(2) = w(3) = 1/3$. If the score F is lower than $\theta = 0.4$, we say, we have detected an event. In general, θ is an arbitrary threshold with $0 < \theta < 1$, and $\theta = 0.4$ is sufficient for our task, as we have verified with some test data (see Figure 4).



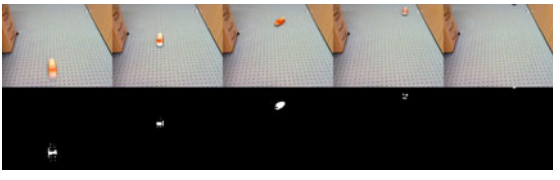
(a) Turning left (b) Straight (c) Turning right

Fig. 2. Frames from the training data: the car turns left(a), drives straight (b) and turns right (c)

3.2.f Training data. The predictors are estimated only for the basic actions. That means, the predictor for straight movement is estimated using only a video where the car moves in one particular direction, and the turns are estimated using videos including only the turns, but no straight parts. A normal activity is any combination of these actions. For each action, less than hundred data vectors were sufficient; for many other models, we would have to use more features, depending on the complexity of the model.



(a) Accident



(b) Normal case

Fig. 3. Several frames from an accident video and the blobs of the tracking (a), and normal activities: car is driving in an S-bend (b)

3.2.g Results. In Figure 3(a), we see several frames of a video we use, that is the first frame with an detected object, immediately before an accident, the frame that has been captured during the accident, the following one and at the last frame of the sequence. The mark on the upper left denotes an event. The event is detected right after the accident.

In Figure 3(b), we see the car, performing an S-bend. This action is correctly classified as normal activity, and no events are detected.

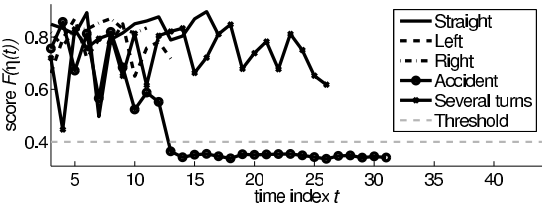


Fig. 4. Score F as described in Section 2.3

In Figure 4, we can see the score F of several normal movements and an accident (an event). As we can see, the score of the normal activities is above the threshold. The same is true for the event video until the accident happens, than the score drops below the threshold and keeps at this low level.

4 Conclusion and Outlook

We have described and tested a method for event detection based on a mixture of linear predictions. Our model outperforms a GMM and a GHMM in a set of tests, despite being less complex than the latter one.

In contrast to GMMs, the LPM uses time dependencies for an improved decision. Furthermore, the LPM is a descriptive model, while the GMM and the GHMM are generative ones. However, LPMs and GMMs have the same simplicity in the inference. The estimation of the LPM is the easiest, because we do not need to estimate covariances, which can be difficult.

Further, we can adapt the LPM easily if the normal case changes by adding new predictors and calculate the weights on some new measurements. This adaption is not useful for GMMs, because it alters the probability of all observations, and HMMs have to be calculated from scratch.

Some problems with the LPM arise from the solution of the Yule Walker equations. For example, in the presence of outliers, the accuracy of the predictor estimation decreases, and if the variance in the data is too low, the number of values to estimate a linear predictor increases. Solutions to these problems are available within the frame of the Yule Walker equations. Because the LPM builds upon these equations, these solutions are available for the LPMs as well.

References

1. Burock, M.A., Dale, A.M.: Estimation and detection of event-related fMRI signals with temporally correlated noise: A statistically efficient and unbiased approach. Human Brain Mapping 11(4), 249–260 (2000)
2. Cline, D.E., Edgington, D.R., Smith, K.L., Vardaro, M.F., Kuhnz, L.: An automated event detection and classifiacation system for abyssal time-series images of station m, ne pacific. In: MTS/IEEE Oceans 2009 Conference Proceedings (2009)
3. Gustafsson, F.: Adaptive Filtering and Change Detection. John Wiley and Sons, Inc., Chichester (2000)
4. Jin, G., Tao, L., Xu, G.: Hidden markov model based events detection in soccer video. In: Campilho, A.C., Kamel, M.S. (eds.) ICIAR 2004. LNCS, vol. 3211, pp. 605–612. Springer, Heidelberg (2004)
5. Liu, W., Lu, X.: Weighted least squares method for censored linear models. Journal on Nonparametric Statistics 21, 787–799 (2009)
6. Piccardi, M.: Background subtraction techniques: a review. Proceedings of the Conference on Systems, Man and Cybernetics 4, 3099–3104 (2004)
7. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
8. Rancher, A.C.: Linear Models in Statistics. John Wiley and Sons, Inc., Chichester (2000)
9. Walker, G.: On periodicity in series of related terms. Proceedings of the Royal Society of London 131, 518–532 (1931)
10. Yule, G.U.: On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London 226, 267–298 (1927)
11. Zhuang, X., Huang, J., Potamianos, G., Hasegawa-Johnson, M.: Acoustic fall detection using Gaussian mixture models and gmm supervectors. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 69–72 (2009)

A Visual Saliency Map Based on Random Sub-window Means

Tadmeri Narayan Vikram[1,2], Marko Tscherepanow[1], and Britta Wrede[1,2]

[1] Applied Informatics Group
[2] Research Institute for Cognition and Robotics (CoR-Lab)
Bielefeld University, Bielefeld, Germany
{nvikram,marko,bwrede}@techfak.uni-bielefeld.de

Abstract. In this article, we propose a simple and efficient method for computing an image saliency map, which performs well on both salient region detection and as well as eye gaze prediction tasks. A large number of distinct sub-windows with random co-ordinates and scales are generated over an image. The saliency descriptor of a pixel within a random sub-window is given by the absolute difference of its intensity value to the mean intensity of the sub-window. The final saliency value of a given pixel is obtained as the sum of all saliency descriptors corresponding to this pixel. Any given pixel can be included by one or more random sub-windows. The recall-precision performance of the proposed saliency map is comparable to other existing saliency maps for the task of salient region detection. It also achieves state-of-the-art performance for the task of eye gaze prediction in terms of receiver operating characteristics.

Keywords: Bottom-up Visual Attention, Saliency Map, Salient Region Detection, Eye Fixation.

1 Introduction

Visual saliency maps are utilized for determining salient regions in images or predicting human eye gaze patterns. Thus they have been exploited extensively for various intelligent interactive systems. There are a wide range of applications from image compression [2], object recognition [3–5], image segmentation [6] and various other computer vision tasks where saliency maps are employed. The intensity of a given pixel in the saliency map corresponds to the attention value attributed to the pixel in the original image.

The first computational model of visual attention was proposed by Koch and Ulmann [21]. They also introduced the concept of a saliency map. Subsequently, a great variety of different bottom-up visual attention models have been proposed in the literature. Methods which are employed to detect salient regions do not emphasize the semantic relevance and the opposite is true in the case of methods which are utilized to predict eye gaze patterns. Despite vast research there has not been a method which could be successfully utilized for both salient region detection as well as eye gaze pattern prediction. Therefore, in this paper we

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 33–40, 2011.

propose a novel saliency map which performs well in salient region detection and eye gaze prediction tasks.

2 Literature Review

Existing saliency maps can be categorized into two fundamental groups: those which rely on local image statistics and those relying on global properties. The popular bottom-up visual attentional model proposed by Itti et al. [1] is based on local gradients, color and orientation features at different scales. It further inspired the application of contrast functions for the realization of bottom-up visual attention. The work of Gao et al. [22] was the first to employ contrast sensitivity kernels to measure center-surround saliencies. It was further improved by local-steering kernels [8] and self information [9]. The method of Bruce and Tsotsos [18] achieved the same level of performance by employing local entropy and mutual information-based features. Local methods are found to be computationally more expensive, and several global and quasi-global methods have been devised to address the issue of computational efficiency. The idea of utilizing the residual Fourier spectrum for saliency maps was proposed in [10, 11]. The authors employ the Fourier phase spectrum and select the high frequency components as saliency descriptors. These methods are shown to have high correlation with human eye-gaze pattern on an image. Frequency domain analysis for image saliency computation warrants the tuning of several experimental parameters. In order to alleviate this issue, several methods which rely on spatial statistics and features [6, 12–15] have been proposed.

Salient region detection and eye gaze prediction are the two significant applications of saliency maps. Salient region detection is relevant in the context of computer vision tasks like object detection, object localization and object tracking in videos [14]. Automatic prediction of eye gaze is important in the context of image asthetics, image quality assessment, human-robot interaction and other tasks which involve detecting image regions which are semantically interesting [17]. The contemporary saliency maps are either employed to detect salient regions as in the case of [6, 12–14], or are used to predict gaze pattern which can be seen in the works of [1, 8–10, 15]. Though these two tasks appear similar, there are subtle differences between them. Salient regions of an image are those which are visually interesting. Human eye gaze which focuses mainly on salient regions is also distracted by semantically relevant regions [3].

Contrast has been the single most important feature for the computation of saliency maps and modelling bottom-up visual attention as it can be inferred from [6, 8, 9, 13, 14]. The method based on global contrast [6] employs absolute differences of pixels to the image mean as saliency representatives. The methods which model the distribution of contrast based on local image kernels [8, 9] need training priors and tuning of a large set of experimental parameters. The local weighting models proposed in [14, 15] are effective, but are computationally expensive. The local symmetric contrast-based method [13] overcomes the many aforementioned shortcomings. Recent research has suggested that contrast detection and normalization in the V1 cortex is carried out in non-linear random

local grids, rather than in linear fashion with regular grids [19, 20]. This property has been exploited in [14, 15] to compute saliency at pixel level and in [6] at global level.

We hereby propose a quasi-global method which operates by computing local saliencies over random regions of an image. This helps in obtaining better computational run-time and also captures local contrast unlike the global methods for computing saliency maps. Furthermore, it does not require any training priors and has only a single experimental parameter which needs tuning. Unlike the existing methods, the proposed saliency map is found to have consistent performance in both salient region detection and eye gaze prediction tasks. The proposed saliency map is determined as follows.

3 Our Method

We consider a scenario where the input I is a color image of dimension $r \times c \times 3$, where r and c are the number of rows and columns respectively. The input image is subjected to a *Gaussian* filter in order to remove noise and abrupt onsets. This is further converted to CIELab space and decomposed into the three (L, a, b) component images of dimension $r \times c$. CIELab space is preferred because of its similarity to the human psycho-visual space [13, 14].

Let n be the number of random sub-windows over the individual L, a and b component images given

$$R_i = \{(x_{1i}, y_{1i}), (x_{2i}, y_{2i})\} \text{ such that } \begin{cases} 1 \leq i \leq n \\ 1 \leq x_{1i} < x_{2i} \leq r \\ 1 \leq y_{1i} < y_{2i} \leq c \end{cases} \tag{1}$$

where R_i is the i^{th} random sub-window with (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}) being the upper left and the lower right co-ordinates respectively.

The final saliency map S of dimension $r \times c$ is thus defined as

$$S = \sum_{i=1}^{n} \left\| R_i^L - \mu(R_i^L) \right\| + \left\| R_i^a - \mu(R_i^a) \right\| + \left\| R_i^b - \mu(R_i^b) \right\| \tag{2}$$

$\|\cdot\|$ denotes the Euclidean norm and $\mu(.)$ the mean of a given input vector, which is a two dimensional matrix in our case. To further enhance the quality of the saliency map S, we subject it to median filtering and histogram equalization. An illustration of the above paradigm is given in Fig. 1. For the sake of illustration we have considered only three random sub-windows and the resulting interim saliency map.

4 Results

First, we illustrate the efficacy of our method on three selected images from the MSR [16] dataset in Fig. 2. The example image in Fig. 1 is also presented in

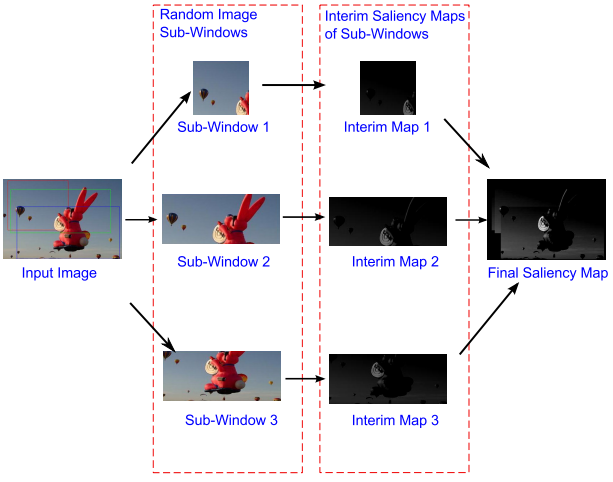


Fig. 1. An illustration of the proposed method where three random windows are generated to produce an interim saliency map for the given input image

Fig. 2. It should be noted that the final saliency map shown in Fig. 2 is obtained by considering a large number of random sub-windows. It can be observed from Fig. 2 that multiscale saliency [1] and local steering kernel-based saliency [8] lay more emphasis on edges and other statistically significant points, rather than salient regions. The local steering kernel-based saliency [8], which is tuned to detect semantically relevant regions like corners, edges and local maxima ends up projecting mild image noise as salient. This can be observed on the upper right image of the moon in Fig. 2, where irrelevant regions are shown as salient. The results due to symmetry-based saliency [13], shows that the images have a sharp contrast. Images which do not consist of smooth signals have found to be bad representatives of eye gaze fixation, as eye gaze fixation function in reality is found to be smooth. The saliency provided by entropy-based methods [18] exhibit low contrast and are found to be inefficient for the task salient region detection in [14]. It can be observed that the proposed saliency does not output spurious regions as salient, has no edge bias, works well on both natual images and images with man made objects, and most importantly is also able to grasp the subjective semantics and context of a given region. The final property makes our method suitable for the task of eye gaze fixation. The experimentation carried out during the course of this research is presented in the section to follow.

In addition to the analysis of examplar images, more comprehensive experiments were carried out on the MSR dataset [16] to validate the performance of the proposed saliency map for the task of salient region detection. In order to evaluate the performance on eye gaze prediction, experiments were conducted on the York University [18] and MIT [17] eye fixation datasets. We compared our method with reference to eight of the existing state-of-the-art methods. The selection of these methods was influenced by the impact factor of the conference and journals in which they were published, popularity of the method in terms

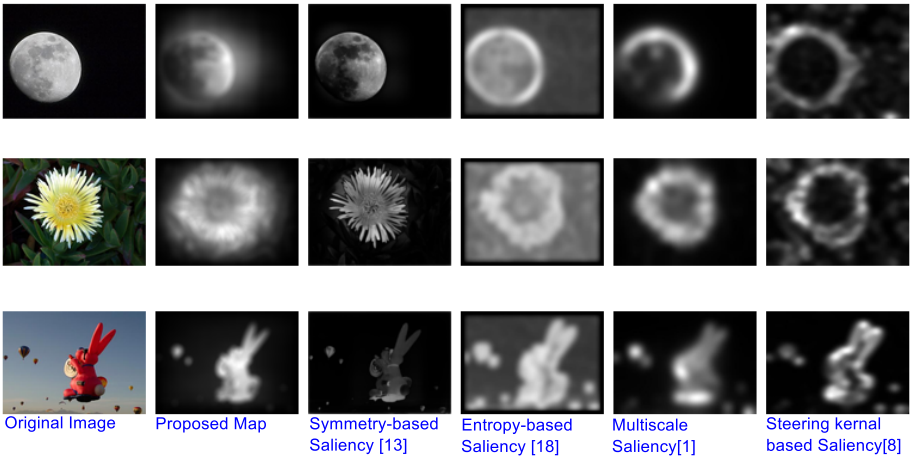


Fig. 2. Illustration of the proposed saliency map on three sample images of the MSR dataset [16]. From left to right: original Image from the MSR dataset [16], followed by the resultant saliency maps of the proposed method, symmetry-based saliency [13], entropy-based saliency [18], multiscale saliency [1] and local steering kernel-based saliency [8].

of citation, and the differences in their approaches. The eight methods are the global saliency-based method [6], symmetric saliency [13], entropy and mutual information-based saliency [18], graph-based saliency [15], multiscale saliency [1], local weighted saliency [14], self information-based saliency [9] and local steering kernel-based saliency [8]. The source codes for the methods were obtained from the homepages of the respective authors, whose links have been mentioned in their articles.

The following settings were used for all the experiments carried out. A Gaussian filter of size 3×3 was used as a pre-processor on the images for noise removal. The number of distinct random sub-windows (n) was set to $0.02 \times r \times c$. We arrived at this value, by varying (n) from (0.005 to 0.03) $\times\, r \times c$ and found that the receiver operating characteristics (ROC) area under the curve (AUC) attained a level of saturation at $0.02 \times r \times c$ on York University [18] and MIT [17] datasets. And finally a median filter of size 11×11 was employed to smooth the resultant saliency map before being enhanced by histogram equalization. All experiments were conducted using Matlab v7.10.0 (R2010a), on an Intel Core 2 Duo processor with Ubuntu 10.04.1 LTS (Lucid Lynx) as operating system.

4.1 Experiments with the MSR Dataset

The original MSR dataset [16] consists of 5000 images with ground truths for salient regions as rectangular regions of interest (ROI). The problems and issues due to such ROIs are explained in [6] and hence the same authors select a subset

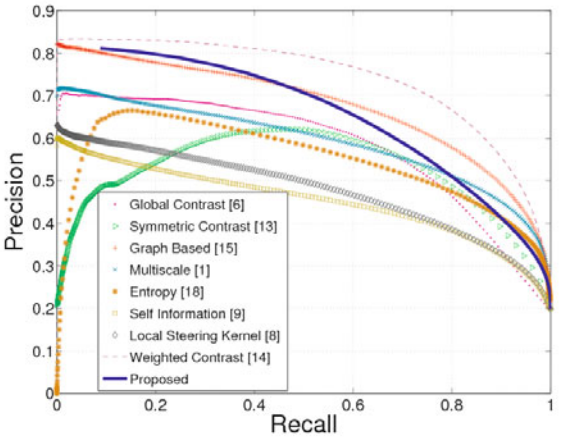


Fig. 3. The Recall-Precision performance of the methods under consideration on MSR dataset [16], with the experimental settings of [6]. Note that our method clearly outperforms the methods based on global contrast [6], symmetric contrast [13], multiscale [1], entropy [18], self information [9] and local steering kernel [8]

of 1000 images from the original set images and create exact segmentation masks. We followed the same experimental settings as described in [6]. In Fig. 3, we show the Recall-Precision performance of the models.

It can be observed that the proposed method clearly has a higher performance than the methods of [1, 6, 8, 9, 13, 18] and comparable performance with that of [14, 15], without having any of their drawbacks. The entropy-based saliency map [18] though promising does not have a high performance because the MSR dataset has a mix of natural images where the entropy is uniformly distributed. Local kernel-based methods [8, 9] also perform moderately because they are biased towards corners and edges than regions. Only the graph based method [15] and weighted distance method [14] perform well, because they have no bias towards edges.

4.2 Experiments Using Eye Fixation Datasets

We benchmarked the performance of the proposed method on York University [18] and the MIT [17] eye fixation dataset. The dataset of York University [18] consists of 120 images and the MIT dataset [17] consists of 1003 images. We followed the experimental method as suggested in [18] and obtained the ROC-AUC on the datasets. It can be observed from Table. 1, that our method has state-of-the-art performance. We omitted the methods of [9, 14] on the MIT dataset [17], as the corresponding Matlab codes required images to be down-sampled to a smaller size.

Table 1. The performance of the methods under consideration in terms of ROC-AUC on the York University [18] and MIT[17] datasets. It can be observed that the proposed method has state-of-the-art performance on both of the eye fixation datasets.

Saliency Map	**York University [18]**	**MIT [17]**
Global Contrast[6]	0.54	0.53
Symmetric Contrast[13]	0.64	0.63
Graph Based[15]	0.84	**0.81**
Multiscale[1]	0.81	0.76
Entropy[18]	0.83	0.77
Self Information[9]	0.67	-NA-
Local Steering Kernel[8]	0.75	0.72
Weighted Contrast[14]	0.75	-NA-
Proposed	**0.85**	**0.81**

5 Discussion and Conclusion

We propose a method which has good performance on both salient region detection and eye gaze prediction tasks. The proposed method does not require training priors, has a minimal set of tunable parameters and relies only on contrast features to compute saliency maps. Our method requires minimal programming effort and achieves state-of-the-art performance despite its simplicity. Like the remaining contrast-based saliency maps, our method also fails to perform well when the color contrast is extremely low. Furthermore, the proposed saliency map fails when the task is to detect regions based on corners, orientation differences, minute differences in shapes etc. The proposed method is well suited in the scenario of human-robot interaction where eye gaze prediction and salient region detection need to be performed concurrently. Generating image specific random sub-windows to boost the proposed saliency map is another topic we wish to address in our future works.

Acknowledgments. Tadmeri Narayan Vikram gratefully acknowledges the financial support from the EU FP7 Marie Curie ITN RobotDoc. Contract No. 235065.

References

1. Itti, L., Koch, C., Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans. Pattern Analysis Machine Intelligence 20(11), 1254–1259 (1998)
2. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Processing 19(1), 185–198 (2010)
3. Rothenstein, A.L., Tsotsos, J.K.: Attention links sensing to recognition. Image and Vision Computing 26(1), 114–126 (2008)
4. Elazary, L., Itti, L.: A Bayesian model for efficient visual search and recognition. Vision Research 50(14), 1338–1352 (2010)

5. Moosmann, F., Larlus, D., Jurie, F.: Learning Saliency Maps for Object Categorization. In: ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision (2006)
6. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Frequency tuned Salient Region Detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (2009)
7. Buschman, T.J., Miller, E.K.: Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. Science 315(5820), 1860–1862 (2007)
8. Seo, H.J., Milanfar, P.: Static and Space-time Visual Saliency Detection by Self-Resemblance. Journal of Vision 9(12), 1–27 (2009)
9. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian Framework for Saliency Using Natural Statistics. Journal of Vision 8(7), 1–20 (2008)
10. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Processing 19(1), 185–198 (2010)
11. Cui, X., Liu, Q., Metaxas, D.: Temporal spectral residual: fast motion saliency detection. In: ACM International Conference on Multimedia, pp. 617–620 (2009)
12. Rosin, P.L.: A simple method for detecting salient regions. Pattern Recognition 42(11), 2363–2371 (2009)
13. Achanta, R., Süsstrunk, S.: Saliency Detection using Maximum Symmetric Surround. In: IEEE International Conference on Image Processing (2010)
14. Vikram, T.N., Tscherepanow, M., Wrede, B.: A Random Center Surround Bottom up Visual Attention Model useful for Salient Region Detection. In: IEEE Workshop on Applications of Computer Vision, pp. 166–173 (2011)
15. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Neural Information Processing Systems, pp. 545–552 (2007)
16. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.: Learning to Detect A Salient Object. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (2009)
18. Bruce, N.D., Tsotsos, J.K.: Attention based on Information Maximization. In: The International Conference on Computer Vision Systems (2007)
19. Mante, V., Frazor, R.A., Bonin, V., Geisler, W.S., Carandini, M.: Independence of luminance and contrast in natural scenes and in the early visual system. Nature Neuroscience 8(12), 1690–1697 (2005)
20. Soltani, A., Koch, C.: Visual Saliency Computations: Mechanisms, Constraints, and the Effect of Feedback. Neuroscience 30(38), 12831–12843 (2010)
21. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
22. Gao, D., Mahadevan, V., Vasconcelos, N.: The discriminant center-surround hypothesis for bottom up saliency. In: Neural Information Processing Systems (2007)

There Is More Than One Way to Get Out of a Car: Automatic Mode Finding for Action Recognition in the Wild*

Olusegun Oshin, Andrew Gilbert, and Richard Bowden

Centre for Vision, Speech and Signal Processing,
University of Surrey, Guildford, Surrey,
United Kingdom GU2 7XH
{o.oshin,a.gilbert,r.bowden}@surrey.ac.uk

Abstract. "Actions in the wild" is the term given to examples of human motion that are performed in natural settings, such as those harvested from movies [10] or the Internet [9]. State-of-the-art approaches in this domain are orders of magnitude lower than in more contrived settings. One of the primary reasons being the huge variability within each action class. We propose to tackle recognition in the wild by automatically breaking complex action categories into multiple modes/group, and training a separate classifier for each mode. This is achieved using RANSAC which identifies and separates the modes while rejecting outliers. We employ a novel reweighting scheme within the RANSAC procedure to iteratively reweight training examples, ensuring their inclusion in the final classification model. Our results demonstrate the validity of the approach, and for classes which exhibit multi-modality, we achieve in excess of double the performance over approaches that assume single modality.

1 Introduction

Human action recognition from video has gained significant attention in the field of Computer Vision. The ability to automatically recognise actions is important because of potential applications in video indexing and search, activity monitoring for surveillance, and assisted living purposes. The task is especially challenging due to variations in factors pertaining to video set-up and execution of the actions. These include illumination, scale, camera motion, viewpoint, background, occlusion, action length, subject appearance and style.

Approaches to action recognition attempt to learn generalisation over all class examples from training, making use of combinations of features that capture both shape and motion information. While this has resulted in excellent results for videos with limited variation, in natural settings, the variations in camera set-up and action execution are much more significant, as can be seen in Figure 1. It is,

* This work is supported by the EU FP7 Project Dicta-Sign (FP7/2007-2013) under grant agreement no 231135, and the EPSRC project Making Sense (EP/H023135/1).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 41–48, 2011.

(a) *GetOutCar* action

(b) *HandShake* action

Fig. 1. Four examples of two actions of the Hollywood2 dataset, all showing the different modes of the same action

therefore, unrealistic to assume that all aspects of variability can be modelled by a single classifier. This motivates our approach.

The method presented in this paper tackles action recognition in complex natural videos by following a different approach. Instead of treating all examples of a semantic action category as one class, we automatically separate action categories into various *modes* or groups, thereby significantly simplifying the training and classification task. We achieve this by applying the tried and tested Random Sampling Consensus (RANSAC) algorithm [2] to training examples of actions, with a novel adaptation based on an iterative reweighting scheme inspired by boosting, and obtain impressive results. Whereas clustering merely groups class examples based on proximity within the input space, our approach groups the positive examples while attempting to exclude negative ones. This ensures less contamination within sub-categories compared to clustering. To our knowledge, our approach is the first to make use of the automatic separation of complex category examples into groups for action recognition. For classes where multi-modality is evident, we achieve a performance increase in excess of 100% over approaches assuming single modality.

The layout for the remainder of this paper is as follows: Section 2 discusses related research. In Section 3, we present our approach in detail. We describe our experimental set-up in Section 4 and present recognition results in Section 5. Finally, Section 6 concludes the paper.

2 Related Work

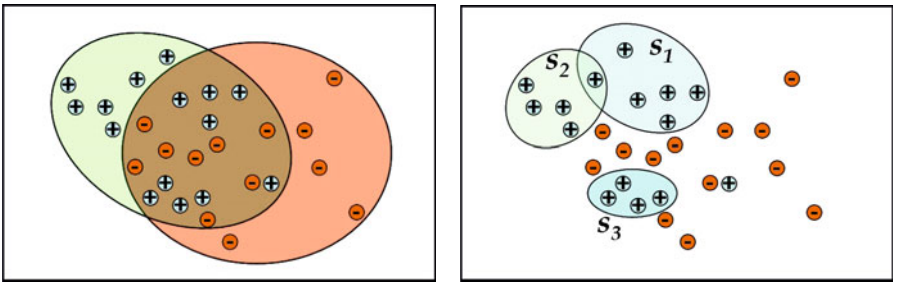
There is a considerable body of work exploring the recognition of actions in video [11,1,6,8,4,13]. While earlier action recognition methods were tested on simulated actions in simplified settings, more recent work has shifted focus to so-called Videos in the Wild, *e.g.* personal video collections available online, and movies. As a result of this increase in complexity, recent approaches attempt to model actions by making use of combinations of feature types. Laptev and Perez [7], distinguish between actions of Smoking and Drinking in movies, combining an optical flow-based classifier with a separately learned space-time classifier applied to a keyframe of the action. The works of [6] and [10] recognise a wider range of actions in movies using concatenated HoG and HoF descriptors in a bag-of-features model, with [10] including static appearance to learn contextual information. Han *et al.* [5] capture scene context by employing object detectors and introduce bag-of-detectors, encoding the structural relationships between object parts, whereas Ullah *et al.* [13] combine non-local cues of person detection, motion-based segmentation, static action detection, and object detection with local features. Liu *et al.* [8] also combine local motion and static features and recognise actions in videos obtained from the web and personal video collections.

In contrast to these multiple-feature approaches, our method makes use of one feature type. Then, instead of seeking to learn generalisation over all class examples, we argue that a single action can be split into subsets, which cover the variability of action, environment and viewpoint. For example, the action of *Getting Out of a Car* can be broken into sets of radically different actions depending on the placement of the camera with respect to the car and individual. We automatically discover these modes of action execution or video set-up, thereby simplifying the classification task. While extensive work exist on local classification methods for object category recognition [15], human pose estimation [14], etc [12], the assumption of multimodality has not so far been applied to action recognition. We employ RANSAC [2] for this grouping and introduce a reweighting scheme that increases the importance of difficult examples to ensure their inclusion in a mode.

3 Action Modes

The aim of this work is to automatically group training examples in natural action videos, where significant variations occur, into sub-categories for improved classification performance. The resulting sub-categories signify different modes of an action class, which when treated separately, allow for better modelling of training examples, as less variations exist within each mode.

To illustrate, Figure 2(a) shows positive and negative examples in a simple binary classification problem. It can be observed that, using a Gaussian classifier, there exists a great deal of overlap between the classes. This is as a result of both positive and negative examples occupying regions within the classification space that make separation impossible with a single classifier.



(a) Classification problem with overlap between classes

(b) Classification problem with identified modes, simplifying the classification of the positive data

Fig. 2. Binary classification problem for classes with overlap due to large variability in the data

While this is an obvious problem in classification that could be solved using a mixture model, for the task of action recognition in natural videos, this phenomenon is still observed, yet mostly ignored. Figure 1 shows six different modes of the action class *GetOutCar*, and four modes of the action category *HandShake*, respectively, taken from the Hollywood2 dataset [10]. It can be seen that, while the same action is being performed, all the examples appear radically different due to the differences in camera setup and in some cases, action execution. Despite these variations, the examples are given one semantic label, making automatic classification extremely difficult. We propose that, for cases such as this, it should be assumed that the data is multi-modal, and the use of single classifiers for such problems is unrealistic.

Our method is based on the notion that there exists more compact groupings within the classification space that when identified, reduces confusion between classes. Figure 2(b) shows examples of such groupings when applied to the binary classification problem. The grouping also enables the detection of outliers, which are noisy examples that may prove detrimental to the overall classifier performance, as can be observed by the single ungrouped positive example in Figure 2(b).

3.1 Automatic Grouping Using Random Sampling Consensus

For a set, Φ of training examples belonging to a particular class C, we iteratively select a random subset, $\varphi \subset \Phi$ of the examples. We then train a binary classifier of the subset φ against all training examples from other classes. This forms the hypothesis stage. The resulting model is then evaluated on the remainder of the training example set, $\psi \subset \Phi$, where $\Phi = \varphi \cup \psi$. For each iteration, $t = \{1...T\}$, a consensus set is obtained, labelled Group ς_t, which is made up of φ and the correctly classified examples from ψ. This procedure identifies examples in the subset ψ where the mode is similar to examples in φ.

3.2 Sub-category Selection

After several iterations of this random training and evaluation, the sub-categories $\mathcal{S}_j, j = \{1...J\}$ of the class C are selected from the groups $\varsigma_t, t = \{1...T\}$, where T is the number of iterations, and J is the number of sub-categories. We apply AdaBoost [3] in the selection process in an attempt to ensure that all training examples are represented in at least one of the sub-categories. Each group, ς_t, is given a score which is the sum of weights $W_t(i)$ associated with each example i in the group. The process is initialised by assigning equal weights, $W_1(i) = \frac{1}{|\Phi|}$ to all training examples. Hence, in the first instance, we find the group that results in the highest number of correctly classified examples in subset ψ, labelled $\mathcal{S}_1$. This is the first sub-category.

For subsequent sub-categories, the weight of each example is given by

$$W_{t+1}(i) = \frac{W_t(i)}{Z_t} exp(-\alpha_t y_i h_t(x_i)), \tag{1}$$

given that,

$$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t}) \tag{2}$$

and the term $y_i h_t(x_i) = \{-1, +1\}$ denotes the absence or presence of a particular example in the previously selected sub-categories. Z_t is a normalisation constant, and ϵ_t is the error rate. This process is repeated until all examples are selected, or the maximum number of sub-categories is exceeded. In some cases outliers are discovered. Also, there is often overlap of examples between the sub-categories $\mathcal{S}_j$ as shown in 2(b).

During training, the sub-categories are trained separately against examples of other classes. Examples of class C that do not belonging to the sub-category being trained are not included in the training. During classification, results of all sub-categories $\mathcal{S}_j$ are combined, with all true positives of $\mathcal{S}$ counted as belonging to one class C. The computational complexity of our approach is $O(T(\mathcal{C}_{hypo}(|\varphi|) + |\psi|\mathcal{C}_{test}) + J\mathcal{C}_{boost}(|\Phi|))$, where $\mathcal{C}_{hypo}$ and $\mathcal{C}_{test}$ are the costs of RANSAC hypothesis and test phases respectively, $\mathcal{C}_{boost}$ is the cost of the reweighting procedure, and $|.|$ denotes cardinality.

4 Experimental Setup

We evaluate our method on the Hollywood2 Human Action dataset [10]. The dataset contains 12 action classes: AnswerPhone, DriveCar, Eat, FightPerson, GetOutCar, HandShake, HugPerson, Kiss, Run, SitDown, SitUp and StandUp. Obtained from 69 different movies, this dataset contains the most challenging collection of actions, as a result of the variations in action execution and video set-up across the examples. The examples are split into 823 training and 884 test sequences, where training and test sequences are obtained from different movies.

For our experiments, we follow the experimental set-up of Laptev *et al.* [6]. We detect interest points using the spatio-temporal extension of the Harris detector

and compute descriptors of the spatio-temporal neighbourhoods of the interest points, using the specified parameters. The descriptor used here is Histogram of Optical Flow (HoF). We cluster a subset of 100,000 interest points into 4000 visual words, using k-means with the Euclidean distance, and represent each video by a histogram of visual word occurrences.

As in [6], we make use of a non-linear support vector machine with a χ^2 kernel given by, $K(H_i, H_j) = exp(-\frac{1}{2A}\sum_{n=1}^{V}\frac{(h_{in}-h_{jn})^2}{h_{in}+h_{jn}})$, where V is the vocabulary size, A is the mean distance between all training examples, and $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are histograms.

Performance is evaluated as suggested in [10]: Classification is treated as a number of one-vs-rest binary problems. The value of the classifier decision is used as a confidence score with which precision-recall curves are generated. The performance of each binary classifier is thus evaluated by the average precision. Overall performance is obtained by computing the mean Average Precision (mAP) over the binary problems.

For our RANSAC implementation, the size of the training subset, φ is chosen as one-fifth of the number of training examples, Φ. We set the number of RANSAC iterations $T = 500$, and train T *one-vs-rest* binary classifiers using φ_t against all other class examples. As detailed, reweighting is employed to select the N most inclusive groups.

Having trained using the more compact sub-categories, during testing, we obtain confidence scores from all sub-category binary classifiers for each test example. In order to obtain average precision values which combine results of multiple sub-categories within a class, we normalise the scores, such that the values are distributed over a range of $[0, 1]$, and make use of a single threshold across the multiple sub-category scores within that range. Precision-Recall curves which combine the results of the sub-categories are generated by varying this single threshold, and using the logical-OR operator across sub-categories, on the label given to each test example. In particular, for each increment of the threshold, positives, from which precision and recall values are obtained, are counted for the class if *any* one of its sub-category scores is above the threshold. Hence, a classification for a sub-category within a class is a classification for that class.

5 Results

Table 1 shows average precision obtained for each class using our method, compared with the results of Marszalek *et al.* [10]. The table shows average precision obtained using number of sub-categories $J = \{1...7\}$, and highlights the optimal value of J for each class. The table also shows the improvement obtained over [10] by splitting examples into sub-categories.

It can be seen that while six of the classes appear to be more uni-modal in their execution or setup, the remaining six benefit from the discovery of additional modes, with the actions *Eat*, *HugPerson* and *SitUp* showing best results with two modes, and *HandShake* and *GetOutCar* giving best performance with 5 and 7 modes, respectively.

Table 1. Average Precision on the Hollywood2 dataset

Action	**HoG/ HoF**[10]	**HoF + Mode Selection (Our Method)**							
		Number of Sub-categories, J							Best
		1	2	3	4	5	6	7	(#Groups)
AnswerPhone	0.088	0.086	0.130	0.144	0.153	**0.165**	0.162	0.152	**0.165**(5)
DriveCar	0.749	**0.835**	0.801	0.681	0.676	0.676	0.643	0.643	**0.835**(1)
Eat	0.263	0.596	**0.599**	0.552	0.552	0.552	0.525	0.525	**0.599**(2)
FightPerson	0.675	**0.641**	0.509	0.545	0.551	0.551	0.549	0.549	**0.641**(1)
GetOutCar	0.090	0.103	0.132	0.156	0.172	0.184	0.223	**0.238**	**0.238**(7)
HandShake	0.116	0.182	0.182	0.111	0.092	**0.190**	**0.190**	0.111	**0.190**(5)
HugPerson	0.135	0.206	**0.217**	0.143	0.129	0.134	0.134	0.120	**0.217**(2)
Kiss	0.496	**0.328**	0.263	0.239	0.253	0.263	0.101	0.091	**0.328**(1)
Run	0.537	**0.666**	0.255	0.267	0.267	0.269	0.267	0.241	**0.666**(1)
SitDown	0.316	**0.428**	0.292	0.309	0.310	0.239	0.255	0.254	**0.428**(1)
SitUp	0.072	0.082	**0.170**	0.135	0.134	0.124	0.112	0.099	**0.170**(2)
StandUp	0.350	**0.409**	0.342	0.351	0.295	0.324	0.353	0.353	**0.409**(1)
Mean	0.324								**0.407**

It should be noted that, where the number of sub-categories $J = 1$, the method simply reduces to outlier detection. In this case, examples not belonging to the largest consensus set are treated as noisy examples. As with categories which exhibit multi-modality, seeking generalisation over these examples may prove detrimental to the overall classifier performance. They are therefore discarded.

It can be observed that the improvements in performance are made for the worst performing classes without grouping. In the case of *AnswerPhone*, *GetOutCar* and *SitUp*, more than 100% improvement is observed. This shows that the low performance is due to the multi-modal nature of the examples in these classes, which is ignored without the grouping procedure. Discovering these modes and training them separately results in better performance. Conversely, breaking down of classes which performed well without grouping resulted in reduction in performance in most cases. This suggests that, in these cases, most of the actions are uni-modal. We obtain a mean average precision of 0.407 having discovered the modes of the actions, compared to 0.324 obtained in [10].

6 Conclusion

We present an approach to improving the recognition of actions in natural videos. We argue that treating all examples of a semantic action category as one class is often not optimal, and show that, in some cases, gains in performance can be achieved by identifying various modes of action execution or camera set-up. We make use of RANSAC for this grouping, but add a boosting-inspired reweighting procedure for the selection of optimal groups. Our results show that, for poorly performing classes, when different modes are trained separately, classification

accuracy is improved. This is attributed to the learning of multiple classifiers on smaller, better-defined sub-categories within each of the classes. Our approach is generic, and can be used in conjunction with existing action recognition methods, and complex datasets. Future work will include finding the optimal number of modes for each action category.

References

1. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72 (2005)
2. Fischler, M.A., Bolles, R.C.: Ransac: A paradigm for model fitting with applications to image analysis and automated cartography. Comms. of the ACM 24(6), 381–395 (1981)
3. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
4. Gilbert, A., Illingworth, J., Bowden, R.: Action recognition using mined hierarchical compound features. PAMI 99(PrePrints) (2010)
5. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: ICCV, pp. 1–8 (2009)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, pp. 1–8 (2008)
7. Laptev, I., Perez, P.: Retrieving actions in movies. In: ICCV, pp. 1–8 (2007)
8. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos "in the wild". In: CVPR, pp. 1–8 (2009)
9. Liu, J., Shah, M.: Learning human actions via information maximization. In: CVPR, pp. 1–8 (2008)
10. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR, pp. 1–8 (2009)
11. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: ICPR, pp. 32–36 (2004)
12. Szepannek, G., Schiffner, J., Wilson, J., Weihs, C.: Local modelling in classification. In: Perner, P. (ed.) ICDM 2008. LNCS (LNAI), vol. 5077, pp. 153–164. Springer, Heidelberg (2008)
13. Ullah, M.M., Parizi, S.N., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: BMVC, pp. 95.1–95.11 (2010)
14. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity-independent human pose inference. In: CVPR (2008)
15. Zhang, H., Berg, A.C., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: CVPR, pp. 2126–2136 (2006)

The Fast and the Flexible: Extended Pseudo Two-Dimensional Warping for Face Recognition

Leonid Pishchulin[1,3], Tobias Gass[2,3], Philippe Dreuw[3], and Hermann Ney[3]

[1] Computer Vision and Multimodal Computing
MPI Informatics, Saarbruecken
leonid@mpi-inf.mpg.de
[2] Computer Vision Laboratory ETH Zurich
gasst@vision.ee.ethz.ch
[3] Human Language Technology and Pattern Recognition Group,
RWTH Aachen University
lastname@cs.rwth-aachen.de

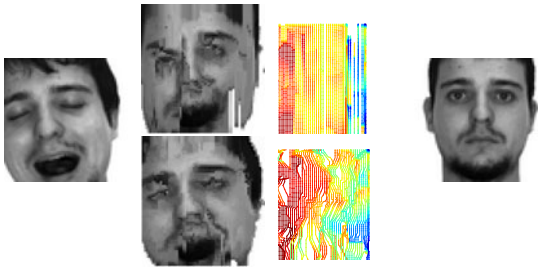
Abstract. In this work, we propose a novel extension of pseudo 2D image warping (P2DW) which allows for joint alignment and recognition of non-rectified face images. P2DW allows for optimal displacement inference in a simplified setting, but cannot cope with stronger deformations since it is restricted to column-to-column mapping. We propose to implement additional flexibility in P2DW by allowing deviations from column centers while preserving vertical structural dependencies between neighboring pixel coordinates. In order to speed up the recognition we employ hard spacial constraints on candidate alignment positions. Experiments on two well-known face datasets show that our algorithm significantly improves the recognition quality under difficult variability such as 3D rotation (poses), expressions and illuminations, and can reliably classify even automatically detected faces. We also show an improvement over state-of-the-art results while keeping computational complexity low.

1 Introduction

Fully automatic reasoning about similarity of facial images is a hard task in computer vision. Strong changes in expression and pose, as well as affine transformations stemming from automatic face detection all contribute to rich intra-class variability which is difficult to tell apart from inter-class dissimilarity.

Many methods approach the problem of intra-class variability by extracting local features from interest points or regular grids and matching them between images. The similarity is then based on the quality or the number of found matches [2, 3, 18, 22]. No geometrical dependencies between matches are considered, which makes these methods fast. However, descriptors must be chosen or trained to carry as much discriminatory information as possible which makes these methods prone to overfitting on a certain task. Even more task-specific are methods like Elastic Bunch Graph Matching [21], where faces are represented as

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 49–57, 2011.



(a) test (b) (c) (d) reference

Fig. 1. The reference image (d) is aligned to the query image (a) using P2DW (top row) and the proposed P2DW-FOSE approach (bottom row). The aligned reference image (b) shows vertical artifacts for P2DW while the proposed approach allows for much better alignment due to the flexible warping; (c) shows respective warping grids, where the colour/intensity represents the magnitude of the local deformation.

labelled graphs, and the approach of [23] who obtain pose projections by creating 3D head models from two training images per class.

Recently, increased research focus has been put on finding geometrically smooth, dense correspondences between images, which is alleviated by the availability of relatively fast, approximative energy minimization techniques for (loopy) graphs [1, 5, 9]. The complexity of these approaches is high, and the impact of the approximative optimization on the classification performance remains unclear. Contrarily, relaxing the *first-order* dependencies between neighbouring pixels leads to optimally solvable problems. [14] developed a pseudo-2D hidden Markov model (P2DHMM), where column-to-column mappings are optimised independently, leading to two separate 1D alignment problems. This idea has been extended to trees [13], allowing for greater flexibility compared to P2DHMMs at the cost of great computational complexity.

In this work, we present a novel algorithm for finding dense correspondences between images. Our approach is based on the ideas of pseudo-2D warping (P2DW) motivated by [4, 10, 14]. We show that the restriction to column-to-column mapping is insufficient for recent face recognition problems and extend the formulation to allow strip-like deviations from a central column while obeying first-order smoothness constraints between vertically neighbouring pixels (c.f. Fig. 1). This leads to an efficient formulation which is experimentally shown to work very well in practise.

We will first introduce a general formulation of two-dimensional warping (2DW) before discussing P2DW and introducing our novel algorithm. Then, we will present an experimental evaluation and finally provide concluding remarks.

2 Image Warping

In this section, we briefly recapitulate the two-dimensional image warping (2DW) as described in [19]. In 2DW, an alignment of a reference image $R \in F^{U \times V}$ to

a test image $X \in F^{I \times J}$ is searched so that the aligned or *warped* image $R' \in F^{I \times J}$ becomes as similar as possible to X. F is an arbitrary feature descriptor. An alignment is a pixel-to-pixel mapping $\{w_{ij}\} = \{(u_{ij}, v_{ij})\}$ for each position $(i, j) \in I \times J$ to a position $(u, v) \in U \times V$. This alignment defines a dissimilarity E as follows:

$$E(X, R, \{w_{ij}\}) = \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_h(w_{i-1,j}, w_{ij}) + T_v(w_{i,j-1}, w_{ij}) \Big] \ , \quad (1)$$

where $d(X_{ij}, R_{w_{ij}})$ is a distance between corresponding pixel descriptors and $T_h(\cdot)$, $T_v(\cdot)$ are horizontal and vertical smoothness functions implementing first-order dependencies between neighboring pixels. An alignment is obtained through minimization of the energy function $E(X, R, \{w_{ij}\})$. Unfortunately, finding a global minimum for such energy functions was shown to be NP-complete [7] due to cycles in the underlying graphical model representing the image lattice.

2.1 Pseudo Two-Dimensional Warping (P2DW)

In order to overcome the NP-completeness of the problem, P2DW [4, 10, 14] decouples horizontal and vertical displacements of the pixels. This decoupling leads to separate one-dimensional optimization problems which can be solved efficiently and optimally. In this case, the energy function (1) is transformed as follows:

$$\begin{aligned} E(X, R, \{w_{ij}\}) &= \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_v(v_{ij}, v_{i,j-1}) + T_h(u_i, u_{i-1}) \Big] \\ &= \sum_{i} J \cdot T_h(u_i, u_{i-1}) + \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_v(v_{ij}, v_{i,j-1}) \Big] \ , \quad (2) \end{aligned}$$

where the horizontal smoothness is only preserved between entire columns by the slightly changed term T_h. Dynamic programming (DP) techniques have been used to separately find optimal alignments between column matching candidates, and then perform an additional DP optimization in order to find the globally optimal column-to-column mapping [4].

3 Extended Pseudo-2D Warping

The simplification of horizontal dependencies not only reduces complexity of P2DW, but also decreases the flexibility of the approach since all pixels in a column are forced to have the same horizontal displacement. An example of such an alignment is demonstrated in Fig. 1(b) (top row) revealing the inability of P2DW to cope with rotation. Furthermore, scan-line artifacts are clearly visible. Column-to-column mapping degrades discriminative qualities of P2DW, which can lead to an overall decrease of recognition performance. In the following we present a flexible extension of P2DW which intends to overcome the explained shortcomings with a reasonable raise of complexity.

Strip extension. In order to overcome the limitations of the column-to-column mapping in P2DW, we propose to permit horizontal deviations from the column centers. This allows for more flexible alignments of local features within a *strip* of neighbouring columns rather than within a single column. The degree of flexibility is controlled through parameter Δ restricting the maximal horizontal deviation. This parameter is task-dependent and can be adjusted in each particular case. Setting Δ to 0 results in the original P2DW, while large values of Δ allow to compensate for noticeable image misalignments.

Especially in the last case it is important to enforce structure-preserving constraints within a strip, since otherwise one facilitates matching of similar but non-corresponding local features, which degrades the discriminative power. Therefore, we propose to model horizontal deviations from column centers while retaining the first-order dependencies between alignments in a strip, which results in a **f**irst-**o**rder **s**trip **e**xtension of P2DW (P2DW-FOSE). The first-order dependencies are modeled by hard structure-preserving constraints enforcing monotonicity and continuity of the alignment. This type of constraints was introduced in [19] in order to prevent mirroring and large gaps between aligned neighbouring pixels. Formally these constraints are expressed as follows:

$$0 \leq v_{i,j} - v_{i,j-1} \leq 2, \quad \mid u_{i,j} - u_{i,j-1} \mid \leq 1 \ . \tag{3}$$

The constraints (3) can easily be implemented in the smoothness penalty function T_v by setting the penalty to infinity if the constraints are violated. In order to decrease the complexity, we hardcode the constraints in the optimization procedure, which prevents the computation of all alignments by considering only those permitted by the constraints. This helps to greatly reduce the number of possible alignments of a coordinate given the alignments of its neighbours.

Energy function. According to the explained changes, we rewrite Eq. (2) as

$$\begin{aligned} E(X, R, \{w_{ij}\}) = &\sum_{i} J \cdot T_h(u_i, u_{i-1}) \\ &+ \sum_{ij} \Big[d(X_{ij}, R_{w_{ij}}) + T_{cv}(w_{ij}, w_{i,j-1}) + T_{\Delta}(u_i, u_{i,j}) \Big] \ . \end{aligned} \tag{4}$$

Here, T_{Δ} penalizes the deviations from the central column u_i of a strip, and $T_{\Delta} = \infty$ if $|u_i - u_{i,j}| > \Delta$; T_{cv} is the smoothness term with continuity and monotonicity constraints. In comparison to P2DW, minimization of (4) is of slightly increased complexity which is linearly dependent on the choice of parameter Δ.

Absolute displacement constraints. In order to reduce the overall complexity of the proposed approach, we restrict the absolute displacement between ij and its matching candidate w_{ij} [16]. Formally these constraints are expressed as

$$0 \leq \mid i - u_{i,j} \mid \leq W, \quad \mid j - v_{i,j} \mid \leq W \ . \tag{5}$$

The warp-range parameter W can be adjusted for each task. It can be relatively small assuming pre-aligned faces, while more challenging conditions of misaligned

Fig. 2. Sample images from AR Face (top row) and CMU-PIE (bottom row) datasets. Faces in the top row were detected by VJ, faces in the bottom were manually aligned.

faces require sufficiently large W. Absolute displacement constraints help to reduce the complexity from $O(IJUV\Delta)$ to $O(IJW^2\Delta)$ providing a significant speed-up even for a large W which is viewed as an accuracy/complexity trade-off.

Fig. 1(b) (bottom row) exemplifies the advantages of the proposed approach over the original P2DW. It can clearly be seen that the deviations from columns allow to compensate for local and global misalignments, while the implemented monotonicity and continuity constraints preserve the geometrical structure of the facial image. Both improvements lead to a visibly better quality of alignment.

We accentuate that preserving structural constraints within a strip does not guarantee global smoothness, since strips are optimised independently. The latter can lead to intersecting paths in neighbouring columns, especially for large Δ.

4 Results

We evaluate the proposed algorithm on two challenging databases with varying expressions, illuminations, poses and strong misalignments.

AR Face. Following [3], we use a subset of 110 individuals of the AR Face [12]. We use four different expressions and three illuminations, all fully taken in two sessions two weeks apart. The first session is for training, the second for testing. Simulating a real world environment we detect and crop the faces automatically to 64x64 pixels using the Viola&Jones (VJ) detector [20]. See Fig. 2 for samples.

CMU-PIE. The CMU-PIE [17] database consists of over 41000 images of 68 individuals. Each person is imaged under 43 different illumination conditions, 13 poses and 4 various facial expressions. In order to evaluate our algorithm on 3D transformations, we use a subset of all individuals in 13 poses with neutral facial expression. The original face images were manually aligned by eye-centre locations [6] and cropped to 64×64 resolution. Fig. 2 shows sample images.

Experimental Setup. We extract an 128-dimensional SIFT [11] descriptor at each position of the regular pixel grid. As proposed by [8], we reduce the descriptor to 30 dimensions by means of PCA estimated on the respective training data and subsequently normalize each descriptor to unit length. We use a NN classifier for recognition directly employing the obtained energy as dissimilarity measure and the L_1 norm as local feature distance. Similar to [5], we include

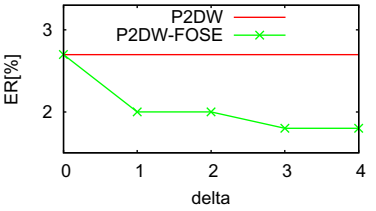


Fig. 3. Error rate on automatically detected faces for different strip widths Δ, where $\Delta = 0$ is equivalent to the P2DW

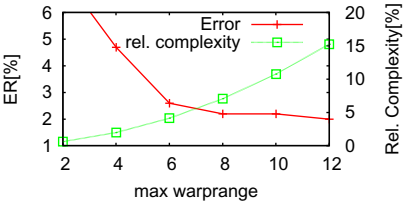


Fig. 4. Error rate on VJ detected faces and relative complexity compared to P2DW-FOSE with different warping ranges

a context of 5×5 neighboring pixels in the distance, which is also thresholded with an empirically estimated threshold value of $\tau = 1$. This makes our approach robust to unalignable pixels. Additionally, we speed up the computation of the alignments using local distance caching, and track the smallest energy obtained to stop if it is surpassed by a rough lower bound on the current energy [5]. For comparison, we use our own re-implementation of P2DW [4].

Evaluation on the AR Face database. First, we show the effects of strip width on the recognition error. Fig. 3 shows the error rate for increasing Δ where the biggest improvement is seen at $\Delta = 1$. Although the error decreases further afterwards, the return is diminishing quickly. This gives rise to two interpretations: on the one hand, it seems most important to allow (even slight) horizontal movements of individual pixels. On the other hand, big strip widths increase the chance of intersecting column paths, making the deformation less smooth.

In order to study means of speeding up the recognition, we fix $\Delta = 3$ (c.f. Fig. 3) and vary the warp-range parameter W restricting the maximum absolute displacement. Fig. 4 shows the influence of W on both recognition accuracy and computational complexity. As the total number of possible alignments grows quadratically with increasing W, the recognition error decreases until the accuracy of the unconstrained version is reached (c.f. Fig. 3). For $W = 8$, the relative complexity is 7.1%, corresponding to a speed up by a factor of 15 (in comparison to $W = \infty$) while leading to only a slight increase of the error.

In Table 1, we summarise our findings and compare relative run-times and performance of the proposed approach with basic methods and results from the literature. The last column shows a computing-time factor (CTF) relative to P2DW, which therefore has a CTF of 1 (26 s per image). It can be seen that increasing the flexibility of P2DW by means of the proposed strip extension greatly improves the accuracy. The proposed speedup allows us to use 64x64 pixels resolution, while the energy minimization technique presented in [5] operates on 32x32 pixels due to much higher complexity. Our method also greatly outperforms state-of-the-art feature matching approaches [2, 3, 18] which are though more efficient. Moreover, [3, 18] used manually pre-registered faces.

Evaluation on CMU-PIE database. To demonstrate the robustness of our approach w.r.t. to pose deformation, we evaluate our algorithm on the pose

Table 1. Results for VJ-detected faces and comparison of run-times

Model	ER [%]	CTF
No warping	22.3	-
P2DW	2.7	1
P2DW-FOSE	**1.8**	2.3
$+ W = 8$	2.0	**0.2**
CTRW-S [5]	3.7	0.4
SURF-Face [2]	4.15	-
DCT [3]	4.70*	-
Av-SpPCA [18]	6.43*	-

* with manually aligned faces

Table 2. Average error rates [%] on CMU-PIE groups of poses by our algorithms

Model	near frontal	near profile	avg.
No warping	40.69	86.27	63.48
P2DW	0.25	17.63	8.94
P2DW-FOSE	0.25	10.39	**5.32**
Hierarch. match. [1]	1.22	10.39	5.76
3D shape mod. [23]	0.00	**14.40	**6.55
Prob. learning [15]	* 7	* 32	19.30

* estimated from graphs, ** missing poses

Table 3. Qualitative evaluation of the proposed approach

Query	Ref	Deformed Ref.		Query	Ref	Deformed Query	
		P2DW	P2DW-FOSE			P2DW	P2DW-FOSE
							

subset of the CMU-PIE database, using the frontal image as reference and the remaining 12 poses as testing images. As the reference is much more accurately cropped compared to the testing images (see Fig. 2 (bottom row)), we reverse the alignment procedure and align the test image to the reference one. This helps to minimize the impact of background pixels in the test images. We also follow [1] and additionally use left and right half crops of the reference image. We choose $\Delta = 3$ and set no absolute constraints for P2DW-FOSE, as this setup was shown to lead to the best performance on the AR Face database. Recognition results on the CMU-PIE database are listed in Table 2. In order to highlight the specific difficulties of the task, we divide the test data in near frontal and near profile poses. For the former, most approaches are able to achieve error rates near to 0%, while the latter is very difficult. A clear improvement is achieved compared to P2DW, and we also obtain the best result compared to the literature, where [1] uses a much more complex warping algorithm and [23] even use an additional profile shot as training data in order to generate a 3D head model. [15] uses automatically cropped images, which make the task even harder.

Table 3 shows qualitative results on an expression and pose image: in both cases the alignment by our method is much smoother compared to P2DW.

5 Conclusion

In this work, we have shown that a flexible extension of pseudo-2D warping helps to significantly improve recognition results on highly misaligned faces with different facial expressions, illuminations and strong changes in pose. Interestingly, even small deviations from the strict column-to-column mapping allow for much smoother alignments, which in turn provides more accurate recognitions. One interesting result from our evaluation is that it pays of to sacrifice a little of the global smoothness for tractable run-time on higher-resolution images. Also, we show that our globally optimal solution to a simplified problem outperforms an hierarchical approximation of the original problem, which might suffer from local minima. We believe this is an important road to explore, since quite often problems in computer vision are made tractable by introducing heuristics such as hierarchies without clearly investigating the impact of the hidden assumptions.

References

[1] Arashloo, S., Kittler, J.: Hierarchical image matching for pose-invariant face recognition. In: BMVC (2009)
[2] Dreuw, P., Steingrube, P., Hanselmann, H., Ney, H.: Surf-face: Face recognition under viewpoint consistency constraints. In: BMVC (2009)
[3] Ekenel, H.K., Stiefelhagen, R.: Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In: CVPRW, Washington, DC, USA, p. 34 (2006)
[4] Eickeler, S., Mller, S., Rigoll, G.: High performance face recognition using pseudo 2-d hidden markov models. In: ECCV (1999)
[5] Gass, T., Dreuw, P., Ney, H.: Constrained energy minimisation for matching-based image recognition. In: ICPR, Istanbul, Turkey (2010) (in press)
[6] Gross, R.: `http://ralphgross.com/FaceLabels`
[7] Keysers, D., Unger, W.: Elastic image matching is np-complete. Pattern Recognition Letters 24, 445–453 (2003)
[8] Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: CVPR, vol. (2), pp. 506–513 (2004)
[9] Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE TPAMI 28, 1568–1583 (2006)
[10] Kuo, S.S., Agazzi, O.E.: Keyword spotting in poorly printed documents using pseudo 2-d hidden markov models. IEEE TPAMI 16(8), 842–848 (1994)
[11] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
[12] Martinez, A., Benavente, R.: The AR face database. Technical report, CVC Technical report (1998)
[13] Mottl, V., Kopylov, A., Kostin, A., Yermakov, A., Kittler, J.: Elastic transformation of the image pixel grid for similarity based face identification. In: ICPR (2002)
[14] Samaria, F.: Face Recognition Using Hidden Markov Models. PhD thesis, Cambridge University (1994)
[15] Sarfraz, M.S., Hellwich, O.: Probabilistic learning for fully automatic face recognition across pose. Image and Vision Computing 28(5), 744–753 (2010)

[16] Smith, S.J., Bourgoin, M.O., Sims, K., Voorhees, H.L.: Handwritten character classification using nearest neighbor in large databases. IEEE TPAMI 16(9), 915–919 (1994)
[17] Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: AFGR (2002)
[18] Tan, K., Chen, S.: Adaptively weighted sub-pattern pca for face recognition. Neurocomputing 64, 505–511 (2005)
[19] Uchida, S., Sakoe, H.: A monotonic and continuous two-dimensional warping based on dynamic programming. In: ICPR, pp. 521–524 (1998)
[20] Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
[21] Wiskott, L., Fellous, J.M., Krger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE TPAMI 19, 775–779 (1997)
[22] Wright, J., Hua, G.: Implicit elastic matching with random projections for pose-variant face recognition. In: CVPR, pp. 1502–1509 (2009)
[23] Zhang, X., Gao, Y., Leung, M.K.H.: Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. IEEE TIFS 3(4), 684–697 (2008)

On Importance of Interactions and Context in Human Action Recognition

Nataliya Shapovalova, Wenjuan Gong, Marco Pedersoli, Francesc Xavier Roca, and Jordi Gonzàlez

Computer Science Department and Computer Vision Center
Universitat Autònoma de Barcelona (UAB),
08193 Barcelona, Catalonia, Spain
{shapovalova,wenjuan,marcopede,xavir,poal}@cvc.uab.es

Abstract. This paper is focused on the automatic recognition of human events in static images. Popular techniques use knowledge of the human pose for inferring the action, and the most recent approaches tend to combine pose information with either knowledge of the scene or of the objects with which the human interacts. Our approach makes a step forward in this direction by combining the human pose with the scene in which the human is placed, together with the spatial relationships between humans and objects. Based on standard, simple descriptors like HOG and SIFT, recognition performance is enhanced when these three types of knowledge are taken into account. Results obtained in the PASCAL 2010 Action Recognition Dataset demonstrate that our technique reaches state-of-the-art results using simple descriptors and classifiers.

Keywords: Scene Understanding, Action Recognition, Spatial Interaction Modeling.

1 Introduction

The enormous amount of images daily generated by millions of Internet users demands robust and generic image understanding techniques for the automatic indexing and annotation of those human events displayed in pictures, for a further search and retrieval. In essence, the main goal of this Image Understanding process is to assign automatically semantic labels to images in which humans appear. This process tries to bridge the semantic gap between low-level image representation and the high-level (Natural Language) descriptions given by humans [1]. In this domain, the recognition of human activities in static images becomes of great importance, since (i) humans appear the most in the images (and videos), and (ii) knowledge about the scene and objects nearby can exploited for inferring the human action [4].

This progress on such an automatic recognition of human events in image databases have led to recent interesting approaches which take into account multiple sources of knowledge that can be found in an image, namely (i) the pose of the human body [2], (ii) the kind of scene in which the human is performing her/his action [3], and (iii) the objects with which the human interacts in order to perform the action [5].

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 58–66, 2011.

On one hand, a popular strategy for pose estimation is based on finding the proper parameters of an articulated model of the human body, like for example in [8]. The task of pose estimation is to find a parameters of the articulated model which correspond to a human in the image/video of interest. However, these type of approaches are computationally expensive and require good quality images of humans in both training and testing data, which is not always achievable. That is why other approaches are based on appearance matching alternatively [11]. This approach is purely bottom-up and the pose is represented by appearance feature vectors.

Pose-based methods for action recognition constitute successful solutions to the recognition of actions such as walking or running [2]. However, pose only is not enough for the analysis of more complex human behaviours like for example riding horse, phoning, or working with computer. In these cases, the knowledge about the scene (outdoor or indoor) and the interacting objects (horses, phones, screens) should be exploited.

The scene in which the human is performing an action can provide discriminant information: for example playing instrument is usually observed in the concerts, while working with computer can be often seen in the office environment. Scene analysis can be done in two main ways: (i) scene region segmentation (sky, road, buildings, etc.), and (ii) holistic scene classification (indoor or outdoor, mountain view or forest, bedroom or kitchen, etc.) using global appearance of the scene [6]. Segmentation methods usually provide detailed information about those regions of the scene which are particularly important for action understanding. Alternatively, obtaining a single label for the whole scene (indoor, outdoor, etc.) has been proven enough for action recognition: Marszalek et al. [3] studied the relevant scene classes and their correlation with human activities. They illustrate that incorporating scene information effectively increases the recognition rate.

The analysis of interactions of the human with other objects is of importance. For example, the aforementioned actions involve not only a human, but also the interaction with object(s) in the scene. Towards this end, there have been presented interesting methods for spatial interactions modeling [4,5,7,8]. The majority of these proposed interaction models are based on Bayesian models. These models provide coherent interference, however they are computationally expensive and require a proper initialization of the model. Authors in [9] proposed a simple and elegant solutions which models spatial interactions "on top, above, below, next to, near, far, overlap". These relationships provides high-level semantic interpretation for human-object and object-object interactions.

In this paper we propose an strategy towards the efficient combination of three sources of knowledge, i.e. human pose, scene label and object interaction. The main contribution of our work relies on taking into account the pose of the human, the interactions between human and objects, and the context. Also, an important requirement of our technique is that we do not make any assumption about the objects that can appear in the scene. In particular, this goal extends the work proposed in Li and Fei-Fei [4], where a generic model is proposed that incorporate several sources of knowledge, including event, scene and objects. Their model, restricted to sports activities only, ignores the spatial and interactive

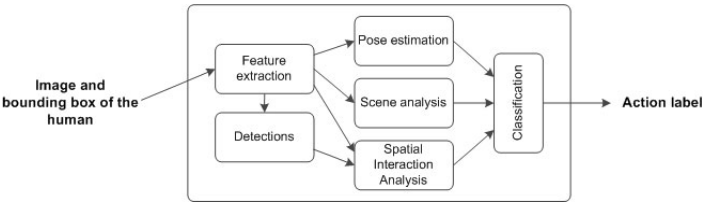


Fig. 1. Human action recognition

relationships among the objects in the scene. Our approach also extends another technique for action recognition presented by Gupta et al. [5]. They describe how to apply spatial constrains on location of objects in the action recognition. However, their model requires a strong prior knowledge in order to distinguish between manipulable and scene objects, and their approach is tested only on sport datasets, where there is no context information available.

The rest of the paper is divided as follows: Section 2 presents our novel framework for action recognition, and details the models used for pose, scene, and interaction analysis. Since we aim to demonstrate the benefits of combining these three types of knowledge, we restrict our models to be based on standard descriptors like HOG or SIFT in order to better evaluate the gain in their combination. Section 3 provide the experiment results and shows that our performance achieves state-of-the-art results in the PASCAL 2010 VOC Challenge [13]. Finally, section 4 draws the conclusions and scope for future research.

2 Human Action Recognition

The overall pipeline for the action recognition is illustrated in Fig. 1. Initially, given an image and bounding box of the human, salient features are extracted. Then, object detection, based on on the Recursive Coarse-to-Fine Localization [10], is done. Next, human pose is extracted, scene analysis is performed and the interactions between human, scene and objects are analysed. Finally, using a classification procedure, an estimation about the human action is computed. While feature extraction and object detection are only used in a straightforward manner, the pose, scene, and spatial interaction analysis are detailed next.

2.1 Pose Estimation

Pose estimation is achieved by fusing knowledge about the local appearance and the local shape of the human, which location is obtained from a bounding box, provided by the dataset. The appearance of a human pose is computed in the area of the bounding box, using the Bag-of-Words (BoW) technique. The shape representation of the human pose is done with histograms of gradient (HOG) [11], which capture edge orientation in the region of interest. In order to keep spatial information, we apply Pyramid of HOG [12]. This allows capturing local contour information as well as keeping a spatial constrains. A final human pose model H_P results from the concatenation of the appearance and shape representations.

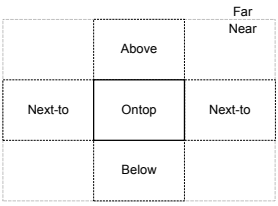


Fig. 2. Spatial histogram

2.2 Scene Model

Scene Analysis is done using SIFT features and BoW approach enhanced with a spatial pyramid presented by [6]. In our work we use a spatial pyramid over the background with two levels: zero level includes entire background region, and first level consists of three horizontal bars, which are defined by bounding box. The global scene of the image is represented with a histogram H_{BG} which is a concatenation of histograms of both levels.

2.3 Spatial Interaction

To handle spatial interactions we combine two interaction models: (i) a local interaction model and (ii) a global interaction model, adapted from [9].

Local Interaction. Local Interaction model H_{LI} is a SIFT based BoW histogram which is calculated over the local neighbourhood around the bounding box. The neighbourhood of the bounding box defines a local context that helps to analyse the interactions between a human and the objects that are being manipulated by the human.

Global Interaction. A basic description of actions in a scene can be done using information about the types of objects that are observed in the scene. Given N_O the number of object detections $O = [O_1, O_2, ..., O_{N_O}]$ in the image I, object occurrence can be represented as a histogram H_O:

$$H_O = \sum_{i=1}^{N_O} P_i u_i, \tag{1}$$

where u_i is such that only one element of u_i is nonzero, and $|u_i|$ is the $L1$-norm of u_i. The index of the only nonzero element in u_i indicates the class of the object O_i with probability P_i.

In addition, it is important incorporate the model about how these objects are distributed in the scene. The model can be obtained by analysing the interactions across all the objects in the scene. The interaction between two objects i and j can be represented by a sparse spatial interaction feature d_{ij}, which bins the relative location of the detection windows of i and j into one of the canonical semantic relations including *above*, *below*, *ontop*, *next-to*, *near*, and *far*, see Fig. 2.

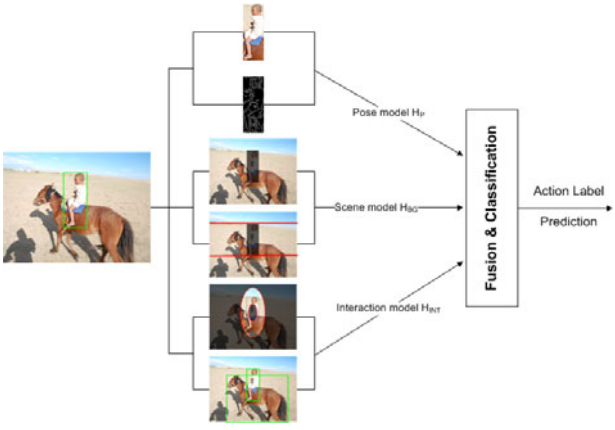


Fig. 3. Classification process

Table 1. Average precision results on PASCAL Action Dataset using different cues

	H_P	H_P& H_{BG}	H_P& H_{BG}& H_{INT}
Walking	**67.0**	64.0	62.0
Running	75.3	75.4	**76.9**
Phoning	**45.8**	42.0	45.5
Playing instrument	45.6	**55.6**	54.5
Taking photo	22.4	28.6	**32.9**
Reading	27.0	25.8	**31.7**
Riding bike	64.5	65.4	**75.2**
Riding horse	72.8	87.6	**88.1**
Using PC	48.9	62.6	**64.1**
Average	52.1	56.3	**59.0**

Therefore, every image I can be represented with an interaction matrix H_I. Every element $h_{I_{kl}}$ of the matrix H_I represents the spatial interaction between classes k and l:

$$h_{I_{kl}} = \sum_{i=1}^{N_{O^k}} \sum_{j=1}^{N_{O^l}} d(O_i^k, O_j^l) min(P_i^k, P_j^l) \tag{2}$$

where O^k and O^l are detections of objects of classes k and l correspondingly.

Therefore, the global interactions model H_{GI} is represented as the concatenation of H_O and H_I, and the final spatial interaction model H_{INT} is defined as the concatenation of the local and global interaction models, H_{LI} and H_{GI}.

2.4 Classification

In this stage, images represented with histograms are classified using a Support Vector Machine (SVM) classifier (see Fig. 3), which was trained and tested using the respective image sets. A histogram intersection kernel is used to introduce non-linearity to the decision functions. In order to fuse multiple image representations H_P, H_{BG}, H_{INT} we use concatenation of normalized histograms.

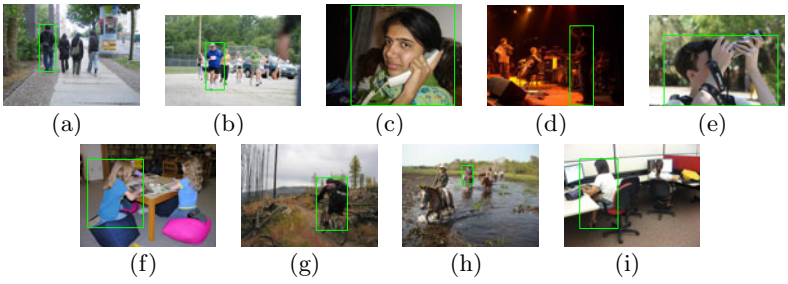


Fig. 4. Correctly classified examples of walking(a), running(b), phoning(c), playing instrument(d), taking photo(e), reading(f), riding bike(g), riding horse(h), using PC(i)

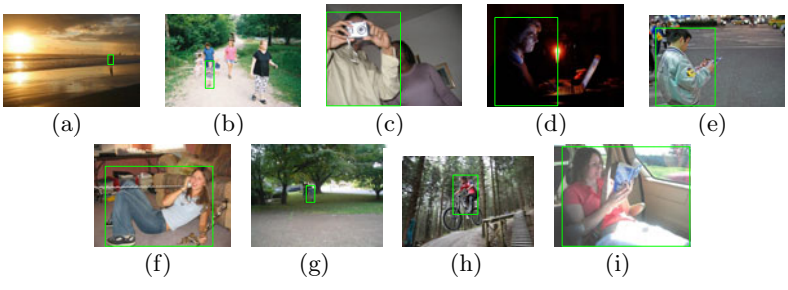


Fig. 5. Misclassified examples of walking(a), running(b), phoning(c), playing instrument(d), taking photo(e), reading(f), riding bike(g), riding horse(h), using PC(i)

3 Experimental Results

Instead of applying our technique in sport datasets as in [4,5], we tested our approach on a more challenging dataset provided by the PASCAL VOC Challenge 2010 [13]. The main feature of this dataset is that each person is annotated with a bounding box together with the activities they are performing: *phoning, playing a musical instrument, reading, riding a bicycle or motorcycle, riding a horse, running, taking a photograph, using a computer*, or *walking*. To train the spatial interaction model based on object detections we used 20 object classes: *aeroplane, bicycle, bird, boat, bus, car, cat, chair, cow, dog, dining table, horse, motorbike, person, potted plant, sheep, sofa, train,* and *tv/monitor*.

To evaluate the importance of context and interactions, three main experiments were accomplished: (i) using only pose model, (ii) using pose model and scene model, and (iii) using pose, scene analysis, and spatial interaction models, see Table 1). A selection of correctly classified and misclassified examples are illustrated in Figures 4 and 5. The complexity of the dataset is that there are *simple actions* (walking, running), *actions with unknown objects* (phoning, playing an instrument, taking a photo, reading), and *actions with known objects* (riding a bike, riding a horse, using a PC). The evaluation of results is accomplished computing precision-recall curves and average precision measures.

Table 2. Comparing average precision scores in PASCAL Action Dataset (details on these methods are found in [13])

	walking	running	phoning	playing instr.	taking photo	reading	riding bike	riding horse	using comp.	average
WILLOW LSVM	41.5	73.6	40.4	29.9	17.6	32.2	53.5	62.2	45.8	44.1
SURREY MK KDA	68.6	86.5	52.6	53.5	32.8	**35.9**	**81.0**	89.3	59.2	45.5
WILLOW SVMSIFT	56.4	78.3	47.9	29.1	26.0	21.7	53.5	76.7	42.9	48.1
WILLOW A SVMSIFT 1-A LSVM	56.9	81.7	49.2	37.7	24.3	22.2	73.2	77.1	53.7	52.9
UMCO DHOG KSVM	60.4	83.0	53.5	43.0	**34.1**	32.0	67.9	68.8	45.9	54.3
BONN ACTION	61.1	78.5	47.5	51.1	32.4	31.9	64.5	69.1	53.9	54.4
NUDT SVM WHGO SIFT C-LLM	71.5	79.5	47.2	47.9	24.9	24.5	74.2	81.0	58.6	56.6
INRIA SPM HT	61.8	84.6	53.2	53.6	30.4	30.2	78.2	88.4	60.9	60.1
CVC SEL	**72.5**	85.1	49.8	52.8	24.9	34.3	74.2	85.5	**64.1**	60.4
CVC BASE	69.2	86.5	**56.2**	56.5	25.4	34.7	75.1	83.6	60.0	60.8
UCLEAR SVM DOSP MULTFEATS	70.1	**87.3**	47.0	**57.8**	32.5	26.9	78.8	**89.7**	60.0	**61.1**
Our method: H_P & H_{BG} & H_{INT}	**62.0**	**76.9**	**45.5**	**54.5**	**32.9**	**31.7**	**75.2**	**88.1**	**64.1**	**59.0**

As we can see from the Table 1, for simple actions (*walking, running*), pose information is the most important. The minor improvements for *running* class can be explained by the fact that *running* is usually observed outdoor with groups of people, while *walking* does not have such a pattern and equally can be both indoor and outdoor, thus, adding context and interaction information decreases recognition rate.

Next, for those actions including interactions with unknown objects there is no single solution. Results of *phoning* are better when pose model are used alone; this has two explanations: (i) the typical pose is discriminative enough for this action, and (ii) the bounding box containing the human usually occupies almost the whole image, so there is not much room for the context and objects in the scene. An action like *playing an instrument* improves significantly with a scene model, since that activity often means "playing in a concert" with quite particular and distinguishable context, e.g. cluttered dark indoor scenes. Even though we can observe the increase of performance for *taking a photo*, its recognition rate is low due to the significant variations in appearance. The recognition results for the *reading* class significantly increase when adding object interaction models, as *reading* is usually observed in an indoor environment, where many objects like sofas, chairs, or tables can be detected.

Finally, the actions like *riding bike, riding horse, using PC* get significant improvement (13.5% in average per class) when we use a complete model (Pose & Scene & Interaction) comparing with results based on the pose model only. This shows the particular importance of using context and spatial object interaction information for action recognition.

Comparing our results with state-of-the art in Table 2, we might notice that our results performs in average around 3% behind the best results reported in [13]. However, in our work we used a simplified model based only on SIFT and HOG features, while the team which achieved the best results developed the action recognition framework based on 18 different variations of SIFT [13].

4 Conclusions and Future Work

In this paper, our main hypothesis is that human event recognition requires modelling the relationships between the humans and environment where the event happens. In order to asses this hypothesis, we propose an strategy towards the efficient combination of three sources of knowledge, i.e. human pose, scene appearance and spatial object interaction. The experimental results on the very recent PASCAL 2010 Action Recognition Challenge show a significant gain in recognition rate in the case our full model is used.

For the future work, we will extend our method by using more features and test our method on other datasets. Moreover, motion information should be added in order to model spatial-temporal interactions, which are of interest for video-based action recognition.

Acknowledgments. The authors wish to thank J. Gonfaus, J. van de Weijer, F.S. Khan, and A.D. Bagdanov, who helped us participating in PASCAL2010. Also, the authors acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133); along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1349–1380 (2010)
2. Ikizler, N., Duygulu, P.I.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. IVC 27(10), 1515–1526 (2009)
3. Marszałek, M., Laptev, I., Schmid, C.: Actions in Context. In: CVPR, Florida (2009)
4. Li, L.-J., Fei-Fei, L.: What, where and who? Classifying event by scene and object recognition. In: ICCV, Rio de Janeiro (2007)
5. Gupta, A., Kembhavi, A., Davis, L.S.: Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1775–1789 (2009)
6. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: CVPR, New York (2006)
7. Kjellström, H., Romero, J., Martínez, D., Kragić, D.: Simultaneous Visual Recognition of Manipulation Actions and Manipulated Objects. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 336–349. Springer, Heidelberg (2008)
8. Bangpeng, Y., Fei-Fei, l.: Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. In: CVPR, San Francisco (2010)
9. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV, Kyoto (2009)

10. Pedersoli, M., Gonzàlez, J., Bagdanov, A.D., Villanueva, J.J.: Recursive Coarse-to-Fine Localization for Fast Object Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 280–293. Springer, Heidelberg (2010)
11. Dalal, N., Triggs, B., Rhone-Alps, I., Montbonnot, F.: Histograms of oriented gradients for human detection. In: CVPR, San Diego (2005)
12. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: ACM ICIVR, Amsterdam (2007)
13. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC 2010) Results (2010), http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html

Detection Performance Evaluation of Boosted Random Ferns*

Michael Villamizar, Francesc Moreno-Noguer,
Juan Andrade-Cetto, and Alberto Sanfeliu

Institut de Robòtica i Informàtica Industrial, CSIC-UPC
{mvillami,fmoreno,cetto,sanfeliu}@iri.upc.edu

Abstract. We present an experimental evaluation of Boosted Random Ferns in terms of the detection performance and the training data. We show that adding an iterative bootstrapping phase during the learning of the object classifier, it increases its detection rates given that additional positive and negative samples are collected (bootstrapped) for retraining the boosted classifier. After each bootstrapping iteration, the learning algorithm is concentrated on computing more discriminative and robust features (Random Ferns), since the bootstrapped samples extend the training data with more difficult images.

The resulting classifier has been validated in two different object datasets, yielding successful detections rates in spite of challenging image conditions such as lighting changes, mild occlusions and cluttered background.

1 Introduction

In the last years a large number of works have been addressed for detecting objects efficiently [1–3]. Some of them use boosting algorithms with the aim of computing a strong classifier from a collection of weak classifiers which could be calculated very fast. One well-known and seminal work is the detector proposed by Viola and Jones [2]. It relies on simple but fast features (Haar-like features) that in combination, by means of AdaBoost, yields a powerful and robust classifier. Nevertheless, this sort of methods depends strongly on the number of image samples used for learning the classifier. In a few words, more training samples (object and non-object samples) means a better detector performance. This fact is an important drawback when datasets with a reduced number of images are used. Traditionally, methods tend to introduce image transformations over the training data in order to enlarge the dataset, and to collect a large number of random patches over background images as negative samples.

Recently, Random Ferns [4] in addition to a boosting phase have shown to be a good alternative for detecting object categories in an efficient and discriminative way [3]. Although this method has shown impressive results in spite of its simplicity, its performance is conditioned to the size of its training data. It is because boosting requires a

* This work was supported by the Spanish Ministry of Science and Innovation under Projects RobTaskCoop (DPI2010-17112), PAU (DPI2008-06022), and MIPRCV (Consolider-Ingenio 2010)(CSD2007-00018), and the EU CEEDS Project FP7-ICT-2009-5-95682. The first author is funded by the Technical University of Catalonia (UPC).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 67–75, 2011.

large number of training samples and because each Fern has an observation distribution that depends on the number of features forming the Fern. As larger this distribution is, more training samples are needed.

In this work, we evaluate the Boosted Random Ferns (BRFs) according to the size and quality of the training data. Furthermore, we demonstrate that adding a bootstrapping phase during the learning step the BRFs improves its detection rates. In summary, the bootstrapping phase allows to having more difficult samples with which to retraining the BRFs and therefore to obtain a more robust and discriminative object classifier.

The remainder of present work is organized as follows: Section 2 describes the approach for learning the object classifier by means of an iterative bootstrapping phase. Subsequently, the computation of the BRFs is described (Section 2.1). The procedure to bootstrapping images from the training data is shown in Section 2.2. Finally, the experimental validation and conclusions are presented in Sections 3 and 4, respectively.

2 The Approach

Detecting object instances in images is addressed by testing an object-specific classifier over images using a sliding window approach. This classifier is computed via Boosted Random Ferns, that are basically a combination of feature sets which are calculated over local Histograms of Oriented Gradients [7]. Each feature set (Random Fern) captures image texture by using the output co-occurrence of multiple local binary features [6].

Initially, the Boosted Random Ferns are computed using the original training data, but subsequently, the classifier is iteratively retrained with the aim of improving its detection performance using an extended and more difficult training data. This is done via a bootstrapping phase that collects new positive and negative samples by testing the previously computed classifier over the initial training data. As a result, a more challenging training dataset with which to train following classifiers is obtained.

2.1 Boosted Random Ferns

The addressed classifier consists on a boosting combination of Random Ferns [4]. This classifier was proposed in the past with the aim of having an efficient and discriminative object classifier [3]. More formally, the classifier is built as a linear combination of weak classifiers, where each of them is based on a Random Fern F which has associated a spatial image location g. The boosting algorithm [5] computes the object classifier $H(x)$ by selecting, in each iteration t, the Fern F_t and location g_t that best discriminates positive samples from negative ones. The algorithm also updates a weight distribution over the training samples in order to focus its effort on the hard samples, which have been incorrectly classified by previous weak classifiers. The result is an assembling of Random Ferns which are computed at specific locations. To increase even more the computational efficiency, a shared pool ϑ of features (Random Ferns) is used [6]. It allows to reduce the cost of feature computation.

Then, the Boosted Random Ferns can be defined as:

$$H(x) = \sum_{t=1}^{T} h_t(x) > \beta \ , \tag{1}$$

Algorithm 1. Boosted Random Ferns Computation

1: Given a number of weak classifiers T, a shared feature pool ϑ consisting of M Random Ferns, and N image samples $(x_1,y_1)...(x_N,y_N)$, where $y_i \in \{+1,-1\}$ indicates the label for object C and background classes B, respectively:
2: Initialize sample weights $D_1(x_i) = \frac{1}{N}$, where $i = 1,2,..N$.
3: **for** $t = 1$ to T **do**
4: **for** $m = 1$ to M **do**
5: **for** $g \in x$ **do**
6: Under current distribution D_t, calculate the weak classifier $h_{(m,g)}$ and its distance $Q_{(m,g)}$.
7: **end for**
8: **end for**
9: Select the h_t that minimizes Q_t.
10: Update sample weights.
$D_{t+1}(x_i) = \frac{D_t(x_i)\exp[-y_i h_t(x_i)]}{\sum_{i=1}^{N} D_t(x_i)\exp[-y_i h_t(x_i)]}$
11: **end for**
12: Final strong classifier.
$H(x) = sign\left(\sum_{t=1}^{T} h_t(x) - \beta\right)$

where x is an image sample, β is the classifier threshold, and h_t is a weak classifier computed by

$$h_t(x) = \frac{1}{2}\log\frac{P(F_t|C,g_t,z_t(x)=k)+\varepsilon}{P(F_t|B,g_t,z_t(x)=k)+\varepsilon}\ , \quad k = 1,..,K\ . \tag{2}$$

The variables C and B indicate the object (positive) and background (negative) classes, respectively. The Fern output z is captured by the variable k, while the parameter ε is a smoothing factor.

At each iteration t, the probabilities $P(F_t|C,g_t,z_t)$ and $P(F_t|B,g_t,z_t)$ are computed under the weight distribution of training samples D. This is done by

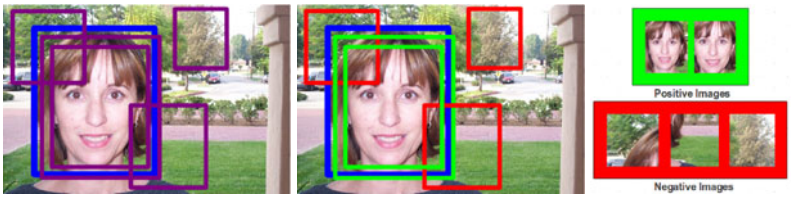
$$P(F_t|C,g_t,z_t=k) = \sum_{i:z_t(x_i)=k\wedge y_i=+1} D_t(x_i)\ , \quad k = 1,..,K\ . \tag{3}$$

$$P(F_t|B,g_t,z_t=k) = \sum_{i:z_t(x_i)=k\wedge y_i=-1} D_t(x_i)\ , \quad k = 1,..,K\ . \tag{4}$$

The classification power of each weak classifier h is measured by means of the Bhattacharyya distance Q between the object and background distributions. The weak classifier h_t that minimizes the following criterion is selected,

$$Q_t = 2\sum_{k=1}^{K}\sqrt{P(F_t|C,g_t,z_t=k)P(F_t|B,g_t,z_t=k)}\ . \tag{5}$$

The pseudocode for computing the BRFs is shown in algorithm 1. For present work, all classifiers are learned using the same classifier parameters, since our objective is to show their detection performances in connection with only the training data. Therefore,



(a) (b) (c)

Fig. 1. The Bootstrapping phase. Once the classifier is computed, it is tested over training images (a). According to the ground truth (blue rectangle) and the bootstrapping criterion (eq. 6), current detections (magenta rectangles) are classified as either positive or negatives samples (b) . Positive samples correspond to object instances (green rectangles), while negative ones are background regions (blue rectangles). They are result of false positives emitted by the classifier (c).

each boosted classifier is built by 100 weak classifiers and using a shared pool ϑ consisting of 10 Random Ferns. Besides, each Fern captures the co-occurrence of 7 random binary features that are computed over local HOGs [7].

2.2 The Bootstrapping Phase

In this section is described how the training data is enlarged with more difficult image samples in order to retraining the object-specific classifier. The procedure is carried out as follows: given the training data, consisting of sets of positive and negative samples, the classifier is first computed using those initial samples. Once the classifier has been computed, it is tested over the same training images with the aim of extracting new positive and negative samples.

The criterion for selecting the samples is based on the overlapping rate r between detections (bounding boxes) and the image ground truth given by the dataset. If that rate is over a defined threshold (0.3 by default), such detections are considered as new positive samples, otherwise, they are considered as false positives are aggregated to set of negative images. The overlapping rate can be written as:

$$r = \frac{|B_{GT} \bigcap B_D|}{|B_{GT} \bigcup B_D|} \tag{6}$$

where B_{GT} indicates the bounding box for the ground truth, and B_D is the bounding box for current detection. New positive samples represent the initial object samples under some image transformations like scales or image shifts (see Figure 3). The robustness of the classifier is then increased thanks to those images since object samples with small transformations are considered in the learning phase. By the other hand, the extracted negative samples correspond to background regions which have been misclassified by the current classifier. Those images force the boosting algorithm to seek out the most discriminative Ferns towards object classification in the following iterations. The extraction of new training samples is visualized in Figure 1.

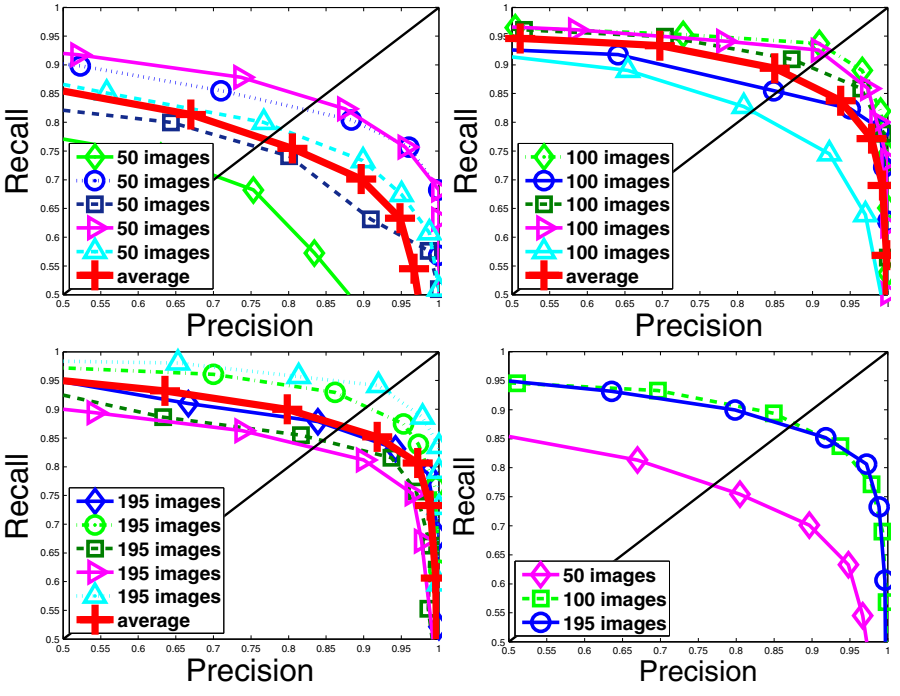


Fig. 2. Detection performance using several sizes of training data. For each case, the classifier has been learned five times due to it is based on random features. Subsequently, the average performance is calculated. Bottom-right figure shows the average curves.

3 Experiments

Several experiments have been carried out in order to show the strong dependency between the performance of the addressed classifier and the size of training data used for its learning. Moreover, it is observed that bootstrapping new samples yields more robustness and discriminative power to the object classifier. For testing the resulting classifier two different object datasets have been used. They are the Caltech Face dataset [8] and the Freestyle Motocross dataset proposed in [3].

Caltech Face Dataset - This dataset was proposed in [8] and it contains frontal faces from a small group of people. The images show extreme illumination changes because they were taken in indoor and outdoor scenes. The set of images is formed by 450 images. For present work, we use the first 195 images for training and the rest ones for validation. It is important to emphasize that images used for validation do not correspond to people used for training the classifier. In addition, this dataset also contains a collection of background images that are used as negative samples.

Freestyle Motocross Dataset - This dataset was originally proposed for validating a rotation-invariant detector given that it has two sets of images, one corresponding to motorbikes with rotations in the image plane, and a second one containing only motorbikes without rotations [3]. For our purposes the second set has been chosen. This set contains 69 images with 78 motorbike instances under challenging conditions such as extreme illumination, multiple scales and partial occlusions. The learning was done using images from the Caltech Motorbike dataset [8].

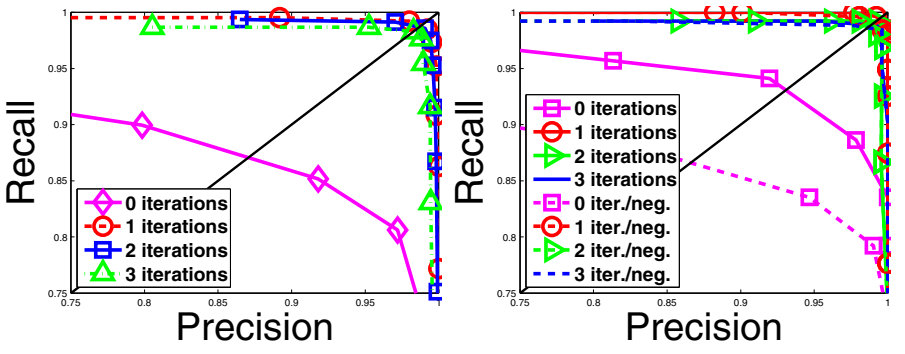
3.1 Training Data

In this section, the detection performance of Boosted Random Ferns is measured in terms of the size of the training data. Here, the classifier is learned without a bootstrapping phase and using different number of training images, both positive and negative images. The classifier aims to detecting faces in the Caltech Face dataset. Figure 2 shows the detection curves (Recall-Precision Curves) for each data size. As the classifier relies on random features (Random Ferns), the learning has been repeated five times and the average performance has been calculated. We can note that detection rates increase with the number of image samples, and that using 100 images the classifier yields similar rates to using the entire training data (195 images).

3.2 Bootstrapping Image Samples

In order to increase the number of training samples and the quality of them for the classification problem, an iterative bootstrapping phase is carried out. In each iteration, the algorithm collects new positive and negative samples from the training data. The result is generating more difficult negative samples in order to force the boosting algorithm to extract more discriminative features. Furthermore, the extraction of new positive images allows to having a wide set of images that considers small transformation over the target object (e.g object scales). To illustrate the extraction of bootstrapped samples,

Fig. 3. Bootstrapped images. Some initial positive images are shown at top row. After one bootstrapping iteration, the algorithm extracts new positive and negative samples. New positive images contain the target object at different scales and shifts (middle row). Contrary, new negative images contain difficult backgrounds and some portion of the object (bottom row).



(a) (b)

Fig. 4. Detection performance in terms of number of bootstrapping iterations. (a) The retrained classifier is evaluated over the original dataset (first experiment). (b) For detection evaluation, 2123 background images are added to set of testing images (second experiment). The classifier without retraining shows a considerable detection rate reduction when background images are added.

Figure 3 shows some new positive and negative samples which are then used for retraining the object classifier. These images were collected after just one bootstrapping iteration. Negative samples correspond to portions of faces and difficult backgrounds.

With the aim of measuring the influence of the bootstrapping phase over the detection performance, two experiments has been elaborated. The first one consists on testing the face classifier over the test images of Caltech dataset, that is, over the remainder 255 images. Similarly, the second experiment consists on evaluating the classifier over the same dataset but adding an additional set of 2123 background images in order to visualize its robustness to false positives. Both experiments have been carried out using different number of bootstrapping iterations. The results are depicted at Figure 4.

Figure 4(a) shows the results for the first experiment. We see that adding a bootstrapping phase the detection rates increase remarkably. After one iteration, the rates are very similar for all cases and achieve an almost perfect classification. For the second experiment, the testing is carried out one more time but using the large set of background images. In Figure 4(b) are shown the detection performances for each case in comparison with the results of not using the additional images. It is realized that the addition of background images for validation affects considerably the detection rate of the object classifier trained without a bootstrapping phase (0 iterations). It demonstrates that trained classifier tends to yield more false positives, while the classifiers computed after bootstrapping are robust to them. Their detection rates are very similar and do not seem to decrease with the extended testing set.

3.3 Freestyle Motocross Dataset

In Figure 5(a) are shown the detection curves for the Freestyle Motocross dataset. It is seen how the number of bootstrapping iterations also improves the detection rates. Figure 5(b) depicts the detection performances of learning the object classifier using

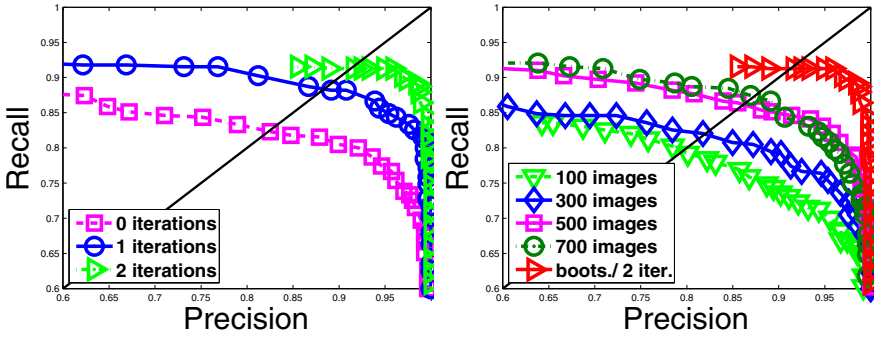


Fig. 5. Detection curves for the Freestyle Motocross dataset. (a) The bootstrapping phase improves the detection classification. (b) The bootstrapped classifier outperforms to traditional classifiers based on creating and collecting a large number of image samples.

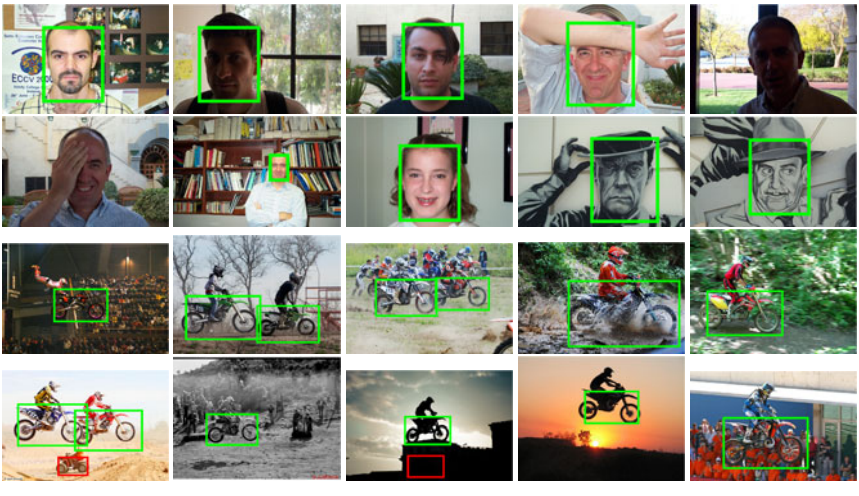


Fig. 6. Some detection results. Green rectangles correspond to object detections. The resulting classifier is able to detect object instances under difficult image conditions.

different sets of training images. In this experiment, the positive images are generated by adding affine transformations over the initial positive samples, while the negative ones are created by collecting random patches over background images. The bootstrapped classifier outperforms in all cases. This demonstrates that bootstrapped images help to the boosting algorithm to build a more robust classifier. Finally, some detection results for this dataset are shown in Figure 6.

4 Conclusions

We have presented an experimental evaluation of the Boosted Random Ferns as an object classifier. The strong dependency between detection performance and the size of training data has been evidenced. To improve detection results, an iterative bootstrapping phase has been presented. It allows to collect new positive and negative samples from the training data in order to force the learning algorithm to compute more discriminative features towards object classification. Furthermore, we can conclude that with just one bootstrapping iteration, the bootstrapped classifier achieves impressive detection rates and robustness to false positives.

References

1. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. PAMI (2007)
2. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
3. Villamizar, M., Moreno-Noguer, F., Andrade-Cetto, J., Sanfeliu, A.: Efficient rotation invariant object detection using boosted random ferns. In: CVPR (2010)
4. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: CVPR (2007)
5. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Machine Learning (1999)
6. Villamizar, M., Moreno-Noguer, F., Andrade-Cetto, J., Sanfeliu, A.: Shared random ferns for efficient detection of multiple categories. In: ICPR (2010)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
8. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)

Feature Selection for Gender Classification

Zhihong Zhang and Edwin R. Hancock

Department of Computer Science, University of York, UK

Abstract. Most existing feature selection methods focus on ranking features based on an information criterion to select the best K features. However, several authors find that the optimal feature combinations do not give the best classification performance [6],[5]. The reason for this is that although individual features may have limited relevance to a particular class, when taken in combination with other features it can be strongly relevant to the class. In this paper, we derive a new information theoretic criterion that called multidimensional interaction information (MII) to perform feature selection and apply it to gender determination. In contrast to existing feature selection methods, it is sensitive to the relations between feature combinations and can be used to seek third or even higher order dependencies between the relevant features. We apply the method to features delivered by principal geodesic analysis (PGA) and use a variational EM (VBEM) algorithm to learn a Gaussian mixture model for on the selected feature subset for gender determination. We obtain a classification accuracy as high as 95% on 2.5D facial needlemaps, demonstrating the effectiveness of our feature selection methods.

1 Introduction

High-dimensional data pose a significant challenge for pattern recognition. The most popular method for reducing dimensionality are variance based subspace methods such as PCA. However, the extracted PCA feature vectors only capture sets of features with a significant combined variance, and this renders them relatively ineffective for classification tasks. Hence it is crucial to identify a smaller subset of features that are informative for classification and clustering. The idea underpinning feature selection is to select the features that are most relevant to classification while reducing redundancy. Mutual information (MI) provides a principled way of measuring the mutual dependence of two variables, and has been used by a number of researchers to develop information theoretic feature selection criteria. For example, Batti [2] has developed the Mutual Information-Based Feature Selection (MIFS) criterion, where the features are selected in a greedy manner. Given a set of existing selected features S, at each step it locates the feature x_i that maximize the relevance to the class $I(x_i; C)$. The selection is regulated by a proportional term $\beta I(x_i; S)$ that measures the overlap information between the candidate feature and existing features. The parameter β may significantly affect the features selected, and its control remains an open problem. Peng et al [9] on the other hand, use the so-called Maximum-Relevance

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 76–83, 2011.

Minimum-Redundancy criterion (MRMR), which is equivalent to MIFS with $\beta = \frac{1}{n-1}$. Yang and Moody's [11] Joint Mutual Information (JMI) criterion is based on conditional MI and selects features by checking whether they bring additional information to an existing feature set. This method effectively rejects redundant features. Kwak and Choi [8] improve MIFS by developing MIFS-U under the assumption of a uniform distribution of information for input features. It calculates the MI based on a Parzen window, which is less computationally demanding and also provides better estimates.

However, there are two limitations for the above MI feature selection methods. Firstly, they assume that every individual relevant feature should be dependent with the target class. This means that if a single feature is considered to be relevant it should be correlated with the target class, otherwise the feature is irrelevant [4]. So only a small set of relevant features is selected, and larger feature combinations are not considered. Hence, although a single feature may not be relevant, when combined with other features it can become strongly relevant. The second weakness is that most of the methods simply consider pairwise feature dependencies, and do not check for third or higher order dependencies between the candidate features and the existing features. For example, [1] there are four features X_1,X_2,X_3,X_4, the existing selected feature subset is $\{X_1,X_4\}$, assume I(X_2,C)= I(X_3,C), I(X_2,X_1|C) = I(X_3,X_1|C), I(X_2,X_4|C)= I(X_3,X_4|C) and I(X_1,X_4,X_2)$\gg$ I(X_1,X_2)+ I(X_4,X_2), which indicates that X_2 has strong affinity with the joint subset $\{X_1,X_4\}$, although has smaller individual affinity to each of them. So in this situation, X_2 may be discarded, and X_3 is selected, although the combination $\{X_1,X_4,X_2\}$ can produce a better cluster than $\{X_1,X_4,X_3\}$.

To overcome this problem in this paper, we introduce the so called multidimensional interaction information (MII) to select the optimal subset of features. This criterion is capable of detecting the relationships between higher order feature combinations. In addition, we apply a mixture Gaussian model and the variational EM algorithm to the selected feature subset to detect clusters. We apply the method to gender classification based on facial needle-maps extracted from the Max-Planck database of range images.

2 Information-Theoretic Feature Selection

Mutual Information: In accordance with Shannon's information theory [10], the uncertainty of a random variable C can be measured by the entropy $H(C)$. For two variables X and C, the conditional entropy $H(C|X)$ measures the remaining uncertainty about C when X is known. The mutual information (MI) represented by $I(X;C)$ quantifies the information gain about C provided by variable X. The relationship between $H(C)$,$H(C|X)$ and $I(X;C)$ is $I(X;C) = H(C) - H(C|X)$.

For training a classifier, we prefer features which can minimize the uncertainty on the output class set C. If $I(X;C)$ is large, this implies that feature vector X and output class set C are closely related. When X and C are independent,

the MI of X and C goes to zero, and this means that the feature X is irrelevant to class C. As defined by Shannon, the initial uncertainty in the output class C is expressed as:

$$H(C) = -\sum_{c \in C} P(c) \log P(c) \ . \tag{1}$$

where $P(c)$ is the prior probability over the set of class C. The remaining uncertainty in the class set C if the feature vector X is known is defined by the conditional entropy $H(C|X)$

$$H(C|X) = -\int_x p(x) \{ \sum_{c \in C} p(c|x) \log p(c|x) \} dx \ . \tag{2}$$

where $p(c|x)$ denotes the posterior probability for class c given the input feature vector x. After observing the feature vector x, the amount of additional information gain is given by the mutual information (MI)

$$I(X;C) = H(C) - H(C|X) = \sum_{c \in C} \int_x p(c,x) log \frac{p(c,x)}{p(c)p(x)} dx \ . \tag{3}$$

Conditional Mutual Information: Assume that S is the set of existing selected features, X is the set of candidate features, $S \cap X = \emptyset$, and C is the output class set. The next feature in X to be selected is the one that maximizes $I(C; x_i|S)$, i.e. the conditional mutual information (CMI) which can be represented as $I(C; x_i|S) = H(C|S) - H(C|x_i, S)$, where C is the output class set, S is the selected feature subset, X is the candidate feature subset, and $x_i \in X$. From information theory, the conditional mutual information is the expected value of the mutual information between the candidate feature x_i and class set C when the existing selected feature set S is known. It can be also rewritten as

$$I(C; x_i|S) = E_S(I(C; x_i)|S) = \sum_S \sum_{c \in C} \sum_{x_i \in X} P(x_i, S, c) \log \frac{P(S)P(x_i, S, c)}{P(x_i, S)P(S, c)} \ . \tag{4}$$

Multidimensional Interaction Information for Feature Selection: The conditioning on a third random variable may either increase or decrease the original mutual information. That is, the difference $I(X;Y|Z) - I(X;Y)$, referred to as the interaction information and represented by $I(X;Y;Z)$, can measure the difference between the original mutual information $I(X;Y)$ when a third random variable is taken into account or not. The difference may be positive, negative, or zero, but it is always true that $I(X;Y|Z) \geqslant 0$.

Given the existing selected feature set S, the interaction information between the output class set and the next candidate feature x_i can be defined as $I(C; x_i; S) = I(C; x_i|S) - I(C; x_i)$. From this equation, the interaction information measures the influence of the existing selected feature set S on the amount of information shared between the candidate feature x_i and the output class set C, i.e. $\{C, x_i\}$. A zero value of $I(C; x_i; S)$ means that the information

contained in the observation x_i is not useful for determining the output class set C, even when combined with the existing selected feature set S. A positive value of $I(C; x_i; S)$ means that the observation x_i is independent of the output class set C, so $I(C, x_i)$ will be zero. However, once x_i is combined with the existing selected feature set S, then the observation x_i immediately becomes relevant to the output class set C. As a result $I(C; x_i|S)$ will be positive. As a result the interaction information is capable of solving *XOR gate* type classification problems. A negative value of $I(C; x_i; S)$ indicates that the existing selected feature set S can account for or explain the correlation between $I(C; x_i)$. As a result the shared information between $I(C; x_i)$ is decreased due to the additional knowledge of the existing elected feature set S.

According to the above definition, we propose the following multidimensional interaction information for feature selection. Assume that S is the set of existing selected feature sets, X is the set of candidate features, $S \cap X = \emptyset$, and C is the output class set. The objective of selecting the next feature is to maximize $I(C; x_i|S)$, defined by introducing the multidimensional interaction information:

$$I(C; x_i|S) = I(C; x_i) + I(\{x_i, s_1 \dots s_m, C\}) \ . \tag{5}$$

where

$$I(C; x_i, S) = I(x_i, s_1, \dots, s_m; C) = \sum_{s_1,\dots,s_m} \sum_{c \in C} P(x_i, s_1, \dots, s_m; c) \times \log \frac{P(x_i, s_1, \dots, s_m; c)}{P(x_i, s_1, \dots, s_m)P(c)} \ . \tag{6}$$

Consider the joint distribution $P(x_i, S) = P(x_i, s_1, \dots, s_m)$. By the chain rule of probability, we expand $P(x_i, S)$, $P(x_i, S; C)$ as

$$P(x_i, s_1, \dots, s_m) = P(s_1)P(s_2|s_1)P(s_3|s_2, s_1) \times P(s_4|s_3, s_2, s_1) \cdots P(x_i|s_1, s_2 \dots s_m) \ , \tag{7}$$

$$P(x_i, S; C) = P(x_i, s_1, s_2 \dots s_m; C) = P(C)p(s_1|C)P(s_2|s_1, C)P(s_3|s_1, s_2, C) \times P(s_4|s_1, s_2, s_3, C) \cdots P(x_i|s_1, \dots, s_m, C) \ . \tag{8}$$

There are two key properties of our proposed definition in Equation 5. The first is that the interaction information term $I(\{x_i, s_1 \dots s_m, C\})$ which can be zero, negative and positive. It can deal with a variety of cluster classification problems including the *XOR gate*. When it taken on a negative value, it can help to select optimal feature sets. The second benefit is its multidimensional form, compared to most existing MI methods which only check for pairwise feature interactions. Our definition can be used to check for third and higher order dependencies between features.

We use a greedy algorithm to locate the k optimal features from an initial set of n features. Commencing with a complete set X and an empty set S, we select a feature which gain a large value of $I(x, C)$. We then apply the greedy strategy to select the feature that maximizes the multidimensional mutual information between the features and the output class set, i.e. the first feature

X'_{max} that maximizes $I(X', C)$, the second selected feature X''_{max} that maximizes $I(X'', X', C)$, the third feature X'''_{max} that maximizes $I(X''', X'', X', C)$. We repeat this procedure until $|S| = k$.

3 Experimental Results

2.5D Facial Needle-maps: The data set used to test the performance of our proposed algorithm is the facial needle-maps extracted from the Max-Planck database of range images, see Fig. 1. It comprises 200 (100 females and 100 males) laser scanned human heads without hair. The facial needle-maps (fields of surface normals) are obtained by first orthographically projecting the facial range scans onto a frontal view plane, aligning the plane according to the eye centers, and then cropping the plane to be 256-by-256 pixels to maintain only the inner part of the face. The surface normals are then obtained by computing the height gradients of the aligned range images. We then apply the principal geodesic analysis (PGA) method to the needle maps to extract features. Whereas PCA can be applied to data residing in a Euclidean space to locate feature subspaces, PGA applies to data constrained to fall on a manifold. Examples of such data are direction fields, and tensor-data.

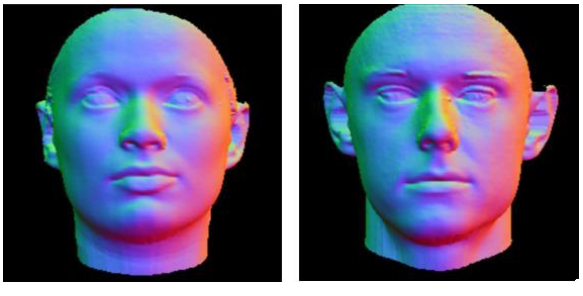
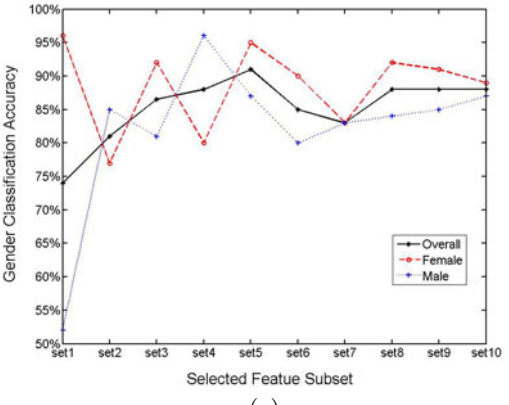


Fig. 1. Examples of Max - Planck needle - maps

Gender Classification Results: Using the feature selection algorithm outlined above, we first examine the gender classification performance using different sized feature subsets for the leading 10 PGA features. Then, by selecting a different number of features from the leading 30 PGA feature components, we compare the gender classification results from our proposed method Multidimensional Interaction Information (MII) with those obtained using alternative existing MI-based criterion methods, namely a) Maximum-Relevance Minimum-Redundancy (MRMR), b) Mutual Information Based Feature Selection (MIFS). The data used are the PGA feature vectors extracted from the 200 Max-Planck facial needle maps.

First, we explore the leading 10 PGA feature components. We have 10 selected feature subsets of size d ($d = 1,\ldots,10$) shown in Fig. 2 (b). For each d, we apply

the variational EM algorithm [3] to fit a mixture of Gaussians model to the d-dimensional feature vectors of the 200 needle maps. After learning the mixture model, we use the a posteriori probability to perform gender classification. The gender classification accuracies obtained on different feature subsets are shown in Fig. 2 (a). From Fig. 2, it is clear that the best classification accuracy for the facial needle-maps is achieved when 5 features are selected. Using fewer or more features both deteriorate the accuracy. The main reason for this is that we select the gender discriminating features only from the leading 10 PGA features. On the base of cumulative reason, there are probably additional non leading features are important for gender discrimination. Therefore, in the following experiments, we extend our attention to the leading 30 features.



(a)

Set1	(1)
Set2	(1,6)
Set3	(1,6,2)
Set 4	(1,6,2,5)
Set5	(1,6,2,5,9)
Set6	(1,6,9,8,10,4)
Set7	(1,6,2,9,3,8,7)
Set8	(1,6,2,9,3,8,7,5)
Set9	(1,6,2,9,3,8,7,5,4)
Set10	(1,6,2,9,3,8,7,5,4,10)

(b)

Fig. 2. Gender Classification on Facial needle-maps using different selected feature subsets

In Fig. 3, we compare the the performance of the three criterion functions. At small dimensionality there is little difference between the different methods. However, at higher dimensionality, the features selected by MII clearly have a higher discriminability power than the features selected by MRMR and MIFS. The ideal size of the feature subsets for MIFS and MRMR is 4 or 5, where an accuracy 87% and 92.5% respectively are achieved. However, for MII we obtain the highest accuracy of 95% when nearly 10 features are selected. This means that with both both MRMR and MIFS there is a tendency to overestimate the redundancy between features, since they neglect the conditional redundancy term $I(x_i, S|C)$. As a result some important features can be discarded, which in turn leads to information loss.

Our proposed method therefore leads to an increase in performance at high dimensionality. By considering higher order dependencies between features, we avoid premature rejection on the basis of redundancy, a well documented problem known to exist in MRMR and MIFS [7]. This gives feature sets which are more informative to our classification problem.

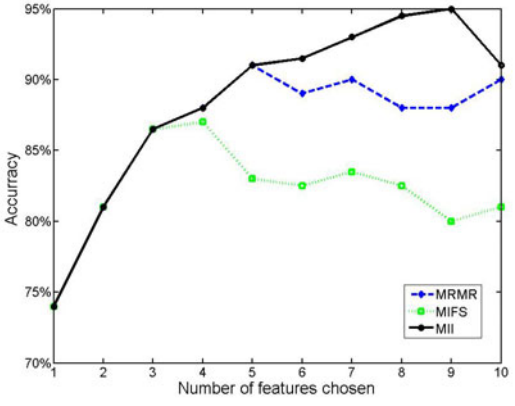
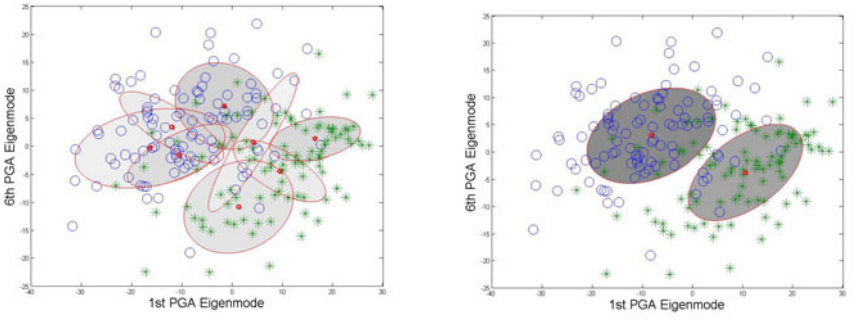


Fig. 3. Average classification accuracies:MII show significant benefit compared to criteria of MRMR and MIFS measuring only pairwise dependencies



(a) Initialization (b) Convergence

Fig. 4. VBEM of $K = 8$ learning process visualized on $1st$ and $6th$ PGA feature components, in which the ellipses denote the one standard-deviation density contours for each components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The initial number of component is $K = 8$, during the learning process, some components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted

We illustrate the learning process and the classification results of using the selected feature subset. The learning process of the EM step in the VBEM algorithm is visualized in Fig. 4 by showing the models learned, a) on initialization, b) on convergence. Each of the two stages is visualized by showing the distribution of the 1st and 6th PGA feature components as a scatter plot. From Fig. 4, it is clear that on convergence of the VBEM algorithm (after 50 iterations), the data are well clustered according to gender. We see that after convergence, only two components survive. In other words, there is an automatic trade-off in

the Bayesian model between fitting the data and the complexity of the model, in which the complexity penalty arises from components whose parameters are forced away from their prior values.

4 Conclusions

This paper presents a new information criteria based on Multidimensional Interaction information for feature selection. The proposed feature selection criterion offers two major advantages. First, the MII criteria takes into account high-order feature interactions, overcoming the problem of overestimated redundancy. As a result the features associated with the greatest amount of joint information can be preserved. Second, the variational EM algorithm and a Gaussian mixture model are applied to the selected feature subset improving the overall classification.

References

1. Balagani, K., Phoha, V., Iyengar, S., Balakrishnan, N.: On Guo and Nixon's Criterion for Feature Subset Selection: Assumptions, Implications, and Alternative Options. IEEE TSMC-A: Systems and Humans 40(3), 651–655 (2010)
2. Battiti, R.: Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Transactions on Neural Networks 5(4), 537–550 (2002)
3. Bishop, C.: Pattern Recognition and Machine Learning, vol. 4. Springer, New York (2006)
4. Cheng, H., Qin, Z., Qian, W., Liu, W.: Conditional Mutual Information Based Feature Selection. In: IEEE International Symposium on Knowledge Acquisition and Modeling, pp. 103–107 (2008)
5. Cover, T.: The Best Two Independent Measurements Are Not The Two Best. IEEE TSMC 4(1), 116–117 (2010)
6. Cover, T., Thomas, J., Wiley, J.: Elements of Information Theory, vol. 1. Wiley-Interscience, Hoboken (1991)
7. Gurban, M., Thiran, J.: Information Theoretic Feature Extraction for Audio-visual Speech Recognition. IEEE Transactions on Signal Processing 57(12), 4765–4776 (2009)
8. Kwak, N., Choi, C.: Input Feature Selection by Mutual Information Based on Parzen Window. IEEE TPAMI 24(12), 1667–1671 (2002)
9. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. IEEE TPAMI 27(8), 1226–1238 (2005)
10. Shannon, C.: A Mathematical Theory of Communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 3–55 (2001)
11. Yang, H., Moody, J.: Feature Selection Based on Joint Mutual Information. In: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis, pp. 22–25 (1999)

Classification of Repetitive Patterns Using Symmetry Group Prototypes*

Manuel Agustí-Melchor, Angel Rodas-Jordá,
and Jose-Miguel Valiente-González

Computer Vision Group, DISCA, Escuela Técnica Superior de Ingeniería Informática,
Universitat Politècnica de València, Camino de Vera, s/n 46022 Valencia, Spain
{magusti,arodas,jvalient}@disca.upv.es

Abstract. We present a novel computational framework for automatic classification method by symmetries, for periodic images applied to content based image retrieval. The existing methods have several drawbacks because of the use of heuristics. These methods have shown low classification values when images exhibit imperfections due to the fabrication or the hand made process. Also, there is no way to give some computation of the *classification goodness-of-fit*. We propose to obtain an automatic parameter estimation for symmetry analysis. Thus, the image classification is redefined as distances computation to the prototypes of a set of defined classes. Our experimental results improves the state of the art in wallpaper classification methods.

Keywords: Symmetry, symmetry groups, image retrieval by symmetry, prototype-based classification, adaptive nearest neighbour classification.

1 Introduction

In industrial sectors, the notion of symmetry is always present as an aesthetic element, indispensable in each new design. These designs, see examples in Fig. 1, are commonly referred to as *regular mosaics*, *wallpaper images*, *wallpaper patterns*, or simply *Wallpapers*. Accumulated over the years, thousands of these samples are stored in company storerooms or museums, in many cases in digital formats. But designers suffer serious limitations when searching and managing these images, and designers are accustomed to using abstract terms related with perceptual criteria. Therefore, some kind of image analysis is necessary to extract information about the internal geometry and structure of patterns. Little effort has been made in their image analysis and classification, and so this work explores this direction. In this way, Content-Based Image Retrieval Systems (CBIR) can be used to find images that satisfies some criteria based on similarity. Therefore, we propose: i) to solve the inherent ambiguities of symmetry computation by using an adaptive classifier without needing learning stages; ii) to use this classifier for CBIR, because the relative response of this classifier allows to recover groups of similar images.

* This work is supported in part by spanish project VISTAC (DPI2007-66596-C02-01).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 84–91, 2011.

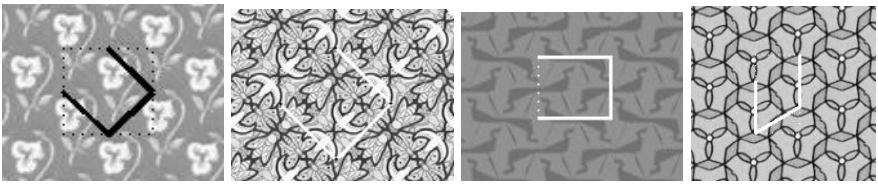


Fig. 1. Details of wallpaper images obtained from [10], [2], and [4] collections. These are images of real textile samples. The geometry of the UL is also shown.

A symmetry of any 2D pattern can be described through a set of geometrical transformations that transforms it in itself: isometries. These are: translations, rotations (n-fold), reflections (specular), and glide reflections (specular plus lateral displacement). Regarding translational symmetry, a wallpaper pattern is a repetition of a parallelogram shaped subimage, called *Unit Lattice* (UL). A lattice extraction procedure is then needed to obtain the lattice geometry. In this work we assume that the lattice geometry has been already obtained and the UL is known, e.g. using autocorrelation [5], Fourier, or wavelets [1]. We rely on the symmetry groups theory [3] to formulate a mathematical description of structures. A symmetry group is the set of isometric transformations that brings a figure in line with itself. In the 2D case, a *Plane Symmetry Groups* (PSG) is defined by means of two translational symmetries (lattice) and some other isometries. For example, Fig. 1 (left) is only translational. In contrast, the other patterns of Fig. 1 have more isometries, such as 180° rotations and reflections about two axes. The last pattern can be reconstructed using 120° rotations. The *crystallographic constraint* limits the number of PSG to 17 cases, helping us to describe the pattern structure. Fig. 2 shows the details of each PSG as well as their standard notation. For example, the patterns in Fig. 1 belong, from left to right, to symmetry groups P1, PMM, PGG and P31M.

The interest in the algorithmic treatment of symmetries has been recognized by a recent tutorial at ECCV 2010 Conference [6]. It included an extended discussion and comparison of the state of the art of symmetry detection algorithms.

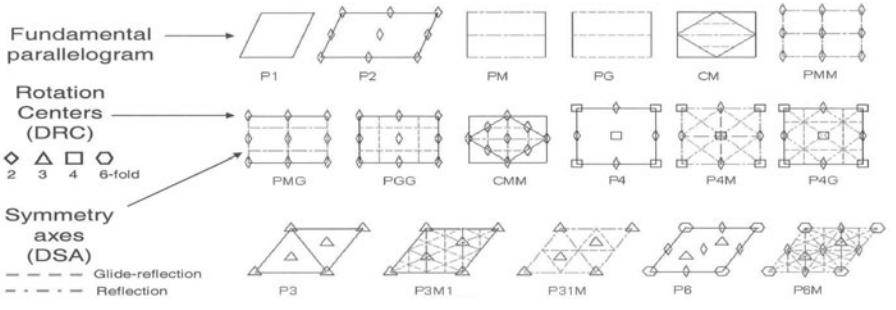


Fig. 2. Representation of the 17 wallpaper groups, their standard notation and their internal symmetries. The UL is referred as a fundamental parallelogram.

A global success of 63% over a test bed of 176 images is reported. The classical algorithms for cataloguing wallpapers [3] are heuristic-based procedures to be used by humans. These are proposed in the form of decision trees whose branches are raised as questions to ask when looking at the design. Because of the ambiguities in the computer reformulation of these questions, their implementation is very complex. The main computer autonomous approach of this kind has been made by Liu et al. [5]. This model expresses the Schattschneider algorithm in the form of a *rule-based classifier* (RBC) where each symmetry group corresponds to a unique sequence of yes/no answers. This model can be seen as a kind of *decision tree* classifier with binary symmetry values. Our experience confirms that this method can be tuned to obtain 100% hits for the Joyce [4] dataset, but this set-up is not successful for other image collections. In addition, the use of RBC obtains only one result without an associate measure of confidence. An enhancement of this solution is necessary, as indicated in [7].

2 Proposed Method

We propose a novel computational framework based on continuous measurements of symmetries for a distance-based classification of the symmetry groups approach applied to real and synthetic images of two-dimensional repetitive patterns. The use of binary class prototypes describing the PSG mentioned above, adaptively adjusted for image conditions, assumes a high degree of variance of symmetry features, due to noise, deformations or just the nature of the hand made products. As this classification results in an ordered list of content similarity based on symmetry, it can be used as an result for Context-Based Image Recovery (CBIR) applications. We started by using a Nearest Neighbour Classifier (NNC) as this enabled us to obtain a measure of goodness for the classification. This kind of method requires the use of continuous-value feature vectors.

2.1 Feature Computation and Symmetry Groups Classification

A close view to PSG description in Fig. 2 reveals that the minimum number of symmetry features needed to distinguish every PSG is twelve: four (R_2, R_3, R_4, and R_6) related to rotational symmetries; four (T_1, T_2, T_{1G}, and T_{2G}) to describe reflection symmetries (non-glide and glide) along axes parallel to the sides of UL; and four more features (D_1, D_2, D_{1G}, and D_{2G}) for reflection (non-glide and glide) with respect to the two UL diagonals. We defined a symmetry feature vector (SFV) of twelve elements that identifies the presence/absence of these symmetries as $(f_1, f_2, ..., f_{12})$. To obtain a symmetry feature f_i for a specific isometry, e.g 2-fold rotation, we apply this transformation to the original image $I(x, y)$ obtaining the transformed image $I^T(x, y)$. A piece of the transformed image, of the size of the bounding box of the UL (m), is selected. A score map is then computed as $Map(x, y) = 1 - SAD(x, y)$, where:

$$SAD(x,y) = \frac{1}{m} \sum_{x_0 \ y_0} |I(x - x_0, y - y_0) - BBox(x_0, y_0)| \qquad (1)$$

If symmetry is present in the image, this map peaks at several positions indicating the presence of that symmetry, while revealing lower values in areas where the symmetry is not hold. The $|maximum - minimum|$ difference should then be a good measure to quantify the feature. However, there are patterns without internal symmetries, such as P1 (Fig. 1), so that max-min difference should be relative to any other value representing the presence of symmetry. The only symmetry always present in every pattern is the translational symmetry (S_T). Finally, we compute the normalized components of the SFV as follows:

$$f_i = \frac{max(Map) - min(Map)}{S_T - min(Map)} \quad 1 \leq i \leq 12 \tag{2}$$

The higher the value of f_i, the more likely the image contains symmetry. Table 1 shows the SFV vectors obtained for the four wallpaper samples in Fig. 1. As expected, these results partially confirm high values that indicate the presence of symmetry and low values otherwise. The bold values means a value that has to be considered as presence of symmetry when considering each vector as the group that it belongs to, while the others mean absence of others symmetries. Because these features were computed as gray level differences between image patches, their values will strongly depend on the particular arrangements of image pixels: the *image complexity*. As a consequence SFV requires a higher level of adaptation to the image conditions, i.e. taking into account the contents of each image separately. This idea will be used later by an adaptive NNC.

Table 1. Symmetry feature vectors of the four wallpapers shown in Fig. 1

Sample	SFV=(R_2, R_3, R_4, R_6, T_1, T_2, D_1, D_2, T_{1G}, T_{2G}, D_{1G}, D_{2G})	PSG
1	(0.62, 0.47, 0.69, 0.34, 0.65, 0.67, 0.80, 0.59, 0.37, 0.43, 0.80, 0.59)	$P1$
2	(**0.82**, 0.09, 0.20, 0.09, **0.88**, **0.83**, 0.20, 0.19, 0.27, 0.26, 0.2, 0.19)	PMM
3	(**0.95**, 0.42, 0.33, 0.46, 0.39, 0.45, 0.31, 0.48, **0.98**, **0.99**, 0.31, 0.48)	PGG
4	(0.46, **0.69**, 0.28, 0.49, 0.74, 0.65, 0.48, **0.72**, **0.74**, **0.65**, 0.48, 0.72)	$P31M$

To classify a wallpaper image, featured by SFV, we need a set of class samples. Fortunately, the number of classes and their structure are known in advance. For the sake of simplicity, we start by proposing the use of binary prototypes representing each one of the classes. Table 2 shows the resulting 23 prototypes. Some classes have two prototypes because there are two possible places where reflection symmetry can appear. After applying the NNC to several image collections we did not found significant improvements in comparison with RBC (see the Experiments section). This is probably due to the bias of the feature values: minimum values are not near '0', nor maximum values are near S_t. In that situation, the use of binary prototypes, with inter-class boundaries equidistant to each class, does not fit the problem. However, some advantage have been achieved. Firstly, the Euclidean distance to the class prototype can be used as a measure of confidence. Secondly, the NNC produces an ordered set of outputs describing the class membership of each sample. This latter consideration can enable an automatic adjustment of the prototypes in order to adapt them to the image variability.

Table 2. Binary prototypes for the 17 PSG classes

Classes	Prototype Feature vectors
$P1$	(0,0,0,0,0,0,0,0,0,0,0,0)
$P2$	(1,0,0,0,0,0,0,0,0,0,0,0)
PM_1	(0,0,0,0,1,0,0,0,0,0,0,0)
PM_2	(0,0,0,0,0,1,0,0,0,0,0,0)
PG_1	(0,0,0,0,0,0,0,0,1,0,0,0)
PG_2	(0,0,0,0,0,0,0,0,0,1,0,0)
CM_1	(0,0,0,0,0,0,1,0,0,0,1,0)
CM_2	(0,0,0,0,0,0,0,1,0,0,0,1)
PMM	(1,0,0,0,1,1,0,0,0,0,0,0)
PMG_1	(1,0,0,0,1,0,0,0,0,1,0,0)
PMG_2	(1,0,0,0,0,1,0,0,1,0,0,0)
PGG	(1,0,0,0,0,0,0,0,1,1,0,0)
CMM	(1,0,0,0,0,0,1,1,0,0,1,1)
$P4$	(1,0,1,0,0,0,0,0,0,0,0,0)
$P4M$	(1,0,1,0,1,1,1,1,0,0,1,0)
$P4G$	(1,0,1,0,0,0,1,1,1,1,1,0)
$P3$	(0,1,0,0,0,0,0,0,0,0,0,0)
$P31M_1$	(0,1,0,0,1,1,1,0,1,1,1,0)
$P31M_2$	(0,1,0,0,1,1,0,1,1,1,0,1)
$P3M1_1$	(0,1,0,0,0,0,1,0,0,0,1,0)
$P3M1_2$	(0,1,0,0,0,0,0,1,0,0,0,0)
$P6$	(1,1,0,1,0,0,0,0,0,0,0,0)
$P6M$	(1,1,0,1,1,1,1,1,1,1,1,0)

2.2 Adaptive NNC (ANNC)

Recent works on NN classifiers have shown that adaptive schemes [9] outperform the results of classic NNC in many applications. In response to the ambiguities in computed symmetry values, we propose an adaptive approach based on establishing a merit function to adapt the inter-class boundaries to the specific image conditions. Fig. 3-a shows a simplified example of a 2D feature space including 4 binary prototypes. The inter-class boundaries are symmetric with respect to each prototype. In a real context, the $SFV(f_1, f_2)$ vectors never reach certain areas close to the prototypes, Fig. 3-b shows these forbidden areas. The distorted distances force an adaptation of the boundaries between classes.

To achieve this, a transformation of the the feature space can be performed by normalizing these features. In this new space, the null-class $P1$ disappears, therefore this class should be treated separately. The new boundaries between classes can be searched in a way that maximizes a merit function. We use orthogonal boundaries defined by a single parameter U, the *uncertainty boundary of symmetry*. We studied several merit functions and, finally, propose a distance ratio between the reported first and second classes after classifying the sample with respect to binary prototypes using a NN classifier. The result is the boundary

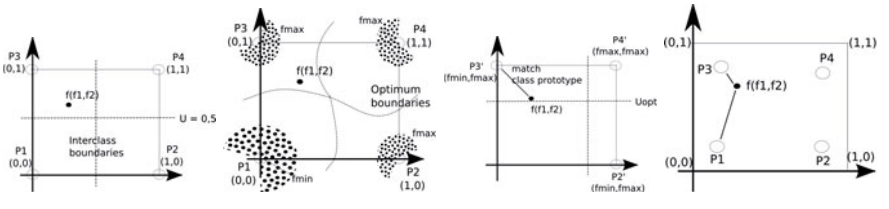


Fig. 3. From left to right: a) 2D feature space and prototypes $P1$, $P2$, $P3$ and $P4$. b) Forbidden areas. c) Adaptation of class boundaries. d) Final disambiguation.

U_{opt} that best separates the classes. Moreover, instead of moving the inter-class boundaries, the problem is reformulated to modify the class prototypes into new values $(H, L) \in [0, 1]$ that are symmetrical with respect to the middle value U (Fig. 3-c). Finally, the closest class to new prototypes (H, L) and the null-class $P1$ are disambiguated (Fig. 3-d). The algorithm is as follows:

Step 1 - The symmetry values are normalized, see Eq. 3, discharging the $P1$ class and resulting in a 16-class problem.

$$\begin{gathered} SFV' = (f'_1, f'_2, ..., f'_{12}); \\ f'_i = \frac{f_i - min(SFV)}{max(SFV) - min(SFV)}; \quad 1 \leq i \leq 12 \end{gathered} \tag{3}$$

Step 2 - The original prototypes are transformed into (H, L) prototypes for each of 16 classes. These values are defined with respect to parameter U as: $H = 1,\ L = 2 \cdot U - 1$ for $U \geq 0,5$ and $L = 0,\ H = 2 \cdot U$ otherwise.

Step 3 - For each U, ranging from 0.2 to 0.8, the H and L limits are computed and a NNC is performed using SFV' and the resulting prototypes. Repeating steps 2-3 for all U values, the value (U_{opt}) that maximizes the merit function is selected, and the corresponding class is also tentatively selected.

Step 4 - Finally, we disambiguate the candidate class from the previously excluded $P1$ class. To achieve this, we again re-classify the SFV but only using the $P1$ and candidate classes.

3 Experiments

As indicated in [6], without a systematic evaluation of different symmetry detection and classification algorithms against a common image set under a uniform standard, our understanding of the power and limitations of the proposed algorithms remains partial. As image datasets reported in literature were not publicly available, we selected several wallpaper images from known websites, to carry out the comparison between the proposed ANNC and the reference RBC methods. We picked image datasets from Wallpaper [4], Wikipedia [10], Quadibloc [8], and SPSU [2], resulting in a test bed of 218 images. All images were hand-labelled to make the ground truth. As the original RBC algorithm source code was unavailable, we implemented it using the original RBC decision tree reported in [5], but using our SFV feature vectors, and binarising the features using a fixed threshold (the average of the better classification results) for all image datasets. The results obtained with RBC, NNC and ANNC classifiers are shown in Table 3. For the shake of brevity, we only include here the percentage of successful classification, i.e. accuracy or precision results.

The first image collection is Wallpaper, a standard collection reported in previous works. In this case, both RBC and ANNC methods obtain a 100% of success. The RBC achieves the same result as reported in [5], which means that our implementation of this algorithm has similar results as the original implementation. To take into account the varying complexity of the images, we separate the next two collections in sub-sets. In the WikiGeom dataset, which is a subset of Wikipedia formed by strongly geometrical patterns, the ANNC and NNC

Table 3. Experimental classification results from RBC, NNC, and ANNC

Collection	#img	Sub-set	RBC	NNC	ANNC	NNC2	ANNC2	NNC3	ANNC3
Wallpaper	17		100	82.35	100	100	100	100	100
Wikipedia	53		54.72	60.38	62.26	73.55	81.13	81.13	83.02
	17	WikiGeom	88.24	88.24	100	94.12	100	100	100
Quadibloc	46		71.74	69.57	82.61	82.61	91.30	91.3	95.63
	29	Quad0102	62.07	75.86	79.31	86.21	89.66	89.66	93.10
	17	Quad03	88.24	58.82	88.24	76.47	94.12	94.12	100
SPSU	102		49.02	62.76	59.80	71.57	76.47	76.47	82.35
Global	218		59.18	65.15	68.35	76.60	82.60	82.60	86.70

results outperformed the RBC. In the case of the entire Wikipedia collection, which includes other distorted images, a decrease in results is evident. Similar results were obtained with Quadibloc image collection, which is of intermediate complexity. We studied it as two subsets: one formed by sketches over a uniform background (Quad0102), and other (Quad03) constituted by more complex motives with many highly contrasted colours. The ANNC obtains a near 80% success rate with this collection, clearly outperforming the NNC and RBC methods. The worse results were obtained with the more complex images in the SPSU collection. In this case, all results are below 60%. This lower values are due to the existence of noise and imprecise details (hand-made) in the images. Also, these exhibit several repetitions and illumination artifacts, which suggest the need for pre-processing. It is remarkable that the ANNC algorithm is still 10 points higher than the RBC algorithm. The latest row is a global hit success considering the number of images in each collection.

The fact of working with a distance-based classifier offers an additional advantage because it delivers a value defining the degree of proximity to the prototype chosen ($d_i = dist(SFV, P_i)$). This (P_i, d_i) description, which can be extended to the whole set of prototypes, can be used as a high level semantic image descriptor, useful in areas such as CBIR. This is particularly helpful in the presence of complex images that, due to various factors (manufacturing defects, noise, small details, etc.), present an ambiguity about the symmetry group to which they belong. Some of these images are difficult for experts to label. Thus taking, the first two (NNC2 & ANNC2) or three (NNC3 & ANNC3) classification results, the success rates are considerably higher (see Table 3) and the distance to the second and third candidate are near the first result. That shows that many of the classification errors were due to the above-mentioned factors. This idea can be conveniently exploited in the context of CBIR.

4 Conclusions

This paper had presented a proposal for a novel computational framework for classification of repetitive 2D pattern images into symmetry groups. The feature vector is composed of twelve symmetry scores, computationally obtained from

image gray level values. The absence of symmetry is never computed as '0', nor the presence of symmetry is computed as '1', even assuming perfect image conditions. The RBC and the NNC behave poorly, because of the ambiguities in symmetry computation. This leads to the use of some adaptive approach, implemented by an adaptive classifier. The ANNC is non-parametric, and it is also non-supervised. The experimental results show that the ANNC outperforms the other methods, even with very complex image collections. Moreover, the results can be useful in recovery tasks (CBIR systems) using an extended version of ANNC - which produces a list of similarity for every group that can be sorted from highest to lower values, and so for example, detect images that are near to several groups. As future work, we are looking for a new way of computing the symmetry features, extending the test beds, and the method for colouring images.

References

1. Agustí, M., Valiente, J.M., Rodas, A.: Lattice extraction based on symmetry analysis. In: Proc.of 3rd. Int. Conf. on Computer Vision Applications (VISAPP 2008), vol. (1), pp. 396–402 (2008)
2. Edwards, S.: Tiling plane & fancy (2009), `http://www2.spsu.edu/math/tile/index.htm`
3. Horne, C.: Geometric Symmetry in Patterns and Tilings. Woodhead Publishing, Abington Hall (2000)
4. Joyce, D.: Wallpaper groups (plane symmetry groups) (2007), `http://www.clarku.edu/~djoyce/`
5. Liu, Y., Collins, R., Tsin, Y.: A computational model for periodic pattern perception based on frieze and wallpaper groups. Trans. on PAMI 26(3) (2004)
6. Liu, Y., Hel-Or, H., Kaplan, C.S., Gool, L.V.: Computational symmetry in computer vision and computer graphics. In: Foundations and Trends in Computer Graphics and Vision, vol. (5), pp. 1–195 (2010)
7. Reddy, S., Liu, Y.: On improving the performance of the wallpaper symmetry group classification. CMU-RI-TR-05-49, Robotics Institute, CMU (2005)
8. Savard, J.G.: Basic tilings: The 17 wallpaper groups, `http://www.quadibloc.com/math/tilint.htm`
9. Wang, J., Neskovic, P., Cooper, L.N.: Improving nearest neighbor rule with a simple adaptive distance measure. Pattern Recognition Letters 28(2), 207–213 (2007)
10. Wikipedia: Wallpaper group, `http://www.wikipedia.org`

Distance Maps from Unthresholded Magnitudes

Luis Anton-Canalis[1], Mario Hernandez-Tejera[1], and Elena Sanchez-Nielsen[2]

[1] SIANI - University of Las Palmas de Gran Canaria
[2] University of La Laguna,
{lanton,mhernandez}@siani.es, enielsen@ull.es

Abstract. A straightforward algorithm that computes distance maps from unthresholded magnitude values is presented, suitable for still images and video sequences. While results on binary images are similar to classic Euclidean Distance Transforms, the proposed approach does not require a binarization step. Thus, no thresholds are needed and no information is lost in intermediate classification stages. Experiments include the evaluation of segmented images using the watershed algorithm and the measurement of pixel value stability in video sequences.

Keywords: Distance Transform, Thresholding, pseudodistances.

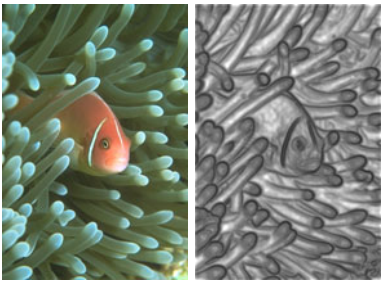
1 Introduction

The Distance Transform (DT), originally proposed by Rosenfeld and Pfaltz [10], has been widely used in computer vision, image processing, pattern recognition and robotics. Applied to an image, a DT assigns a value to each image point that represents the minimum distance to some locus of points, usually belonging to one of the two groups created by a binary partition. Classic DT (i.e. Euclidean Distance Transform) has been applied to object recognition [3] [6], path planning [12], active contour modeling [15] and segmentation [2] [9].

DT usually operates on binary feature maps obtained by thresholding operations like thresholding or edge detection. However, any binarization process based on thresholds sistematically creates groups of points in which binary membership may fluctuate through time due to small light changes or image noise. As video processing applications require features to be detected consistently in the spatiotemporal domain, processes based on thresholds should be avoided. In this context, computing distance maps directly from unthresholded magnitude values should increase the stability of further processing stages relying on them.

Different solutions that do not require a binarization step have been proposed to compute pseudo-distances from a given image point to some locus of points. In [7], pseudo-distance maps are computed applying PDEs to an edge-strength function of a grayscale image, obtaining a robust and less noisy skeletonization; in [11], pseudo-distances are weighted by edge magnitude and length; in [1], an intuitive solution named Continuous Distance Transform (CDT) is proposed, where pseudo-distances to brightness and darkness saturation regions are computed directly from grayscale images.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 92–99, 2011.



(a) (b)

Fig. 1. DMUM pseudo-distance values 1(b) directly computed from the unthresholded Sobel gradients of image 1(a)

In this paper, a novel method for computing distance maps directly from unthresholded magnitude values is proposed. No critical information is lost in an intermediate classification step and different magnitude values can be employed (i.e. depth or gradient, see Fig.1). Output values represent the size of the smaller area around each pixel enclosing a relevant amount of accumulated magnitude values that depends solely on the image being studied. A single pass on the magnitude values image is needed, and the integral image [14] is used to speed up area sums. A formal definition and a concise algorithm to compute this Distance Map from Unthresholded Magnitude values (DMUM) are proposed.

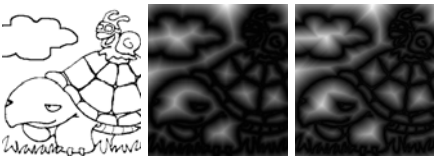
Section 2 describes DMUM, proposing a straightforward algorithm. Section 3 analyzes DMUM values in watershed segmentations and its stability in video sequences and finally Section 4 summarizes the approach and proposes future research lines and applications.

2 Distance Maps from Untresholded Gradients

In classic DT each point receives a value that represents a *distance to the closest object point*. However, a suitable definition of *object* is necessary. It is usually achieved through plain thresholding or edge detection, which leads to a classification step where points are considered object or non-object. This is not a trivial problem: an unsuitable classification mechanism may produce an inaccurate binarization, both removing information and adding noise. In DMUM, the *distance to the closest object point* concept is replaced by the *distance to the closest most relevant considered magnitude value region* in a mapping function that relies directly on unthresholded magnitude values. For the sake of simplicity, the rest of this work will work with gradient magnitudes, but many other features can be used instead, like depth maps.

Let us define the set of points bounded by a closed ball of radius r centered at point p:

$$B_r[p] \triangleq \{x \in \mathbb{S}^2 | d(x,p) \leq r\} \tag{1}$$



(a) (b) (c)

Fig. 2. 2(a) Original binary image, 2(b) classic chessboard DT values and 2(c) DMUM values

where $d(x, p)$ is a distance function. If $\mathbb{S}^2 \equiv \mathbb{R}^2$ then $B_r[p]$ is defined in a continuous metric space, while $\mathbb{S}^2 \equiv \mathbb{Z}^2$ defines $B_r[p]$ in a discrete metric space. For optimization purposes the Chebyshev distance ($L\infty$ norm) is adopted, defining a square around p.

Given an image $i \in \mathbb{Z}^2$ and its normalized gradient image g^n, being its maximum value 1.0, consider the sum of gradients for a given $B_r[p]$ in i as:

$$\phi_r(p) = \gamma \cdot \sum_{\forall x \in B_r[p]} g^n(x) \tag{2}$$

being the γ parameter a scaling factor that will be explained later. Radius q is defined as:

$$q(p) = \arg_r \, sup\{\phi_r(p) > 1.0\} \tag{3}$$

that is, the minimum radius r of a ball $B_r(p)$ that encloses a sum of gradient $\phi_r(p)$ that exceeds 1.0. Then, the value of the Distance Map from Unthresholded Magnitudes at each point p is given by:

$$\Gamma(p) = q - \sum_{r=1}^{q} \phi_r(p) \tag{4}$$

DMUM values $\Gamma(p)$ will be higher for points placed in sparse gradient regions and lower for points with a high gradient or points placed in dense gradient regions. Its application to a binary image produces an output similar to that produced by DT, as seen in Fig.2. Measuring pixel value differences between Fig. 2(b) and Fig. 2(c) returns a dissimilarity smaller than 1.5%.

The accumulation of $\phi_r(p)$ values in Eq.4 introduces a linear weighting factor. The first and smallest area contributes q times to this accumulation, because it is also contained on bigger areas. The second one contributes $q-1$ times, the third one $q-2$ times and so on, while the area that satisfies the condition contributes only once. While raw q values could be used as the desired output, resulting maps would feature isohypses of heights instead of more continuous values. This linear weighting modifies DMUM values according to local magnitude conditions. Fig. 3 shows this effect. The $\gamma, \gamma \, \epsilon(0, 1.0]$ parameter in Eq.2 allows DMUM output values to be softened while computing output values, so no previous filtering is needed. Because region areas grow exponentially (1x1, 2x2, 3x3...qxq), the

(a) (b) (c)

Fig. 3. 3(a) Original image. 3(b) Raw DMUM q values. 3(c) The accumulation of innermost regions results in smoother DMUM maps.

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

Fig. 4. 4(a) Original image. DMUM computed with $\gamma = 1.0$ 4(b), 0.1 4(c) and 0.01 4(d).

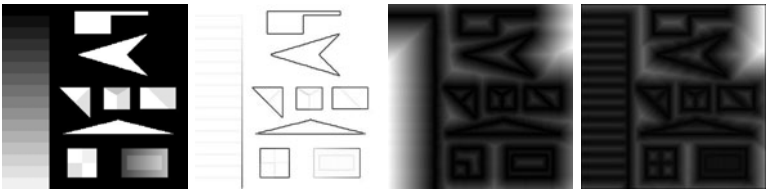
γ parameter dampening effect will be much stronger in small areas with low gradient accumulations. Fig.4 shows the effect of different γ values on the same image. Notice how high gradient values are preserved, effectively working as an anisotropic filter.

DMUM values represent the pseudo-distance from a given image point to the closest most relevant gradient accumulation. Due to the limit condition in Eq. 3, $\phi_r(p) <= 1.0, \phi_r(p) \in [0, q]$ and $\Gamma(p) >= 0$, being zero only for those points which gradient equals 1.0. This satisfies the positive definiteness condition of a metric. However no symmetry and thus no triangle inequality can be defined. Therefore, DMUM map values can only be considered a pseudo-distance, though most works consider both terms interchangeable.

2.1 DMUM Algorithm

Computing the DMUM value for each pixel p in i consists of finding the smaller area $B_q(p)$ which gradient sum $\phi_q(p)$ is equal or higher than 1.0. As only squared regions around p are considered, due to the adoption of the Chebyshev distance, the sum of gradient values can be costlessly computed using the integral image [14] of g. The following pseudocode summarizes the whole process:

Require: RGB image $source$ sized n x m
Require: $\gamma \in (0..1]$
Ensure: Floating point image $DMUM$ sized n x m
  $DMUM[i, j] \leftarrow 0, i \in [0..n), j \in [0..m)$
  $gradient \leftarrow gradient(source)/max(gradient(source))$
  $integral \leftarrow integral(gradient)$
  $max\_r \leftarrow min(n/2, m/2)$
  for $i \in [0..n)$ **do**
    for $j \in [0..m)$ **do**



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

Fig. 5. 5(a) Image with both binary and grayscale regions. 5(b) Sobel gradient. 5(c) DT values. 5(d) DMUM map.

```
        r ← 1
        enclosed_gradient ← gradient[i, j] · γ
        gradient_accum ← enclosed_gradient
        while enclosed_gradient ≤ 1.0 and r ≤ max_r do
            gradient_accum ← gradient_accum + enclosed_gradient
            r ← r + 1
            enclosed_gradient ← (integral(i + r, j + r) + integral(i − r, j − r)) − (integral(i + r, j −
            r) + integral(i − r, j + r))
            enclosed_gradient ← enclosed_gradient · γ
        end while
        DMUM[i, j] ← r − gradient_accum
    end for
end for
return DMUM
```

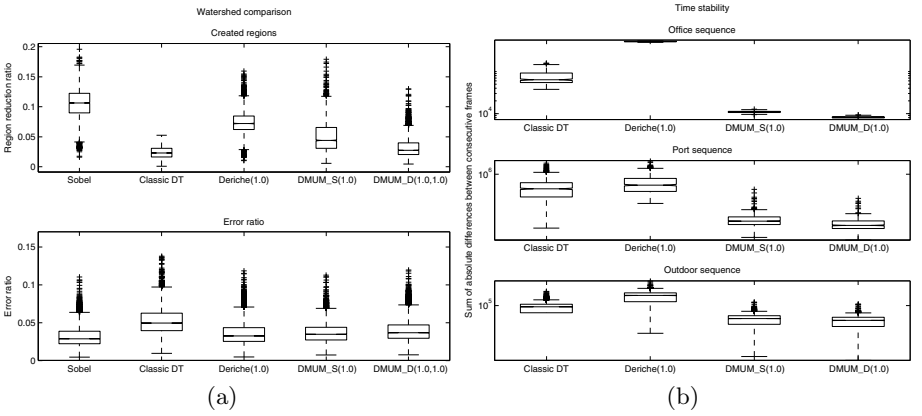
3 Results

The main difference between classic DT (computed from a Canny edge map) and the proposed approach is shown in Fig.5. Although both operations behave similarly in binary regions, details on grayscale parts are removed in the DT image. Ideally, the perfect set of Canny parameters would produce an optimal distance map, but finding them is not easy. Each image, even consecutive frames of a video sequence, may need different values. Besides, unevenly lighted images may require parameters to be adjusted locally. However, DMUM is computed from local gradient values (see Eq.3) so the original image structure is better preserved (See Fig. 5).

It was already shown in [1] that distance maps applied directly to grayscale values increase object recognition rates using Chamfer distances. The solution proposed in [**?**] was specifically designed for OCR applications and computationally expensive. DMUM offers a more general solution that is two orders of magnitude faster and introduces an anisotropic filtering mechanism. Experiments were focused on the analysis of the structure of distance maps and their stability in video sequences.

3.1 Spatial Structure Analysis

Watershed algorithms[13] [2] perform an oversegmentation of a given image, creating groups of pixels that share similar features and reducing the amount of



(a) (b)

Fig. 6. Watershed region-error ratios and time stability measures

data needed to describe the image. The sorting and region growing mechanism in the basic watershed algorithm reveals relevant morphological relations between pixels.

A total of 5074 images were watershed-segmented, including the Berkeley Segmentation Dataset [8], showing a wide range of indoor and outdoor images and different sizes and noise factors. Sobel gradients, chessboard DT images, Deriche [5] gradients and DMUM values computed both from Sobel ($DMUM_S$) and Deriche ($DMUM_D$) were used as input for the watershed algorithm. Both the number of final regions and the sum of absolute differences between pixels from the segmented image and the original image were measured. These values were normalized with respect to the maximum number of regions (that equals the number of pixels) and the maximum possible error (sum of maximum error for every pixel on each channel) respectively, obtaining the region reduction ratio and the error ratio respectively.

Region reduction and error ratios for each method are depicted in Fig.6. Classic DT creates the lowest number of regions, but also the highest error, which reveals the loss of information suffered in the binarization step required to compute the classic DT. As DMUM computes pixel values from local regions, differences between neighbouring pixels tend to be smoother, while noisy pixels have a smaller influence on the final outcome.

While $DMUM_S$ creates less regions that *Deriche*, the application of a Tukey-Kramer multiple comparison test on error ratio values reveals that Deriche and $DMUM_S$ error ratios are significantly similar, with a confidence level of 99%. $DMUM_D$ creates even less regions with a slightly higher error, but introduces the computational complexity of computing Deriche gradients.

3.2 Temporal Stability Results

It is important for most operators working on video sequences to be stable in time, meaning that the same values should be returned if operational conditions

have not changed significantly. Even small light changes like those coming from the almost imperceptible flicker of a fluorescent bulb may affect the scene, and they certainly affect some image operators like the Canny edge detector.

Three different video sequences were used for measuring stability. The first one is an indoor sequence with no visible changes. As nothing moves, the difference between consecutive frames would be ideally zero, although visually unnoticeable light changes introduce pixel value variations. The second sequence shows a video from a static security camera placed above a harbor, showing images from dawn until dusk. Light conditions change smoothly and some objects cross the scene. Finally, the third sequence is an outdoor video that includes dynamic objects and backgrounds, different settings and changes in lighting conditions.

Classic DT, Deriche gradients, $DMUM_S$ and $DMUM_D$ were computed on each frame. Temporal coherence was measured computing the sum of absolute differences between pixel values of consecutive frames for each method, avoiding prior filtering. Fig. 6 depicts measured differences. DMUM values are clearly more stable in video sequences than Deriche and classic DT on the three sequences, showing also a smaller variance.

A new Tukey-Kramer multiple comparison test applied to the time stability test results reveals that $DMUM_S$ and $DMUM_D$ values are significantly similar to each other, and different from DT and Deriche. Once again, $DMUM_S$ seems more appropriate for real time applications due to the computational cost of computing Deriche gradients.

4 Conclusions

This paper describes a new method to compute distance maps from unthresholded magnitudes that includes an inexpensive anisotropic filter. It is suitable for still images and real time applications. The approach is conceptually simple and can be easily reproduced following the proposed algorithm. Similarly to classic DT, DMUM computes a pseudo-distance from any pixel to the closest most relevant gradient accumulation.

Two different experiments were perfomed. A first test compared watershed segmentations created from five different methods: Sobel gradients, DT, Deriche gradients and both DMUM computed from Sobel and Deriche values. The number of created regions and the per-pixel error between segmented images and their original values was measured. It was statistically proved that the proposed approach obtains a better region-to-error ratio than the rest of considered methods, suggesting that pixel value relations are more natural in DMUM images and supporting the goodness of unthresholded methods. The proposed operator is also more stable in video sequences, obtaining the lowest pixel value differences between consecutive frames. This stability is critical in object detection or tracking schemes. It was also shown that $DMUM_S$ was statistically as stable as $DMUM_D$. Being Sobel gradients much simpler to compute, $DMUM_S$ results appropriate for real-time applications. Further image processing stages would certainly benefit from DMUM increased stability, as there exists a stronger certainty that changes in values correspond to true changes in the scene.

Different optimizations are being considered in order to improve overall speed, considering that DTs usually take place in the first stages of a visual system. Further research related to DMUM includes the application of values to Hausdorff matching for object classification and tracking, and its application to depth maps to guide object segmentation.

References

1. Arlandis, J., Perez-Cortes, J.-C.: Fast handwritten recognition using continuous distance transformation. In: Sanfeliu, A., Ruiz-Shulcloper, J. (eds.) CIARP 2003. LNCS, vol. 2905, pp. 400–407. Springer, Heidelberg (2003)
2. Beucher, S., Meyer, F.: The morphological approach of segmentation: the watershed transformation. Mathematical Morphology in Image Processing, 433–481 (1993)
3. Borgefors, G.: Hierarchical chamfer matching: A parametric edge matching algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(6), 849–865 (1988)
4. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698 (1986)
5. Deriche, R.: Using canny's criteria to derive a recursively implemented optimal edge detector. International Journal of Computer Vision 1(2), 167–187 (1987)
6. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9), 850–863 (1993)
7. Jang, J.H., Hong, K.S.: A pseudo-distance map for the segmentation-free skeletonization of gray-scale images. In: Proceedings of ICCV, pp. II18–II23 (2001)
8. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of ICCV, vol. 2, pp. 416–423 (July 2001)
9. Roerdink, J.B.T.M., Meijster, A.: The watershed transform: definitions, algorithms and parallelization strategies. Fundamenta Informaticae 41(1-2), 187–228 (2000)
10. Rosenfeld, A., Pfaltz, J.L.: Distance functions on digital pictures. Pattern Recognition 1(1), 33–61 (1968)
11. Rosin, P.L., West, G.A.W.: Salience distance transforms. Graphical Models and Image Processing 57(6), 483–521 (1995)
12. Taylor, T., Geva, S., Boles, W.W.: Directed exploration using a modified distance transform. In: Proceedings of DICTA, p. 53. IEEE Computer Society, Washington, DC (2005)
13. Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. 13(6), 583–598 (1991)
14. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of CVPR, pp. I511–I518 (2001)
15. Yang, R., Mirmehdi, M., Xie, X.: A charged active contour based on electrostatics. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2006. LNCS, vol. 4179, pp. 173–184. Springer, Heidelberg (2006)

Scratch Assay Analysis with Topology-Preserving Level Sets and Texture Measures

Markus Glaß[1], Birgit Möller[2], Anne Zirkel[1], Kristin Wächter[1], Stefan Hüttelmaier[1], and Stefan Posch[2]

[1] Zentrum für Angewandte Medizinische und Humanbiologische Forschung (ZAMED), Martin Luther University Halle-Wittenberg, Heinrich-Damerow-Str. 1, 06120 Halle
[2] Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06099 Halle/Saale, Germany
markus.glass@medizin.uni-halle.de

Abstract. Scratch assays are widely used for cell motility and migration assessment in biomedical research. However, quantitative data is very often extracted manually. Here, we present a fully automated analysis pipeline for detecting scratch boundaries and measuring areas in scratch assay images based on level set techniques. In particular, non-PDE level sets are extended for topology preservation and applied to entropy data of scratch assay microscope images. Compared to other algorithms our approach, implemented in Java as ImageJ plugin based on the extension package MiToBo, relies on a minimal set of configuration parameters. Experimental evaluations show the high-quality of extracted assay data and their suitability for biomedical investigations.

Keywords: Scratch assay, level set segmentation, texture, topology.

1 Introduction

Cell migration is an essential mechanism for various processes of multicellular organisms. Examples of such processes include wound healing, immune responses, and tissue formation. Migration is characterized by the synchronized movement of cells in particular directions, often caused by and aimed at external signals. Dysfunction of this mechanism may have serious impact on an organism causing, e.g., mental retardation or vascular diseases. Also for the course of cancer, cell migration plays a central role as it is the mechanism underlying cell scattering, tissue invasion, and metastasis.

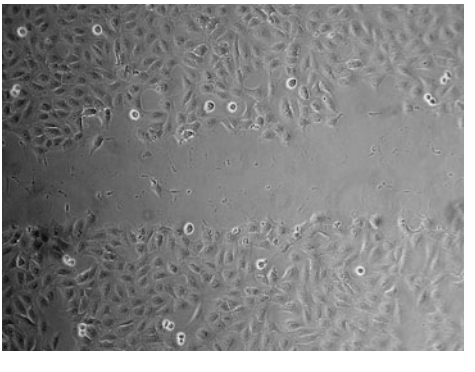


Fig. 1. Example scratch assay image with a horizontal scratch in the middle of the image. For display, the contrast of the image was enhanced.

One experimentally well-developed and easy protocol for analyzing cell migration in vitro is the *scratch assay* [10]. Here, a mono-layer of cells is grown in vitro and an

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 100–108, 2011.

artificial wound or *scratch* is mechanically created. The scratch's shape is typically a region with parallel borders (cf. Fig. 1). Subsequently, the cells will migrate to close this wound, and the time course of this migration process is monitored by capturing images, typically for a period of several hours to a few days.

There are several levels of interest in scratch assay analysis (see, e.g., [6]). In this paper we concentrate on the migration of cell populations and aim at measuring the dynamic invasion of the scratch area which gives insight into a cell population's overall migration capacity. In order to facilitate high-throughput experiments as well as objective evaluation we propose a fully automated image analysis pipeline for this problem. Compared to other techniques we focus on a minimal set of configuration parameters offering easy handling of the software also for non-experts. At the core we apply a level set approach to segment the images into foreground (scratch) and background (no scratch). The underlying energy functional is based on the common Chan-Vese energy, however, applied to entropy images derived from the scratch microscope images. For minimizing the energy, a non-PDE approach is used including topology-preservation as only one connected scratch area is desired as result. To the best of our knowledge, topology-preservation was not applied to non-PDE level set methods so far.

The quality of our approach is proved using data from a total of 10 scratch assay experiments with the cell line *U2OS* grown in two different experimental conditions. Precision and recall of the segmentation results as well as the sizes of the scratch areas are compared to ground-truth data independently labeled by two biologist experts.

The paper is organized as follows. After reviewing some related work in Sec. 2 we introduce our automated image analysis pipeline to detect scratches in microscopic images from mono-layer cells. An evaluation comparing the results to ground-truth data is presented in Sec. 4, and we conclude with some final remarks.

2 Related Work

Although scratch assays have gained increasing interest in biomedical research quantitative assessment of cell migration is in many cases still done semi-automatically or even completely manually. In [11] the quantitative assay analysis is supported by the image processing software package *ImageJ*[1], however, mainly depends on ad hoc threshold settings on fluorescent intensities. Other software packages available for scratch assay evaluation are *CellProfiler*[2] and *TScratch* [7]. While the first one is not specifically dedicated to scratch assays and mainly relies on image brightness for wound classification, *TScratch* considers also brightness-independent cues based on curvelets [2]. However, it requires a time-consuming adjustment of several configuration parameters for different applications as suggested defaults often do not lead to satisfactory results.

From a computer vision point of view, quantitative assessment of wound healing in scratch assays refers to a two-category classification task where regions populated by cells and the gap in between have to be distinguished from each other. As image intensities are usually quite similar in both regions (cf. Fig. 1) the classification requires brightness-independent features like texture measures. To characterize texture a huge

[1] `http://rsbweb.nih.gov/ij/`
[2] `http://www.cellprofiler.org`

amount of different measures has emerged over the years starting from image entropy and co-occurence matrices, subsuming various filter techniques like Law and Gabor features [15], and ending up with local measures like local binary patterns [12].

Independent of the chosen texture representation, however, the distinction between cells and background requires the application of appropriate classifiers. In the simplest case binary threshold classification is done on extracted features [11]. More sophisticated classification techniques consider statistical models like generative Gaussian mixtures [14], or discriminant criteria for SVMs [4].

Interpreting the classification task as a segmentation problem also active contour techniques like level sets are applicable [13,5]. Level sets implicitly define image contours between target regions or classes as the zero level of an embedding function on the image. Starting from an initialization they evolve according to an energy functional into a local minimum yielding the final segmentation. The energy rates the segmentation in an application dependent way and may include measures of various image characteristics, like average region intensities and variances, gradient magnitudes, or texture measures. These may be, e.g., derived from structure tensors [1] or subsumed in feature histograms [9]. Level set methods can be configured to either adapt to an a-priori unknown region topology or to keep the initially specified topology [8].

3 Entropy-Based Level Set Segmentation

Level set principles. Level set methods represent contours implicitly by an embedding function $\phi(x,y)$ defined on the image domain. The contours are defined as the zero level of this level set function, while the foreground is represented by positive and the background by negative values. In the following, the scratch region is considered as the foreground, whereas the cell areas are considered as background. Starting from a given initialization, these contours evolve in time according to an energy functional to be minimized. This energy functional determines the characteristics of the segmentation as given by the application under consideration. For the analysis of scratch assays we choose the widely used Chan-Vese energy [3].

This Chan-Vese model assumes the image to be composed of two not necessarily connected regions. These foreground and background regions are assumed to have different mean gray values being approximately constant within the regions. Deviations from these means are penalized by the energy functional where in general the deviations may be weighted differently for both regions. For our application the weights are assumed to be identical and set to 1. In addition, the Chan-Vese model contains two regularization terms regarding perimeter and area of the foreground region, which are however not required for our application. Accordingly, the energy to be minimized is

$$\begin{aligned} E(\phi(x,y),c_1,c_2) = &\int_\Omega |I(x,y)-c_1|^2 H(\phi(x,y))dxdy \\ &+ \int_\Omega |I(x,y)-c_2|^2 (1-H(\phi(x,y)))dxdy \end{aligned} \tag{1}$$

where Ω denotes the image domain, c_1 and c_2 are the average values of the foreground and background, respectively, and $H(.)$ denotes the Heaviside function. For minimization we do not solve the Euler-Lagrange equation associated with the energy functional.

Following the work of [16] we directly evaluate the discretized energy, and iteratively check, if changing the sign of the level set function of a pixel decreases the energy. If so, the pixel is flipped to the respective region. Iterations are performed until the pixel assignment to both regions converges.

Preserving topology. Usually, level set methods inherently allow for changes in the topology, e.g. splitting or merging of connected components. However, for our application this is not desirable as holes in the cell layer, which are not connected to the scratch, should not be recognized as wound. In addition, the scratch area is to contain no holes as cells migrate from the border of the initial scratch during healing. Thus, we devise a topology-preserving minimization strategy for the non-PDE approach. We initialize the scratch region as one connected component without holes using 4-connectedness for the scratch region and 8-connectedness for the background. To preserve this topology during minimization we use a strategy which is related to, but slightly simpler than the method proposed in [8] for PDE-based minimization. A pixel is allowed to change the sign of its level set function if it is a contour pixel and its 8-neighborhood contains only one connected component of foreground pixels 4-connected to the pixel under consideration. The latter property is identical to the first of two conditions used in [8] to verify a pixel to be simple, namely the topological numbers $T_4(x, fg)$ to equal one.

Entropy. Average intensity values within cell and scratch area are approximately identical and far from being constant. Thus the energy (1) cannot be applied to the intensity values directly. However, the distribution of gray values within the scratch area is far more homogeneous than inside the cell regions. One property that can reflect this difference is the Shannon entropy,

$$-\sum_{l=0}^{L-1} P(l) \cdot log_2\, P(l), \tag{2}$$

where $P(l)$ denotes the probability of the intensity values l, $l = 0, ..., L-1$. The entropy is estimated for each pixel using a window centered at this pixel to approximate intensity probabilities as their relative frequencies. The resulting entropy values are used as intensities in (1) to discriminate between scratch and cell areas.

Application. The outline of our algorithm to detect scratch areas is as follows. After smoothing the input image using a Gaussian filter with $\sigma = 5$ to reduce noise, for every pixel of the image the discretized entropy is calculated using a window size of 31×31. This size was determined using an additional set of seven images, also manually labeled by two biologist experts. The F-measure for the resulting precision and recall were calculated and the window size chosen that maximized the F-measure's mean. For details on image data, ground truth labeling, and performance measures see Sec. 4.

By design, the level set segmentation always yields exactly one scratch area, even if none is present any more, Thus we scrutinize the histogram of entropy values to detect these cases. If the lower 20% of the entropy interval found in the image is populated with less than 2.5% of all pixels it is assumed that no scratch is present. In detail, let $[e_{\min}; e_{\max}]$ be the entropy interval of an image, than we assume no scratch present if the 0.025 percentile is larger than $e_{\min} + 0.2(e_{\max} - e_{\min})$. Otherwise, the segmentation is performed applying the topology-preserving level set method on the entropy image.

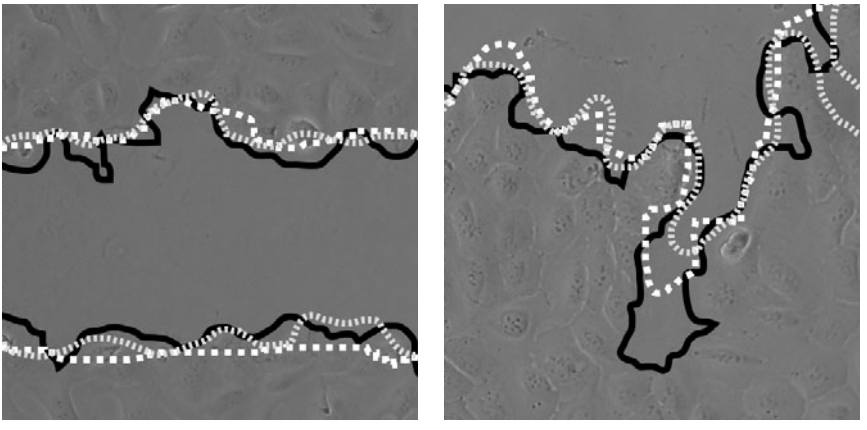


Fig. 2. Two sample segmentation results for clips of scratch assay images. Both figures compare expert labelings (dotted white and gray) and our automated segmentation results (solid black).

Assuming the scratch to be centered in the image the scratch region is initialized as a rectangle in the middle of the image spanning its complete width. After segmentation the result is post-processed to remove small intrusions of cell regions inside the scratch region caused by cell debris. Such intrusions are linked to the surrounding cell region by only very thin corridors due to topology preservation. They are removed by morphological closing with a squared structuring element sized 15×15 pixels. The closing may give raise to holes in the scratch area, which are finally removed by hole filling.

The algorithm was implemented as a plugin for the open-source package *MiToBo*[3], an extension of the Java image processing software *ImageJ*.

4 Results

Image data. Cultivated osteosarcoma derived *U2OS* cells were used to evaluate the algorithm. The cells where divided into two populations. Cells from the first population where starved, whereas the others where fed with fetal calf serum. Consequently, for the latter population it is assumed that scratches should be closed faster.

Five assays per population were prepared and monitored at seven distinct time points (0h, 2h, 4h, 6h, 8h, 12h, 24h) after inducing the scratches, resulting in 70 images each sized 2048×1536 pixels. In addition, one separate assay was used to optimize configuration parameters in advance, i.e. the entropy mask's size (Sec. 3). Images where acquired by a digital camera linked to a bright-field microscope, both from Nikon. For ground-truth comparison each image was labeled manually by two biologist experts.

Evaluation metrics. The scratch segmentation task defines a two-category classification problem where each pixel in an image is labeled as scratch (foreground, FG) or non-scratch (background, BG). Common measures for the quality of a classification

[3] `http://www.informatik.uni-halle.de/mitobo/`

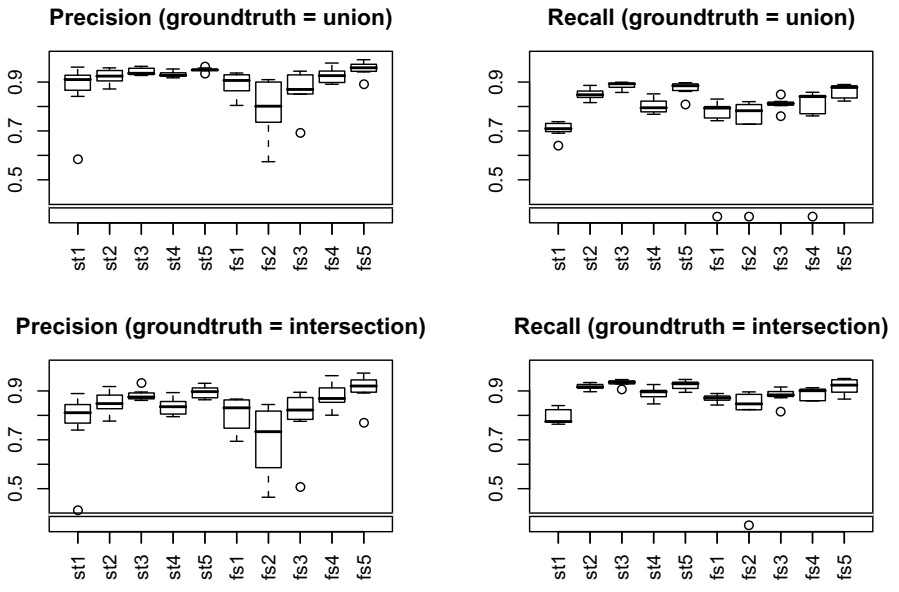


Fig. 3. Boxplots of recall and precision achieved for the two ground-truth definitions. On the abscissa the 10 experiments are arranged, i.e. the 5 starved (st1 to st5) and the 5 fetal calf serum experiments (fs1 to fs5). On the ordinate recall and precision averages over all seven time points for each experiment are shown. Outliers with a recall of 0 are shown in the bottom rectangles.

task are *recall* and *precision*. The recall defines the ratio of ground-truth FG pixels in an image correctly classified as FG. In contrast, the precision defines the ratio of pixels classified as FG that actually belong to FG. Since we have two expert labelings available for evaluation we use two different ways to define the segmentation ground-truth, namely the union of the FGs of both labelings and their intersection. We decided to use both ground-truth definitions in order to get a fair view of both experts' opinions. For the sake of completeness we present recall and precision for both of the ground-truth definitions.

Segmentation results. In Fig. 2 examples of segmentation results for our algorithm are shown for two clips from different assay images. Ground-truth segmentations are marked by dotted and automated segmentation results by solid black lines. In the left image the scratch is centered with boundaries running almost horizontally. While the expert labelings are relatively smooth in this case the automated segmentation shows larger sensitivity to local structures. In the right image a more curved scratch border is present with the scratch area at the top of the image. This gives rise to larger differences in segmentation results. Especially in the image center our algorithm includes larger areas without cells into the scratch area resulting in larger scratch areas (cf. quantitative results below). At the top right corner of the image an example for significant differences between all three segmentations can be seen. Here both expert labelings differ significantly while the automated result lies in between.

To quantitatively assess our results, recalls and precisions are presented in Fig. 3. For the two boxplots in the top row the ground-truth was derived from the union of both expert segmentations, for the plots in the bottom row their intersection was used. In the top row we observe median recalls between 0.71 and 0.89, and median precisions between 0.80 and 0.96. In comparison, in the bottom row medians of recall are in the range of 0.77 to 0.94, while the precision medians for all experiments range from 0.73 to 0.92. The latter recall values are slightly higher than in the union case which is to be expected as the ground-truth definition is less strict, i.e. includes pixels in the FG that are labeled by just one of the experts. Moreover, this is also the reason for the precision rates tending to be slightly smaller using ground-truth derived from intersection as using the intersection of expert labelings reduces the number of FG pixels and, thus, leads to an increased chance for the algorithm to wrongly include BG pixels into the FG.

Our precondition (see Sec. 3) detected the absence of a scratch in four images. For two of these one expert labeled some FG pixels, but these labelings comprised only 0.2 % and 0.4 % of the image areas, respectively. In another image both experts labeled about 5 % of the image area as FG, whereas the precondition stated no scratch. These cases explain the recall values of 0 and suggest slight improvements of our precondition. The other outliers mainly stem from image regions along the border of scratches where the recovery of the cell layer caused intrusions of scratch areas into the cell area. These are only partially included in the ground-truth by any of the experts, however, completely labeled by our algorithm due to their proximity to the scratch area (see Fig. 2, right, for an example).

From a biological point of view the scratch area's size dynamics over time is of interest to quantify the cell population's motility. In Fig. 4 the average values of measured scratch areas are given for both populations and all seven time-points. In both plots ground-truth results are shown for comparison. A clear tendency for faster closing of the scratch in the population fed with fetal calf serum is observable for all three segmentations. More important, however, is the comparison between ground-truth and our automated segmentation as accurate data is essential for biomedical investigations. From both plots it is obvious that the experts' segmentations correlate quite well with each other as indicated by Pearson correlation

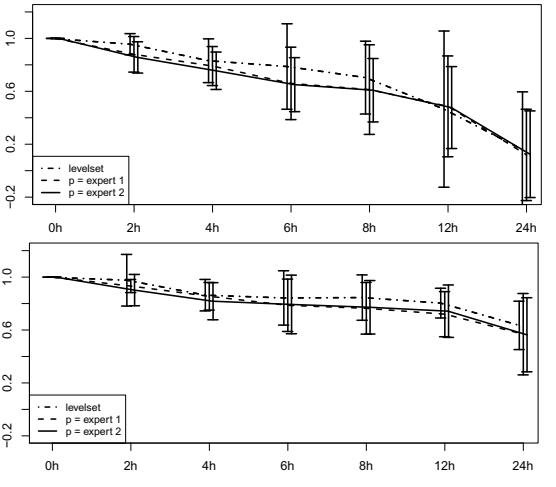


Fig. 4. Mean areas with errorbars of two-times the standard deviation over time for the fetal calf serum (top) and the starving populations (bottom). Ground-truth results are shown solid and dashed, our results dashed-dotted.

coefficients of 0.989 and 0.999, respectively. Our algorithm tends to segment slightly more scratch area than the experts, but recovers the overall tendency of scratch development very well. This is emphasized by the Pearson correlation comparing automated and expert segmentations with an average of 0.984. In addition, the difference between automatically and manually measured scratch area is small compared to the degree of variance within the populations for each time point.

5 Conclusion

The evaluation results show that our algorithm is well-suited for scratch assay analysis. Temporal evolution of scratch area is recovered with high quality allowing for biologically meaningful interpretation. In addition, our approach based on topology-preserving level sets combined with entropy-based texture measures relies on a minimal set of configuration parameters, particularly promoting the software to be used by non-experts.

References

1. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Colour, texture, and motion in level set based segmentation and tracking. Image Vision Comput. 28, 376–390 (2010)
2. Candès, E., Donoho, D.: Curvelets - a surprisingly effective nonadaptive representation for objects with edges. In: Cohen, A., Rabut, C., Schumaker, L. (eds.) Curves and Surface Fitting, pp. 105–120. Vanderbilt University Press, Nashville (2000)
3. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 10(2), 266–277 (2001)
4. Chethan, H.K., Raghavendra, R., Kumar, G.H.: Texture based approach for cloud classification using SVM. In: Proc. of Int. Conf. on Advances in Recent Technologies in Comm. and Comp., pp. 688–690 (2009)
5. Cremers, D., Rousson, M., Deriche, R.: A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. IJCV 72(2), 195–215 (2007)
6. Debeir, O., Adanja, I., Kiss, R., Decaestecker, C.: Models of cancer cell migration and cellular imaging and analysis. In: Lambrechts, A., Ampe, C. (eds.) The Motile Actin System in Health and Disease, pp. 123–156. Transworld Research Network (2008)
7. Gebäck, T., Schulz, M.M., Koumoutsakos, P., Detmar, M.: TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays. Biotechniques 46(4), 256–274 (2009)
8. Han, X., Xu, C.Y., Prince, J.L.: Topology preserving level set method for geometric deformable models. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(6), 755–768 (2003)
9. Karoui, I., Fablet, R., Boucher, J.-M., Augustin, J.-M.: Region-based image segmentation using texture statistics and level-set methods. In: Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, pp. II817–II820 (May 2006)
10. Liang, C.C., Park, A.Y., Guan, J.L.: In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. Nat. Protoc. 2(2), 329–333 (2007)
11. Menon, M.B., Ronkina, N., Schwermann, J., Kotlyarov, A., Gaestel, M.: Fluorescence-based quantitative scratch wound healing assay demonstrating the role of mapkapk-2/3 in fibroblast migration. Cell Motility and the Cytoskeleton 66(12), 1041–1047 (2009)

12. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition 29(1), 51–59 (1996)
13. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. Journal of Comp. Physics 79(1), 12–49 (1988)
14. Permuter, H., Francos, J., Jermyn, I.: A study of Gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recog. 39, 695–706 (2006)
15. Sandberg, B., Chan, T., Vese, L.: A level-set and Gabor-based active contour algorithm for segmenting textured images. Technical Report 39, Math. Dept. UCLA, USA (2002)
16. Song, B., Chan, T.F.: A fast algorithm for level set based optimization. CAM-UCLA 68, 02–68 (2002)

Level Set Segmentation with Shape and Appearance Models Using Affine Moment Descriptors

Carlos Platero, María Carmen Tobar, Javier Sanguino, José Manuel Poncela, and Olga Velasco

Applied Bioengineering Group, Technical University of Madrid

Abstract. We propose a level set based variational approach that incorporates shape priors into edge-based and region-based models. The evolution of the active contour depends on local and global information. It has been implemented using an efficient narrow band technique. For each boundary pixel we calculate its dynamic according to its gray level, the neighborhood and geometric properties established by training shapes. We also propose a criterion for shape aligning based on affine transformation using an image normalization procedure. Finally, we illustrate the benefits of the our approach on the liver segmentation from CT images.

1 Introduction

Level set techniques have been adapted to segment images based on numerous low-level criteria. Several edge-based and region-based models have been proposed without information priors. More recently, shape prior has been integrated into the level set framework [1–3]. There are two topics in this area: a) shape alignment and b) shape variability. The first issue is to calculate the set of pose parameters (rotation, translation and scaling) used to align the template set, and hence remove any variations in shape due to pose differences. And the second question is the deformation of the shape, which is typical derived from a training set using Principal Component Analysis (PCA) on the distance surface to the object [1, 4].

In this paper, we introduce a new approach for shape variability, which combines a parametric registration of shapes by affine moment descriptors with shapes encoded by their signed distance functions. We solve the shape alignment using the image normalization procedure [5], which avoids increasing the number of coupled partial differential equations of the problem. Finally, a new active contour evolves using a combination the appearance terms and shape terms. The paper is organized as follows: in Section 2, we show the problem of the shape alignment and our approach based on image normalization. Section 3 describes the problem of building a shape term and present one based on the two shape distance measure. Section 4 presents our framework for image segmentation. Finally, in Section 5, we apply our procedure for liver segmentation from CT images.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 109–116, 2011.

2 Shape Alignment

Consider a training set consisting of N binary images $\{S_i\}_{i=1,..N} : \Omega \subset \mathbb{R}^n \rightarrow \{0,1\}$, $n = 2$ or 3. Our first aim is to align it, in order to avoid artifacts due to different pose. The new aligned images are defined as $\tilde{S}_i = S_i \circ T_i^{-1}$, where T_i is an affine transformation, given by the composition of a rotation, a scaling transformation and a translation. Equivalently, S_i can be written in terms of $\tilde{S}_i$ as: $S_i = \tilde{S}_i \circ T_i$. Traditionally, the pose parameters have been estimated minimizing the following energy functional, via gradient descent [2]:

$$E_{align} = \sum_{\substack{i,\,j\,=\,1 \\ i \neq j}}^{N} \frac{\int_\Omega (\tilde{S}_i - \tilde{S}_j)^2 \, dx}{\int_\Omega (\tilde{S}_i + \tilde{S}_j)^2 \, dx}. \tag{1}$$

Minimizing (1) is equivalent to simultaneously minimizing the difference between any pair of binary images in the training database. We propose to improve this approach by using a normalization procedure over the shape priors. An advantage is that the affine transformation is defined by closed-form expressions involving only geometric moments. No additional optimization over pose parameters is necessary. This procedure will be applied both to the N aligned training shapes and to the problem of aligning the active contour. Specifically, given a reference image S_{ref}, we propose a criterion for alignment based on a shape normalization algorithm. It is only necessary to computer the first and the second order moments. The first-order moments locate the centroid of the shape and the second-order moments characterize the size and orientation of the image. Given a binary image S_i, we compute the second-order moment matrix, and the image is rotated using the eigenvectors and it is scaling along the eigenvectors according to the eigenvalues of the second-order moment matrix of S_i and S_{ref}. Then, it is translated to the centroid. We do not consider the problem of reflection (for this see [6]).

If we only consider moments up to order two, S_i is approximated to an ellipse/ellipsoids centered at the image centroid. The ellipse/ellipsoids rotate angles and the semi-axes are determined by the eigenvalues and the eigenvectors of the second-order moment matrix [5]. We denote R as the rotation matrix.

Let S_{ref} be a normalized binary image as reference and $\{\lambda_j^{ref}\}_{j=1,..,n}$ be the eigenvalues for the reference image. We consider one of the following scale matrices: a) $W = \sqrt{\frac{\lambda^{ref}}{\lambda^i}} \cdot I$ where $\lambda = (\prod_{j=1}^n \lambda_j)^{1/n}$ and I is the identity matrix or b) W is diagonal matrix where $w_{j,j} = \sqrt{\frac{\lambda_j^{ref}}{\lambda_j^i}}$. In the first case is a scaling identical in all directions, while in the second case the size fits in each principal axis as the reference. The first option is used for shape priors without privileged directions otherwise the second option should be used. Finally, if the reference centroid is $\overline{x}_{ref}$, the affine transformation translates the origin of the coordinate system to the reference centroid. Then, the affine transformation is defined as follows:

$$T_i^{-1}(x) = R \cdot W \cdot (x - \overline{x}_i) + \overline{x}_{ref} \tag{2}$$

This affine transformation aligns from S_i to S_{ref}. If we use a scaling identical in all directions, S_{ref} will be only a numeric artefact for the pose algorithm. The alignment error does not depend on the reference, S_{ref}. But when each principal axis is adjusted to the reference, the alignment error depends on the choice of the reference. We can not guarantee the optimal pose for any shape. But neither the gradient descent method guaranteed to find the optimum because there is no evidence that the functional (1) is convex. Our procedure is fast and optimum if the shapes are closed to ellipses or ellipsoids. Section 5 will compare our approach with the variational method of (1).

3 Handling Shape Variability

Early work on this problem involves the construction of shapes and variability based on a set of training shapes via principal component analysis (PCA). In recent years, researchers have successfully introduced prior information about expected shapes into level set segmentation. Leventon et al. [1] modeled the embedding function by principal component analysis of a set of training shapes and added appropriate driving terms to the level set evolution equation. Tsai et al. [2] suggested a more efficient formulation, where optimization is performed directly within the subspace of the first few eigenvectors. Following these works, suppose now that the N aligned shapes $\tilde{S}_i$ define N objects, whose boundaries are embedded as the zero level set of N signed distance functions $\{\tilde{\phi}_1, \ldots, \tilde{\phi}_N\}$ respectively, which assign positive distances to the inside of the object, negative to the outside. Now, we constrain the level set function $\tilde{\phi}$ to a parametric representation of the shape variability [2]:

$$\tilde{\phi}_{\boldsymbol{\alpha}}(x) = \tilde{\phi}_0(x) + \sum_{i=1}^{k} \alpha_i \tilde{\psi}_i(x) \tag{3}$$

where $k \leq N$ is empirically chosen and $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ collects the weights for the first k eigenvectors $\tilde{\psi}_i(x)$, $i = 1, \ldots, k$. We also have the mean level set function $\tilde{\phi}_0(x) = \frac{1}{N} \sum_{i=1}^{N} \tilde{\phi}_i(x)$ of the aligned shape database.This will drive the shape variability, where the parameter vector $\boldsymbol{\alpha}$ models shape deformations. Experimentally, we have observed that given any parameter vector $\boldsymbol{\alpha}$, the shape generated by (3) is also a normalized shape. It preserves the centroid, the orientation and the product of the eigenvalues remains constant.

3.1 Shape Model

Each aligned training shape $\tilde{\phi}_i$ can be represented by its corresponding shape parameter $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ik})$. Cremers et al. [3] have introduced nonlinear statistical shape priors based on kernel density estimation. The goal of statistical shape learning is to infer a statistical distribution $\mathcal{P}(\alpha)$ from these samples. Following [3], it considers a nonparametric density approximation:

$$\mathcal{P}(\boldsymbol{\alpha}) = \frac{1}{N\sigma} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{\alpha} - \boldsymbol{\alpha_i}}{\sigma}\right) \tag{4}$$

where $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{|z|^2}{2})$, being $\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \min_{i \neq j} |\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j|^2$ the average nearest neighbor distance . It combines the nonparametric shape prior and a data term within a Bayesian framework to form the energy functional for segmentation. The data term use to minimize the probability of misclassified pixels for two regions [7]. Let $u : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be the image to be segmented and let $\tilde{\phi}$ be the level set function with $\tilde{\phi}(x) > 0$ if $x \in \Omega_{in}$ and $\tilde{\phi}(x) < 0$ if $x \in \Omega_{out}$. We also consider the regularized Heaviside function $H(s)$. The data energy functional is written as

$$\begin{aligned} E_{data}(\boldsymbol{\alpha}) = &- \int_{\Omega} H(\tilde{\phi}_{\boldsymbol{\alpha}}(x)) \log(p_{in}(u(T^{-1}(x)))) dx \\ &- \int_{\Omega} (1 - H(\tilde{\phi}_{\boldsymbol{\alpha}}(x))) \log(p_{out}(u(T^{-1}(x)))) dx \end{aligned} \tag{5}$$

being T^{-1} the affine transformation which accommodates shape variability due to differences in pose, it is also calculated using geometric moments. With the nonparametric models for shape and intensity introduced above, this leads to an energy of the form

$$E(\boldsymbol{\alpha}) = E_{data}(\boldsymbol{\alpha}) - \log(\mathcal{P}(\boldsymbol{\alpha})). \tag{6}$$

This approach is quite robust with respect to initialization and noise. However, it has also been observed that the evolution becomes inefficient when the object shape to be segmented varies significantly with respect to the training base. Indeed, when the dimension of the shape parameter vector is much lower than the number of elements of the active contour, then the decision of each element of the boundary can not be taken with a vector so generic. Therefore, we propose a signed distance-based measure. Let Φ be the level set function for segmentation, and $\tilde{\phi}_{\boldsymbol{\alpha}}$ be the one embedding the shape model as (3). Both are signed distance functions. Hence, we propose the following shape term:

$$E_{shape}(\Phi, \tilde{\phi}_{\boldsymbol{\alpha}}) = \frac{1}{2} \int_{\Omega} (\Phi(x) - \tilde{\phi}_{\boldsymbol{\alpha}}(T(x)))^2 dx. \tag{7}$$

It minimizes the dissimilarity measure between the target $\Phi(x)$ and the shape model $\tilde{\phi}_{\boldsymbol{\alpha}}(T(x))$. The pose parameter is calculated with the moments of $\Phi(x) \geq 0$ and $\tilde{\phi}_{\boldsymbol{\alpha}}(x) \geq 0$. The optimization criterion produces local pixel-wise deformation.

4 Our Model

Several edge-based and region-based models have been proposed. These two types of models are usually based on the fact that the regions are piecewise constant and the borders have high slopes. However, the regions of interest are not usually statistically homogeneous; noise, weak edges and small objects are also presented in most of the real images. We propose a new procedure that takes all these features into account. The basic idea is to pre-process the image with a filter based on the nonlinear diffusion techniques. The segmentation problem was

formalized by Mumford and Shah as the minimization of an energy functional that penalizes deviations from smoothness within regions and the length of their boundaries. Under certain assumptions on the discretization, the nonlinear diffusion filters are connected with minimizing the Mumford Shah functional [8]. Following this idea, we have used a nonlinear diffusion filter and we have also applied a stopping time criteria for obtaining piecewise smooth images. We can certainly assume that the nonlinear diffusion produces a filtered image where the intensity distributions are closed to Gaussian functions, modeled with the mean and variance for each region. Let us consider a image made up of two regions and integrating the maximization of the a posteriori probability and the regularity constraint, we obtain the following energy [7]:

$$E_{region}(\Phi) = -\int_{\Omega} H(\Phi(x)) \log p_{in}(u(x)|\mu_{in}, \sigma_{in}^2)dx$$
$$-\int_{\Omega} (1 - H(\Phi(x))) \log(p_{out}(u(x)|\mu_{out}, \sigma_{out}^2))dx + \nu \int_{\Omega} |\nabla \Phi(x)|dx \quad (8)$$

where $p(u|\mu_i, \sigma_i)$ denotes the probability of observing the filtered image u when Ω_i is a region of interest and ν is a positive constant chosen empirically. On the other hand, Kimmel and Bruckstein [9] have developed a novel and robust edge integration scheme. The edge-based stopping term serves to stop the contour on the desired object boundary. The active contour evolves along the second-order derivative in the direction of the image gradient. The functional is given by

$$E_{edge}(\Phi) = -\int_{\Omega} \delta(\Phi(x)) \nabla u \cdot \boldsymbol{n} dx + \int_{\Omega} H(\Phi(x)) \text{div}\left(\frac{\nabla u}{\|\nabla u\|}\right) \|\nabla u\| dx \quad (9)$$

being $\delta()$ a regularized version of the Dirac function such as $\delta(s) = H'(s)$ and $\boldsymbol{n}$ is the normal unit vector to zero level set. Finally, the global functional is a weighted sum of the above ones where is combined data terms and shape priors:

$$E(\Phi, \tilde{\phi}_{\boldsymbol{\alpha}}) = \varrho_1 E_{region}(\Phi) + \varrho_2 E_{edge}(\Phi) + \varrho_3 E_{shape}(\Phi, \tilde{\phi}_{\boldsymbol{\alpha}}) \quad (10)$$

where ϱ_i are positive constants chosen empirically. Obviously, there are two evolutions, the first one is the active contour for the object segmentation and the second one represents the deformation model evolution. The two evolutions are related through the proposed affine transformation between $\Phi(x) \geq 0$ and $\tilde{\phi}_{\boldsymbol{\alpha}}(x) \geq 0$. Moreover, both of them use the same statistical data improving the algorithm.

4.1 Numerical Algorithms

In this subsection, we show the numerical algorithms for minimizing the functionals presented. We can not guarantee that our functionals are convex. Therefore, gradient descent process stops at a local minima. One challenge is tracking down the significant minimum. This is done by initializing the active contour near of the object of interest. Following these observations, we propose a method of

two main stages: (i) an initial segmentation is generated using a combination of traditional techniques, and (ii) we deform locally this initial solution using a level set approach which combines edge alignment, homogeneity terms and shape dissimilarity measures.

The level set evolution algorithm uses an efficient distance preserving the narrow band technique[10]. The reinitialization of the level set is not longer necessary. The algorithm is implemented using a simple finite difference scheme. It is based on the following internal energy term: $\gamma \int_\Omega \frac{1}{2}(\|\nabla \Phi(x)\| - 1)^2 dx$. Here $\gamma > 0$ is a parameter that controls the effect of penalizing the deviation of Φ from a signed distance function. Finally, the global functional is a weighted sum of the above functionals. The resulting evolution of the level set function is the gradient flow that minimizes the overall energy functional:

$$\partial_t \Phi(x) = \delta(\Phi(x)) \left[\varrho_1 \cdot \log \frac{p_{in}(u(x))}{p_{out}(u(x))} + \varrho_1 \cdot \nu \cdot \operatorname{div}\left(\frac{\nabla \Phi(x)}{\|\nabla \Phi(x)\|} \right) - \varrho_2 \cdot u_{\eta\eta}(x) \right] + \varrho_3 \cdot (\Phi(x) - \tilde{\phi}_{\boldsymbol{\alpha}}(T(x))) + \gamma \cdot \left(\Delta \Phi(x) - \operatorname{div}\left(\frac{\nabla \Phi(x)}{\|\nabla \Phi(x)\|} \right) \right) \quad (11)$$

where $u_{\eta\eta}$ is the second derivative of u in the gradient direction. Recall that the affine transformation $T(x)$ is defined by the pose parameter. It has connected the pose from the normalized shape model to the target $\Phi(x) \gtrsim 0$. This property makes more efficient the algorithm since it allows to pre-load $\tilde{\phi}_0(x)$ and $\tilde{\psi}(x)_{i=1,\ldots,k}$. Moreover, $\tilde{\phi}_{\boldsymbol{\alpha}}(x)$ is calculated using the above data about $p_{in}(u(x))$ and $p_{out}(u(x))$. Gradient descent method is now used to find the shape parameter $\boldsymbol{\alpha}$ that minimizes $E(\boldsymbol{\alpha})$ in equation (6):

$$\partial_t \boldsymbol{\alpha} = \int_\Omega \delta(\tilde{\phi}(x)) \tilde{\boldsymbol{\psi}}(x) \left[\log \frac{p_{in}(u(T^{-1}(x)))}{p_{out}(u(T^{-1}(x)))} \right] dx + \frac{\sum_{i=1}^N (\boldsymbol{\alpha}_i - \boldsymbol{\alpha}) K_i}{\sigma^2 \sum_{i=1}^N K_i} \quad (12)$$

with $K_i = K(\frac{\alpha - \alpha_i}{\sigma})$ and $\tilde{\boldsymbol{\psi}} = \{\tilde{\psi}_1, \tilde{\psi}_2, ..., \tilde{\psi}_k\}$.

5 Liver Segmentation from 3D CT Images

Liver segmentation from 3D CT images is usually the first step in the computer-assisted diagnosis and surgery systems for liver diseases. Algorithms relying solely on image intensities or derived features usually fail. To deal with missing or ambiguous low-level information, shape prior information has been successfully employed. However, the lack of knowledge about the location, orientation and deformation of the liver, due to diseases or different acquisition procedures, adds another major challenge to any segmentation algorithm. In order to address these problems, we have applied our framework to liver segmentation. The proposed method has been trained on 20 patient CT slice set, and tested on another 10 specified CT datasets. The shape model is composed of 20 segmented livers. The segmented livers are aligned by the proposed procedure. In this case, each principal axis is adjusted to the reference. Experimentally, S_{ref} has been

tuned by minimizing (1). Using *Similarity Index*, $\left(SI = \frac{1}{N}\sum_i \frac{(S_i \circ T_i^{-1}) \cap S_{ref}}{(S_i \circ T_i^{-1}) \cup S_{ref}}\right)$, our approach gets a SI of 0.66 while variational method of (1) gives 0.54 over the training set. From aligned training set, we calculate and save the mean level set function $\tilde{\phi}_0(x)$, the first k eigenvectors $\tilde{\psi}(x)$ and the shape parameter $\boldsymbol{\alpha}_i$, for each sample. In this application, we use $k = 10$. Our approach starts filtering the CT image by a nonlinear diffusion filter with selection of the optimal stopping time. Once the image has been processed, an intensity model is built, $p_{in}(u)$ and $p_{out}(u)$. We also calculate $u_{\eta\eta}$. We load the mean level set function $\tilde{\phi}_0(x)$, the first eigenvectors $\tilde{\psi}(x)$ and the shape parameter $\boldsymbol{\alpha}_i$, $i = 1, \ldots, 20$. Then, region growing and 3D edge detector are applied to the filtered image. Morphological post-processing merges the previous steps, giving the initial solution and initializing the contour. The zero level of Φ is employed as the starting surface. The evolution of $\Phi(x)$ follows (11) and $\boldsymbol{\alpha}$ is calculated as (12). The constant parameters of the active contour were tuned to segmentation scores using the leave-one-out technique. The pose parameter is calculated with the moments of $\Phi(x) \geq 0$ and $\tilde{\phi}_{\boldsymbol{\alpha}}(x) \geq 0$. Fig. 1 shows slices from two cases, drawing the result of the method (in blue) and the reference (in red). The quality of the segmentation and its scores are based on the five metrics[11]. Each metric was converted to a score where 0 is the minimum and 100 is the maximum. Using this scoring system one can loosely say that 75 points for a liver is comparable to human performance. Table 1 lists the average values of the metrics and their scores over the test data set.

Table 1. Average values of the metrics and scores for all ten test case: volumetric overlap error (m_1), relative absolute volume difference (m_2), average symmetric surface distance (m_3), root mean square symmetric surface distance (m_4) and maximum symmetric surface distance (m_5)

Type	m_1	m_2	m_3	m_4	m_5
metrics	12.6%	4.7%	1.84 mm	3.86 mm	21.9 mm
scores	51	75	54	46	71

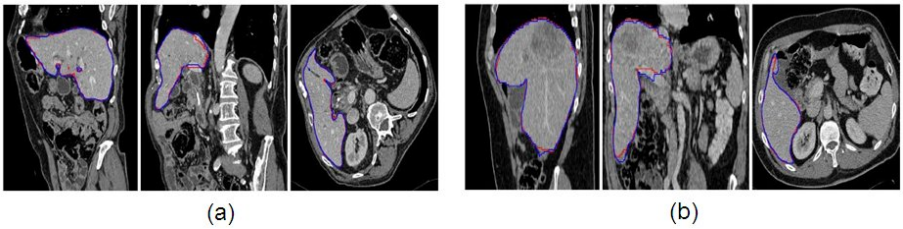


Fig. 1. From left to right, a sagittal, coronal and transversal slice for an easy case (a) and a difficult one (b). The outline of the reference standard segmentation is in red, the outline of the segmentation of the method described in this paper is in blue.

6 Conclusion

We have presented two main contributions. Firstly, the shape alignment has been solved using an image normalization procedure. An advantage is that the proposed affine transformation is defined by closed-form expressions involving only geometric moments. No additional optimization over pose parameters is necessary. This procedure has been applied both to the training shapes and to the problem of aligning the active contour. Secondly, we have proposed a level set based variational approach that incorporates shape priors into edge-based and region-based models. Using the Cremers' shape model, we have integrated a shape dissimilarity measure, a piecewise smooth region-based model and an edge alignment model. For each boundary pixel, our approach calculates its dynamic according to its gray level, the neighborhood and geometric properties established by training shapes.

References

1. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape influence in geodesic active contours. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE Computer Society, Los Alamitos (2000)
2. Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. IEEE Transactions on Medical Imaging 22(2), 137–154 (2003)
3. Cremers, D., Rousson, M.: Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: Suri, J.S., Farag, A. (eds.) Parametric and Geometric Deformable Models: An application in Biomaterials and Medical Imagery. Springer, Heidelberg (2007)
4. Cootes, T., Taylor, C., Cooper, D., Graham, J., et al.: Active shape models-their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
5. Pei, S., Lin, C.: Image normalization for pattern recognition. Image and Vision Computing 13(10), 711–723 (1995)
6. Heikkila, J.: Pattern matching with affine moment descriptors. Pattern Recognition 37(9), 1825–1834 (2004)
7. Rousson, M., Brox, T., Deriche, R.: Active unsupervised texture segmentation on a diffusion based feature space. In: CVPR, pp. 699–706. IEEE Computer Society, Los Alamitos (2003)
8. Kawohl, B.: From Mumford-Shah to Perona-Malik in image processing. Mathematical Methods in the Applied Sciences 27(15), 1803–1814 (2004)
9. Kimmel, R., Bruckstein, A.: Regularized laplacian zero crossings as optimal edge integrators. International Journal of Computer Vision 53(3), 225–243 (2003)
10. Li, C., Xu, C., Gui, C., Fox, M.: Level set evolution without re-initialization: A new variational formulation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 430–436 (2005)
11. van Ginneken, B., Heimann, T., Styner, M.: 3D segmentation in the clinic: A grand challenge. In: Proceedings of MICCAI Workshop 3D Segmentation in the Clinic: A Grand Challenge, pp. 7–15 (2007)

Automatic HyperParameter Estimation in fMRI*

David Afonso, Patrícia Figueiredo, and J. Miguel Sanches

Institute For Systems and Robotics, Instituto Superior Técnico,
Lisboa, Portugal
{jmrs,dafonso}@isr.ist.utl.pt,
patricia.figueiredo@ist.utl.pt

Abstract. *Maximum a posteriori* (MAP) in the scope of the Bayesian framework is a common criterion used in a large number of estimation and decision problems. In image reconstruction problems, typically, the image to be estimated is modeled as a Markov Random Fields (MRF) described by a Gibbs distribution. In this case, the Gibbs energy depends on a multiplicative coefficient, called *hyperparameter*, that is usually manually tuned [14] in a trial and error basis.

In this paper we propose an automatic *hyperparameter* estimation method designed in the scope of *functional Magnetic Resonance Imaging* (fMRI) to identify activated brain areas based on *Blood Oxygen Level Dependent* (BOLD) signal.

This problem is formulated as classical binary detection problem in a Bayesian framework where the estimation and inference steps are joined together. The prior terms, incorporating the a priori physiological knowledge about the *Hemodynamic Response Function* (HRF), drift and spatial correlation across the brain (using edge preserving priors), are automatically tuned with the new proposed method.

Results on real and synthetic data are presented and compared against the conventional *General Linear Model* (GLM) approach.

Keywords: HyperParameter, Estimation, Bayesian, fMRI, HRF.

1 Introduction

The detection of neuronal activation based on BOLD signals measured using fMRI is one of the most popular brain mapping techniques.

The data are classically analyzed using the *Statistical Parametric Mapping* (SPM) technique [11,10] where the *General Linear Model* GLM is used to describe the observations at each voxel and the corresponding coefficients are estimated by using the Least Square Method [14]. The detection of the activated

* This work was supported by the Portuguese Government – FCT (ISR/IST plurianual funding. We acknowledge GinoEco, Porto, for collaboration in collecting the fMRI data).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 117–125, 2011.

regions is performed from the estimated *explanatory variables* (EV's) by using a second step of classical inference approach based on the Neyman-Pearson theorem.

Still, Bayesian approaches have been gaining popularity since they provide a formal method of incorporating prior knowledge in data analysis [6,12]. In [8], Groutte et al. propose a non-parametric approach where a *Finite Impulse Response* (FIR) filter is used to describe the HRF and smoothing constraints are imposed at the solution by using a regularization matrix. Ciuciu et al. describe another non-parametric approach for the Bayesian estimation of the HRF in [3]. The authors make use of temporal prior terms to introduce physiological knowledge about the HRF. Basic and soft constraints are incorporated in the analysis, namely the considerations that the HRF starts and ends at zero and that the HRF is a smooth function. In [16] the authors propose a Bayesian approach in which the data noise is estimated using a spatio-temporal model and propose a half-cosine functions HRF model based on their experimental findings. In Yet another Bayesian approach based on the mathematical formalism of the GLM is proposed in [1]. The authors describe an SPM algorithm based on the maximum a posteriori (MAP) criterion to jointly estimate and detect the activated brain areas characterized by binary coefficients. The prior term introduced for these parameters comprises a bimodal distribution defined as the sum of two Gaussian distributions centered at zero and one.

In this paper we further improve this last method [1] by implementing an automatic HyperParameter estimation method to automatically set the prior strength in the Bayesian estimation of a MRF described by a Gibbs distribution. Additionally, data drift estimation is incorporated and the spatial correlation between neighbors is taken into account by using *edge preserving* priors that promote piecewise constant region solutions. The optimization of the overall energy function with respect to the activation binary variables is performed by using the *graph-cuts* (GC) based algorithm described in [2], which is computationally efficient and is able to find out the global minimum of the energy function.

2 Problem Formulation and Method

By making use of the problem formulation and variables defined in [1] we further incorporate a slow time data drift variable (a.k.a. baseline) $\mathbf{d_i}$ of time dimension N into the observation model, yielding eq. 1 at each i^{th} voxel, when L stimuli are applied simultaneously.

$$y_i(n) = \overbrace{h_i(m) * \underbrace{\sum_{k=1}^{L} \beta_i(k) p_k(n)}_{x_i(n)}}^{z_i(n)} + d_i(n) + \eta_i(n), \tag{1}$$

where $y_i(n)$ is the N length observed BOLD signal at the i^{th} voxel, $h_i(m)$ is the $M \leq N$ dimensional HRF and $\eta_i(n)$ is noise signal. The activity unknowns

$\beta_i(k) \in \{0,1\}$ are binary with $\beta_i(k) = 1$ if the i^{th} voxel is activated by the k^{th} $p_k(n)$ stimulus. $\eta_i(n) \sim \mathcal{N}(0, \sigma_y^2)$ is assumed *Additive White Gaussian Noise* (AWGN) which is an acceptable assumption mainly if a prewhitening preprocessing [9] of the data is performed.

The *maximum a posteriori* (MAP) estimation of the unknown column vectors $\mathbf{b}_i = \{\beta_i(1), ..., \beta_i(L)\}^T$, $\mathbf{h}_i = \{h_i(1), ..., h_i(N)\}^T$ and $\mathbf{d}_i = \{d_i(1), ..., d_i(N)\}^T$ is obtained, in matrix form, by minimizing the following energy function

$$\begin{aligned} E(\mathbf{y}_i, \mathbf{b}_i, \mathbf{h}_i, \mathbf{d}_i) &= \overbrace{E_y(\mathbf{y}_i, \mathbf{b}_i, \mathbf{h}_i, \mathbf{d}_i)}^{\text{Data fidelity term}} + \overbrace{E_h(\mathbf{h}_i) + E_d(\mathbf{d}_i)}^{\text{Prior terms}} \\ &= -\log(p(\mathbf{y}_i|\mathbf{h}_i, \mathbf{h}_i,)) - \log(p(\mathbf{h}_i)) - \log(p(\mathbf{d}_i)) \end{aligned} \tag{2}$$

where the prior terms incorporate the *a priori* knowledge [13] about the temporal behavior of $\mathbf{h}_i$ and $\mathbf{d}_i$ - the HRF ($C1$) starts and ends at 0; ($C2$) is smooth; ($C3$) has similar magnitude to the HRF one gamma function ($h^c(t)$) proposed in [5,9]; and that the $\mathbf{d}_i$ is ($C4$) a slow varying signal with a smaller *bandwidth* than the one of $\mathbf{h}_i$.

By the *Hammersley-Clifford* theorem [7] and *Markov Random Fields* theory, these constraints may be imposed in the form of the following *Gibbs distributions*

$$p(\mathbf{h}_i) = \frac{1}{Z_h} e^{-\alpha U(\mathbf{h}_i)} \tag{3}$$

$$p(\mathbf{d}_i) = \frac{1}{Z_d} e^{-\gamma U(\mathbf{d}_i)} \tag{4}$$

where Z_h and Z_d are partition functions and the *Gibbs* energies $U(\mathbf{h}_i)$ and $U(\mathbf{d}_i)$ are designed in the following way, where $[\alpha, \gamma]$ are regularization parameters to tune the degree of smoothness of the estimated vectors.

$$U(\mathbf{h}_i) = \overbrace{w_h(1)}^{\text{C3}} \underbrace{h(1)^2}_{\text{C1}} + \overbrace{w_h(M)}^{\text{C3}} \underbrace{h(M)^2}_{\text{C1}} + \underbrace{\sum_{n=2}^{M-1} \overbrace{w_h(n)}^{\text{C3}} \left[\overbrace{(h_i(n+1) - h_i(n)) - (h_i(n) - h_i(n-1))}^{\text{Discrete version of the } 2^{nd} \text{ derivative}} \right]^2}_{\text{C2}} \tag{5}$$

$$U(\mathbf{d}_i)_i = \underbrace{\sum_{n=2}^{N} \left[\overbrace{d_i(n) - d_i(n-1)}^{\text{Discrete version of the } 1^{st} \text{ derivative}} \right]^2}_{\text{C4}} \tag{6}$$

Here the weigh coefficients $w_h(n) = 1/(|h^d(n)| + 10^{-6})^2$, where the discrete version of the HRF gamma function is $h^d(n) = h^c(t)|_{t=n \times TR}$, are used to compensate for the reduced prior strength when the second derivatives are small.

Its can be shown that the overall energy eq. (2) is rewritten as follows

$$E = \frac{1}{2\sigma_y^2} \|(\mathbf{\Psi}_i \mathbf{b}_i + \mathbf{d}_i - \mathbf{y}_i)\|_2^2 + \alpha \mathbf{h}_i^T \mathbf{H}_0 \mathbf{D}_h \mathbf{h}_i + \gamma \mathbf{d}_i^T \mathbf{D}_d \mathbf{d}_i \tag{7}$$

where $\mathbf{H_0} = diag\{h^d(n)\}$ is a $M \times M$ diagonal matrix containing the HRF. $\mathbf{D}_h$ and $\mathbf{D}_d$ are $M \times M$ and $N \times N$ second and first order difference matrix operators [14], respectively. $\mathbf{\Psi}_i$ is a toeplitz $L \times N$ convolution matrix of $\mathbf{b}_i$ and $\mathbf{b}_i$ as defined in [1]

3 Optimization

The MAP estimation of the unknown vectors $\mathbf{b}_i$, $\mathbf{h}_i$ and $\mathbf{d}_i$ is obtained by minimizing the energy function (7) with respect to each vector, one step at a time. $\mathbf{b}_i$ is first estimated with the drift initialized with the mean of the TC $(\mathbf{d}_i^0 = \bar{\mathbf{y}})_i$ and $\mathbf{h}_i^0$ equal to $h^d(n)$ [5,9].

3.1 Step One: b Estimation

Its easily shown that the binary elements of $\hat{\mathbf{b}}_i^t = \{\hat{\beta}_{k,i}^t\}$ that leads to the minimization of (7) are a simple binarization by thresholding ($thrs = 0$) of the following fields, in matrix notation, where $\psi_i^{t-1}(k)$ is the k^{th} column of $\mathbf{\Psi}_i^{t-1}$:

$$\mathcal{B}_i^t(k) = -\psi_i^{t-1}(k)^T \left[\psi_i^{t-1}(k) + 2(\mathbf{d}^{t-1} - \mathbf{y})\right] \tag{8}$$

To solve this huge combinatorial problem, a fast and computationally efficient *graph-cuts* based algorithm [2] is used to binarize the fields $\mathcal{B}_{r,l}^t(k)$, defined in (8), at each (r, l) pixel location in the data slice, by minimizing the following energy function:

$$\mathbf{\Sigma}(\beta_{r,l}(k), \mathcal{B}_{r,l}(k)) = \overbrace{\sum_{r,l} \mathcal{B}_{r,l}(k)(1 - \beta_{r,l}(k))}^{\text{data fidelity term}} + \underbrace{\hat{\sigma_y^2} \sum_{r,l} \left[V_{r,l}^v(k) + V_{r,l}^h(k)\right] / \tilde{g}_{r,l}}_{\text{spatial regularization term}} \tag{9}$$

where $\hat{\sigma_y^2}$ is the observed signal variance ($var(\mathbf{y})$); $V_{r,l}^v(k)$ and $V_{r,l}^h(k)$ are XOR $\oplus$ operators between $\beta_{r,l}(k)$ and its causal vertical $\beta_{r,l+1}(k)$ and horizontal $\beta_{r+1,l}(k)$ neighbors, respectively; $\tilde{g}_{r,l}$ $(10^{-2} \leq \tilde{g}_{r,l} \leq 1)$ is the normalized (smoothed) filtered gradient of $\mathcal{B}(k)$. Non-uniform solutions to (9) have a higher cost due to the *spatial regularization term*. However, in order to preserve transitions, the division by $\tilde{g}_{r,l}$ reduces this non-uniform cost at locations where the gradient magnitude is large. It can be shown that (9) is convex which guaranties [2] global minimum convergence.

3.2 Step Two: h Estimation

A new estimation of $\hat{\mathbf{h}}_i$ is calculated by finding the null derivative point of (2) with respect to $\mathbf{h}$, yelding:

$$\hat{\mathbf{h}}_i = \left[(\mathbf{\Phi}_i^t)^T \mathbf{\Phi}_i^t + 2\alpha\sigma_y^2 \mathbf{H_0} \mathbf{D}_h^T\right]^{-1} (\mathbf{\Phi}_i^t)^T (\mathbf{y}_i - \mathbf{d}_i^{t-1}) \tag{10}$$

where $\mathbf{\Phi}_i^t$ is calculated with the current $\mathbf{b}_i^t$ vector, estimated at the previous iteration step 3.1. However, the HRF is only estimated in the case of voxel activation by at least one paradigm, i.e., if $\exists_{(r,l)} : \hat{\beta}_{r,l}(k) > 0$.

HyperParameter Estimation. In this method the regularization parameter α is not constant but is automatically and adaptively estimated along the iterative process as follows. Considering (5), we can rewrite eq. (3) as

$$p(\mathbf{h}_i) = \prod_{n}^{N} \underbrace{\frac{1}{Z_h} e^{-\alpha w(n) \delta(n)^2}}_{p(\delta)} \tag{11}$$

where δ^2 is the 2^{nd} derivative operator in (5). By assuming $p(\delta(n))$ to be a probability density function (of unitary area) and $\alpha w(n) = \frac{1}{2\sigma(n)^2}$ we get

$$\frac{\sqrt{2\pi\sigma(n)^2}}{Z_h(n)} \underbrace{\int_{-\inf}^{+\inf} \frac{1}{\sqrt{2\pi\sigma(n)^2}} e^{\frac{\delta(n)^2}{2\sigma(n)^2}} d\delta(n)}_{=1} = 1 \tag{12}$$

which implies that $Z_h(n) = \sqrt{\frac{\pi}{\alpha w(n)}}$, hence the energy term of (2) with respect to h can be rewritten as

$$\mathbf{E}_h(h_i) = -\log(h) = \frac{N}{2}\log\pi - \frac{1}{2}\sum_{n}^{N}\log w(n) - \frac{N}{2}\log\alpha + \alpha\sum_{n}^{N} w(n)\delta(n)^2 \tag{13}$$

By finding the null derivative of (13) we obtain the automatic HyperParameter estimation that is, in each iteration, dependent on the initialization and current estimate of the HRF.

$$\alpha^t = \frac{N}{2U(h)} = \frac{\frac{N}{2}}{(\mathbf{h}_i^{t-1})^T(\mathbf{H_0}\mathbf{D}_h)\mathbf{h}_i^{t-1}} \tag{14}$$

3.3 Step Three: d Estimation

A new estimation of $\hat{\mathbf{d}}_i$ is calculated by finding the null derivative point of (2) with respect to $\mathbf{d_i}$, yelding:

$$\hat{\mathbf{d}}_i = \left[\mathbf{I} + 2\gamma\sigma_y^2\mathbf{D}_d^T\right]^{-1}(\mathbf{y}_i - \mathbf{\Psi}_i^t\mathbf{b}_i^t) \tag{15}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{\Psi}_i^t$ is computed by using the current $\hat{\mathbf{h}}_i^t$ vector, obtained in the previous iteration step.

Since γ is a regularization parameter associated with the drift signal, a much slower frequency signal than HRF, then γ should be higher than α, i.e., $\gamma \gg \alpha$ [15]. Here $\gamma = 100\alpha$.

4 Experimental Results

In this section tests with synthetic and real data are presented to illustrate the application of the algorithm and evaluate its performance.

4.1 Synthetic Data

The synthetic data is based on the well known *Shepp-Logan* image phantom with 256×256 pixels. The paradigm was generated in a block-design basis with 4 epochs, 60 *sec* each (30 *sec* of activation and 30 *sec* of rest) with $TR = 3$ *sec*.

Making use of (1) for generating the $\mathbf{y}_i$ observed data, a Monte Carlo experiment with a total of $3,276,800$ runs was performed with several different noise levels (see Fig. 1 captation). The resulting mean and standard deviation (error bars) values of probability of error (P_e), as a function of σ, SNR_{dB} and SNR, are presented in Fig.1. The results with and without GC are shown, as well as the ratio $P_e(GC)/P_e(w/GC)$.

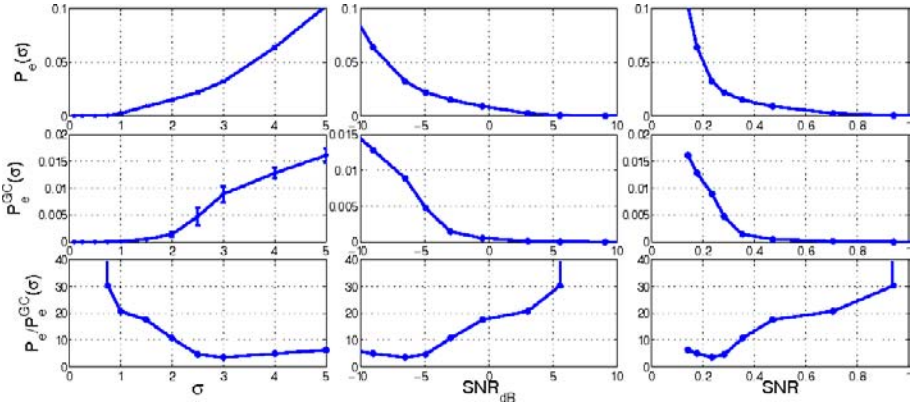


Fig. 1. Monte Carlo results for 50 runs on 256×256 pixels, for $\sigma = \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 5\}$. The mean and standard deviation (error bars) values of P_e, as a function of σ, SNR_{dB} and SNR, are presented on the first, second and third columns, respectively. The results with and without GC are displayed on the top and middle rows, respectively, and the ratio $[P_e(GC)/P_e(w/GC)]$ (σ is displayed on the bottom row.

These results demonstrate the activity detection robustness of the method, even in highly noisy data. They also show that taking into account the spatial correlation among neighboring voxels leads to a significant decrease in P_e. As expected, the improvement increases when the amount of noise increases or, equivalently, the SNR decreases. This is observed by the monotonic increasing behavior of the $[P_e(GC)/P_e(w/GC)]$ (σ).

4.2 Real Data

The real data used in this paper was acquired in the scope of a previous study [4] were two volunteers with no history of psychiatric or neurological diseases participated in a visual stimulation and a motor task fMRI experiment. Functional images were obtained using echo-planar imaging (EPI) with $TR/TE = 2000\ ms/50\ ms$. Datasets were pre-processed and analyzed using the FSL

software (http://www.fmrib.ox.ac.uk/fsl) for: motion correction; non-brain removal and mean-based intensity normalization. The data used in the standard SPM-GLM analysis using FSL (not for the proposed SPM-Drift-GC) was further pre-processed with spatial smoothing (Gaussian kernel, 5 mm FWHM) and high-pass temporal filtering (Gaussian-weighted least squares straight line fitting, 50 sec cut-off).

For the FSL processing a GLM approach with local autocorrelation correction was used on square stimulus functions convolved with the canonical Gamma HRF and it's first derivative [5,9]. Linear stimulus/baseline contrast analysis and t-tests are applied to obtain the SPM, followed by cluster thresholding by the Gaussian Random Fields (GRF) theory. Since the results provided by this "'standard"' method are depend on the inference p-value and clustering Z-score threshold values used, our experienced experimentalist provided two results of SPM-GLM: a relatively *strict* result and a more *loose* result, displayed on columns d) and e) of Fig. 2 respectively. The proposed SPM-Drift-GC method results are displayed on columns a), b) and c) of Fig. 2.

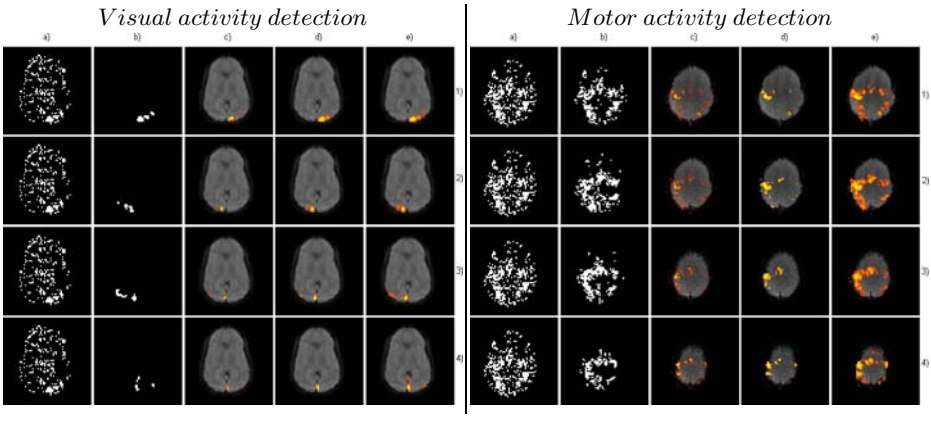


Fig. 2. Activated regions obtained by the new SPM-Drift-GC (a-b-c) and standard SPM-GLM (d-e) methods, on the visual (left) and motor (right) real data, where each row (*1), 2), 3)* and *4)*) corresponds to a different stimulus. Left to right: a) Binary SPM-Drift algorithm results without *GraphCuts*; b) Binary SPM-Drift-GC algorithm results; c) Weighted SPM-Drift-GC algorithm results; d) SPM-GLM algorithm *Strict* results; e) SPM-GLM algorithm *Loose* results. Activation intensity is color coded from red (0) to yellow (1) and is overlaid on the EPI brain image with linearly decreasing transparency from 100% ($activity = 0$) to 0% ($activity \geq 0.5$).

In general, visual inspection of the activation brain maps suggests good agreement between the methods, although the SPM-Drift-GC also detects some regions not present in the *strict* results, but present, most of them, in the *loose* results. However, in some brain slices, there are areas only detected as active by SPM-Drift-GC that correspond to low energy estimated HRF's (coded in transparent red) and somewhat deviant shaped HRF's from the rigid HRF restrictions of SPM-GLM.

5 Conclusions

In this paper, a new data-dependent and automatic estimation method for the HyperParameters of a MRF described by a Gibbs distribution is proposed and applyed in the detection of activated brain areas in fMRI. Here, estimation and inference are joined together and the drift and HRF estimation and iteratively estimated by taking into account the spatial correlation.

Monte Carlo tests with synthetic data are presented to characterize the performance of the algorithm in terms of error probability. The introduction of the final step with *graph-cuts* greatly improves the accuracy of the algorithm, yielding an error probability that is close to zero even at the high noise levels observed in real data.

Real data activation results are consistent with a standard GLM approach, and most importantly, the activation clusters are best matched with the ones obtained at a significance threshold validated by the specialist, but with the advantage that the specification of user-defined subjective thresholds are not required. With the proposed method it also becomes unnecessary to apply spatial smoothing an high-pass temporal filtering as pre-processing steps, while accounting for important physiological properties of the data by estimating the HRF.

References

1. Afonso, D., Sanches, J., Lauterbach, M.: Joint bayesian detection of brain activated regions and local hrf estimation in functional mri. In: Proceedings IEEE ICASSP 2008, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, USA, March 30 - April 4 (2008)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1222–1239 (2001)
3. Ciuciu, P., Poline, J.B., Marrelec, G., Idier, J., Pallier, C., Benali, H.: Unsupervised robust non-parametric estimation of the hemodynamic response function for any fmri experiment. IEEE Trans. Med. Imaging 22(10), 1235–1251 (2003)
4. Cruz, P., ao Teixeira, J., Figueiredo, P.: Reproducibility of a rapid visual brain mapping protocol. In: Proc. of the 15th Annual Meeting of the OHBM, San Francisco, US (June 2009)
5. Friston, K.J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M.D., Turner, R.: Event-related fMRI: characterizing differential responses. Neuroimage 7(1), 30–40 (1998)
6. Friston, K., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J.: Classical and Bayesian inference in neuroimaging: Theory. NeuroImage 16, 465–483 (2002)
7. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. Pattern Analysis and Machine Intelligence 6, 721–741 (1984)
8. Goutte, C., Nielsen, F.Å., Hansen, L.K.: Modelling the haemodynamic response in fmri with smooth fir filters. IEEE Trans. Med. Imaging 19(12), 1188–1201 (2000)
9. Jezzard, P., Matthews, P.M., Smith, S.M.: Functional magnetic resonance imaging: An introduction to methods. Oxford Medical Publications (2006)
10. Friston, K.J.: Analyzing brain images: Principles and overview. In: Frackowiak, R.S.J., Friston, K.J., Frith, C., Dolan, R., Mazziotta, J.C. (eds.) Human Brain Function, pp. 25–41. Academic Press, USA (1997)

11. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C., Frackowiak, R.S.J.: Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Human Brain Mapping 2, 189–210 (1995)
12. Makni, S., Ciuciu, P., Idier, J., Poline, J.B.: Joint detection-estimation of brain activity in functional mri: A multichannel deconvolution solution. IEEE Transactions on Acoustics, Speech, and Signal Processing 53(9), 3488–3502 (2005)
13. Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., Poline, J.B.: Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information. Human Brain Mapping 19, 1–17 (2003)
14. Moon, T.K., Stirling, W.C.: Mathematical methods and algorithms for signal processing. Prentice-Hall, Englewood Cliffs (2000)
15. Sanches, J., Marques, J.S.: A map estimation algorithm using IIR recursive filters. In: Proceedings International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition, Lisbon, Portugal (July 2003)
16. Woolrich, M.W., Jenkinson, M., Brady, J.M., Smith, S.M.: Fully bayesian spatio-temporal modeling of fmri data. IEEE Trans. Med. Imaging 23(2), 213–231 (2004)

Automatic Branching Detection in IVUS Sequences

Marina Alberti[1,2], Carlo Gatta[1,2], Simone Balocco[1,2], Francesco Ciompi[1,2], Oriol Pujol[1,2], Joana Silva[3], Xavier Carrillo[4], and Petia Radeva[1,2]

[1] Dep. of Applied Mathematics and Analysis, University of Barcelona, Spain
[2] Computer Vision Center, Campus UAB, Bellaterra, Barcelona, Spain
[3] Coimbra's Hospital Center, Cardiology Department, Coimbra, Portugal
[4] Unitat d'hemodinàmica cardíaca, Hospital universitari "Germans Trias i Pujol", Badalona, Spain
marina.alberti@cvc.uab.es

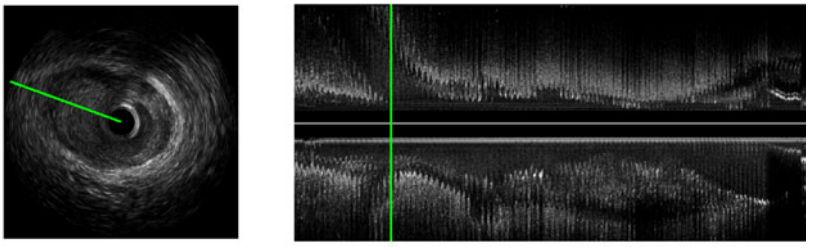
Abstract. Atherosclerosis is a vascular pathology affecting the arterial walls, generally located in specific vessel sites, such as bifurcations. In this paper, for the first time, a fully automatic approach for the detection of bifurcations in IVUS pullback sequences is presented. The method identifies the frames and the angular sectors in which a bifurcation is visible. This goal is achieved by applying a classifier to a set of textural features extracted from each image of an IVUS pullback. A comparison between two *state-of-the-art* classifiers is performed, AdaBoost and Random Forest. A cross-validation scheme is applied in order to evaluate the performances of the approaches. The obtained results are encouraging, showing a sensitivity of 75% and an accuracy of 94% by using the AdaBoost algorithm.

1 Introduction

Atherosclerosis is an inflammatory process affecting the arterial walls, evolving towards the formation of multiple plaques within the arteries. Atherosclerotic plaques can rupture or grow until they narrow the vessel, potentially leading to complications such as unstable angina, myocardial infarction, stroke and sudden cardiac death. It has been shown that specific vessel locations, such as bifurcations, are critical sites for plaque growth and rupture [1]. The treatment of bifurcations by percutaneous coronary intervention (PCI) represents 20% of all PCI procedures.

Intravascular Ultrasound (IVUS) is a catheter-based imaging technique, generally used for guiding PCI and also as a diagnostic technique. IVUS allows the visualization of high resolution images of internal vascular structures. The procedure for the acquisition of an IVUS sequence consists in inserting an ultrasound emitter, carried by a catheter, into the arterial vessel. The standard IVUS frame is a 360-degree tomographic cross-sectional view of the vessel walls (defined as *short-axis view*), which allows an accurate assessment of vessel morphology and tissue composition. Given a certain angle on the *short-axis view* (see Fig. 1-a),

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 126–133, 2011.



(a) (b)

Fig. 1. *Short-axis view* in correspondence of a bifurcation (a); *longitudinal view* of the pullback (b). The two lines indicate the angular and longitudinal bifurcation localizations, in (a) and (b) respectively.

the corresponding *longitudinal view* can be generated by considering the gray-level values of the whole sequence along the diameter at the fixed angle. A typical branching appearance in both *short-axis* and *longitudinal view* is illustrated in Fig. 1.

It has been shown that the use of IVUS, compared to the conventional angiography, reduces the four-year mortality in patients in image-guided bifurcation stenting PCI. Although the topic of automatic bifurcation detection has been investigated in several medical imaging modalities, it has never been addressed in IVUS. In this paper, we present, for the first time, a method for the automatic detection of bifurcations in IVUS pullback sequences. In particular, the frames and the angular sectors in which a bifurcation is visible are identified. This goal is obtained by means of a pattern recognition approach, where a set of textural features extracted from each image of an IVUS pullback provides a feature-space in which the classification is performed. The method takes into account several statistical measures computed on the image texture, calculated along the radius of the frames. The classification task is tackled by using the AdaBoost classifier. A comparison with another *state-of-the-art* discriminative classifier, Random Forest, is provided.

The paper is organized as follows: Section II gives an overview of the method, Section III shows the obtained results and Section IV concludes the paper.

2 Method

The method can be divided into two consecutive stages, as illustrated in the block diagram in Fig. 2. Firstly, the motion artifacts which affect the IVUS sequence due to heart beating are compensated; then, each angular sector is classified as bifurcation or not.

2.1 Motion Artifacts Compensation

During an IVUS pullback acquisition, the catheter is affected by several motion artifacts. The most relevant one is caused by the heart beating, which generates

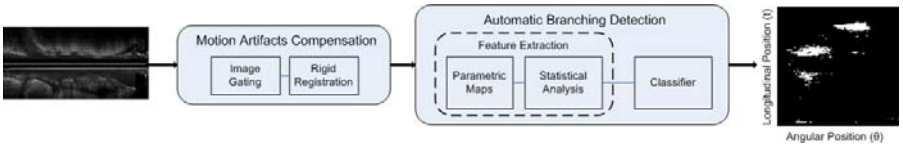
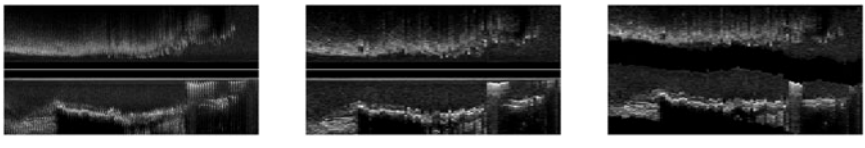


Fig. 2. Block diagram of the proposed approach

a repetitive longitudinal oscillation of the catheter (swinging effect) along the axis of the vessel, resulting in a possible multiple sampling of the same vessel positions. This undesired effect can be compensated by selecting optimally stable frames, for instance by using an image-based gating technique [2]. Another motion artifact is represented by the catheter fluctuation, causing a spatial misalignment of consecutive frames with respect to the real vessel morphology. In order to align the vessel centers in successive frames, we apply an IVUS registration method [3] consisting in a rigid translation of subsequent frames of the gated sequence. Figure 3 illustrates the output of the two stages.



(a) (b) (c)

Fig. 3. Longitudinal views of a pullback sequence before motion artifact compensation (a), after the application of the gating technique (b) and after the application of both gating and registration (c), respectively

2.2 Automatic Branching Detection

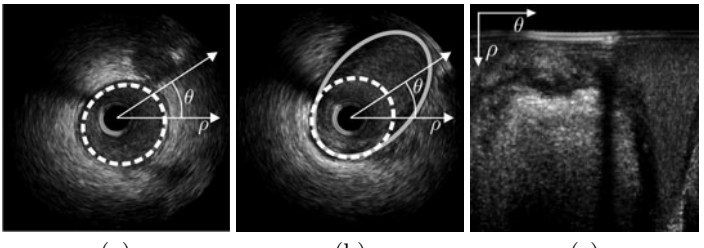
In order to identify bifurcations, we define a binary classification problem aimed at distinguishing between the angular sectors of the IVUS frames containing a bifurcation and the others. The angular analysis is justified by the fact that physicians report bifurcation positions in terms of both frame localization and angular extent. The detection task is based on a pattern recognition technique, in which a classifier is first trained by using a database of IVUS sequences manually labeled by experts and is successively used to identify the presence of a bifurcation in a new frame (test stage). The two main phases of a standard pattern recognition approach are now presented: feature extraction and classification.

Typically, in presence of a bifurcation, the appearance of the visible blood region in a *short-axis* IVUS image tends to an elliptical profile with an eccentricity that is higher with respect to a frame without bifurcation (See Fig. 4-a, -b). The method analyzes radial properties of the vessel texture, for detecting the angular sectors belonging to a bifurcation and in this way it explores the

above mentioned eccentricity property. For this purpose, each of the normalized IVUS pullback images $I(x, y) \in [0, 1]$ which constitutes the pullback sequence $S(x, y, t) \in [0, 1]$ is first converted into polar coordinates:

$$\widetilde{I}(\rho, \theta) = I(\rho \cos\theta, \rho \sin\theta) \tag{1}$$

where x and y are the horizontal and vertical coordinates in the cartesian system, ρ and θ are the radial and angular coordinates in the polar system (See Fig. 4-c), t is the longitudinal (temporal) coordinate along the pullback.



(a) (b) (c)

Fig. 4. *Short-axis view* of a non-bifurcation (a) and a bifurcation (b) frame, in cartesian coordinates. The dotted and continuous curves represent an approximation of the typical geometry of the blood region contour in each case. View of the bifurcation frame in the polar representation (c).

A set of N_T texture descriptors is then defined. Each descriptor specifies a mapping function $F : \widetilde{I}(\rho, \theta) \mapsto M_j(\rho, \theta)$, where $M_j(\rho, \theta) \in \mathbb{R}$ is the parametric map according to the j^{th} textural descriptor, $j = 1, \ldots, N_T$. In order to extract information related to the eccentricity, for each column of the parametric maps, a set of basic statistical features (standard deviation, mean, median and maximum values, position of the maximum value, histogram bins) is computed by means of a second mapping function, D:

$$D : M_j(\rho, \theta) \mapsto f_i(\theta) \tag{2}$$

where $f_i(\theta) \in \mathbb{R}$, $i = 1, \ldots, N_S$, being N_S the total number of statistical descriptors. Since in this step we are considering radial properties of the image, in order to extract homogeneous features the center of the image has to coincide with the vessel center. For this reason, the applied rigid registration proves to be a necessary step. Each column (angular sector) θ is then described by a feature vector, obtained by concatenating all the features, $f_i(\theta) = [f_1(\theta)\ f_2(\theta)\ \ldots\ f_{N_F}(\theta)]$. The used descriptors are: Gabor filters [4], Local Binary Patterns (LBP) [5] and Cross-correlation between successive frames [6]. The gray-level image is considered as one of the maps, as well. The total number of used features is $N_F = 166$.

We propose the application of the AdaBoost classification algorithm [7] with Decision Stump weak classifier, to implement the classification stage. The main

advantages of AdaBoost are its computational simplicity and speed, which make it particularly suitable for clinical applications. Moreover, the classifier is able to work with a large set of features and is not prone to overfitting.

3 Experimental Results

A set of 10 in-vivo pullbacks of human coronary arteries has been acquired with an iLab IVUS Imaging System (Boston Scientific) using a 40 MHz catheter Atlantis SR 40 Pro (Boston Scientific). Each sequence contains an average of 3000 frames, for a total amount of 24 bifurcations. In order to validate our approach, a ground-truth of bifurcation labels has been created by manual segmentation performed by two medical experts. To this aim, an *ad-hoc* interface has been developed. For each pullback, the physicians selected, for each bifurcation, the angle which comprises the bifurcation in the *short-axis view* and the initial and final branching frames in the *longitudinal view*. The procedure for ground-truth collection explains the relevance of the application of a gating technique, which avoids the presence of non-bifurcations samples into a longitudinal vessel segment labeled as bifurcation, otherwise caused by the swinging effect. The ground-truth labels are used for both the training and the validation of the methodology.

The classifier performance is assessed by means of the *Leave-One-Patient-Out* (LOPO) cross-validation technique, over the $N_p = 10$ sequences. At each validation fold, performance is evaluated in terms of accuracy (A), sensitivity (S), specificity (K), precision (P) and false alarm ratio (FAR). The positive and negative classes are strongly unbalanced; therefore, a normalization of the confusion matrix is applied for the computation of all the parameters, with the exception of the accuracy. Given the classification-based nature of the proposed methodology, a comparison with another *state-of-the-art* discriminative classifier is straightforward: we therefore compare the performance of AdaBoost to Random Forest [8]. The AdaBoost classifier has been trained with up to $T = 110$ rounds. The parameters of the Random Forest classifier have been set to a number of trees $N_{trees} = 1000$ and a number of input variables determining the decision at each node, $M_{try} = \log_2 N_F + 1$, as suggested by [8]. Both classifiers have been tuned, during the training process, to optimize the accuracy score, following the standard methodology.

It is worth noticing that in a detection problem, $S = TP/(TP + FN)$ can be regarded as the most relevant parameter, since it expresses the *true positive rate* (the proportion of actual bifurcation samples which are correctly identified as such). We can therefore consider S as the parameter to maximize in the branching detection problem. The precision $P = TP/(TP + FP)$ (*positive predictive value*) is another relevant parameter, representing the proportion of samples assigned to the bifurcation class which are correctly classified.

The results of the automatic classification are presented in Table 1, together with the inter-observer variability. As it can be observed, for the inter-observer variability the sensitivity is lower than the other parameters, thus demonstrating that even the manual bifurcation location performed by expert physicians is

challenging. The results of the AdaBoost automatic classification are superior to manual classification in terms of sensitivity, showing that the algorithm reaches a compromise between the two labeled ground-truths. The Random Forest classifier is ahead of AdaBoost for almost all the considered parameters, with the exception of S, while being comparable with the manual annotation.

In order to corroborate the quality of the achieved results, we perform a statistical analysis. For each pair of approaches, the Wilcoxon signed-ranks test [9] is applied, with a *p-value* $\alpha = 0.05$. Table 2 illustrates, for every comparison, the difference between the mean values of the performance parameters; the asterisk denotes that the difference between the results is statistically significant. The statistical analysis shows that the sensitivity of the AdaBoost classifier is significantly better than both Random Forest and inter-observer variability scores. Moreover, the AdaBoost precision, though lower in mean, is not significantly different from the inter-observer variability. AdaBoost gives a higher false alarm ratio score (+3.55%) than Random Forest, but this drawback is compensated by the high gap in sensitivity (+12.44%). Considering these factors, the AdaBoost classifier turns out as the most appropriate technique for this specific task.

Table 1. Performance of the automatic classifiers and inter-observer variability

$LOPO$	$AdaBoost$	$Random\ Forest$	$Inter-observer$
A	$(94 \pm 4.5)\%$	$(96.78 \pm 3.32)\%$	$(\mathbf{98.77} \pm 0.76)\%$
S	$(\mathbf{75.09} \pm 13.7)\%$	$(62.65 \pm 19.16)\%$	$(61.87 \pm 11.27)\%$
K	$(93.51 \pm 4.71)\%$	$(97.06 \pm 3.46)\%$	$(\mathbf{99.38} \pm 0.39)\%$
P	$(92.56 \pm 3.8)\%$	$(95.96 \pm 3.24)\%$	$(\mathbf{99.05} \pm 0.76)\%$
FAR	$(6.49 \pm 4.71)\%$	$(2.94 \pm 3.46)\%$	$(\mathbf{0.62} \pm 0.39)\%$

Table 2. Difference between the mean performance values $\Delta(mean)$ and statistical significance of the difference assessed with the Wilcoxon signed-ranks test, for the three pairs of approaches

	$AD.\ vs.\ R.F.$	$AD.\ vs.\ I.O.$	$R.F.\ vs.\ I.O.$
A	$-2.78\% *$	$-4.77\% *$	$-1.99\% *$
S	$+12.44\% *$	$+13.22\% *$	$+0.78\%$
K	$-3.55\% *$	$-5.87\% *$	$-2.32\% *$
P	$-3.4\% *$	-6.49%	$-3.09\% *$
FAR	$+3.55\% *$	$+5.87\% *$	$+2.32\% *$

Figs. 5-d, -e, -f, -g illustrate bifurcation frames corresponding to the sequence longitudinal positions highlighted in Fig. 5-a. Fig. 5-c shows a map reporting the pullback classified by using AdaBoost, while Fig. 5-h represents the margin values, produced as an output of the AdaBoost classifier for each classified sample. The margin value indicates how likely a sample is to belong to a class; for this reason, it can be interpreted as an estimate of the probability of bifurcation presence.

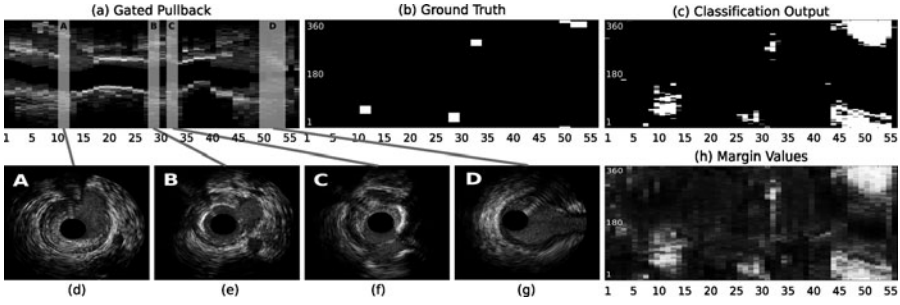


Fig. 5. Pullback after motion artifact compensation (a), ground-truth (b), classification map (c) and map of pseudo-probability of bifurcation presence (h). The maps in (b), (c), (h) represent, on the horizontal and vertical axes, the longitudinal and angular positions along the pullback respectively. In (b) and (c) the white and black colors indicate where bifurcation and non-bifurcation samples (angular sectors) are present, while in (h) pixel intensity represents the probability of bifurcation presence. The frames in (d), (e), (f), (g) correspond to the four bifurcations.

4 Discussion and Conclusion

In this paper, a fully automatic method for the identification of the angular and longitudinal bifurcation position in IVUS sequences is presented. To our knowledge, we are the first to apply bifurcation detection on IVUS images. The novelty of our approach lies in the computation of a set of statistical features on the angular sectors, instead than on the pixels. The task presents a considerable difficulty, due to the high variability of bifurcation dimensions and appearance in IVUS images. Portions of the vessel can often appear like bifurcations, especially if the blood region is not entirely visible in the image; moreover, bifurcations can be hidden by shadows or change significantly their characteristics in correspondence of implanted stents. Nevertheless, the method shows encouraging results. The current method does not use any kind of feature selection strategy, but future work could deal with a feature selection study, which would reduce the computational cost of the methodology and may improve the results. Since AdaBoost proves to be the most suitable classifier for this problem, the margin value produced as an output can be additionally used to refine the detection results. The spatio-temporal continuity of the bifurcation regions could be considered, by exploiting the neighborhood properties of the angular sector samples. Finally, the feasibility of a study on the estimation of the *angle of incidence* (the angle between the main vessel and the side-branch) will be investigated.

Acknowledgment

This work has been supported in part by projects TIN2009-14404-C02, La Marató de TV3 082131 and CONSOLIDER-INGENIO CSD 2007-00018.

References

1. Zarins, C.K., Giddens, D.P., Bharadvaj, B.K., Sottiurai, V.S., Mabon, R.F., Glagov, S.: Carotid bifurcation atherosclerosis. quantitative correlation of plaque localization with flow velocity profiles and wall shear stress. Circulation Research 53, 502–514 (1983)
2. Gatta, C., Balocco, S., Ciompi, F., Hemetsberger, R., Leor, O.R., Radeva, P.: Real-time gating of IVUS sequences based on motion blur analysis: Method and quantitative validation. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 59–67. Springer, Heidelberg (2010)
3. Gatta, C., Pujol, O., Leor, O.R., Ferre, J.M., Radeva, P.: Fast rigid registration of vascular structures in ivus sequences. IEEE Transactions on Information Technology in Biomedicine 13(6), 1006–1011 (2009)
4. Bovik, A., Clark, M., Geisler, W.: Multichannel texture analysis using localized spatial filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 55–73 (1990)
5. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
6. Li, W., van der Steen, A., Lancé, C., Honkoop, J., Gussenhoven, E.J., Bom, N.: Temporal correlation of blood scattering signals in vivo from radio frequency intravascular ultrasound. Ultrasound in Medicine and Biology 22(5), 583–590 (1996)
7. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
8. Statistics, L.B., Breiman, L.: Random forests. Machine Learning, 5–32 (2001)
9. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)

A Region Segmentation Method for Colonoscopy Images Using a Model of Polyp Appearance

Jorge Bernal, Javier Sánchez, and Fernando Vilariño

Computer Vision Center and Computer Science Department, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
jbernal@cvc.uab.es
http://www.cvc.uab.es/

Abstract. This work aims at the segmentation of colonoscopy images into a minimum number of informative regions. Our method performs in a way such, if a polyp is present in the image, it will be exclusively and totally contained in a single region. This result can be used in later stages to classify regions as polyp-containing candidates. The output of the algorithm also defines which regions can be considered as non-informative. The algorithm starts with a high number of initial regions and merges them taking into account the model of polyp appearance obtained from available data. The results show that our segmentations of polyp regions are more accurate than state-of-the-art methods.

Keywords: Colonoscopy, Polyp Detection, Region Merging, Region Segmentation.

1 Introduction

Colon cancer's survival rate depends on the stage that it is detected on, going from rates higher than 95% in stages I or II to rates lower than 35% in stages IV and V [1], elucidating the importance of detecting it on its early stages. In order to do so, several screening techniques are used. One of the most extended techniques is colonoscopy [2], which consists of introducing a probe with a camera mounted on it through the rectum and the colon. The physician can observe the status of the patient as the colonoscope progresses, and it can even remove polyps during the intervention.

Our global objective is to detect polyps in colonoscopy images. Our processing scheme [3] consists of 3 stages: *region segmentation*, *region description* and *region classification*. This detection scheme can be used in several applications such as: 1) real-time polyp detection, 2) off-line quality assessment of the colonoscopy, and 3) quantitative assessment of the trainee skills in training procedures, just to mention a few.

In this paper we present our *region segmentation* stage in which an input colonoscopy image is segmented into a minimum number of informative regions, one of them containing a polyp, therefore reducing the size of the problem. The term of informative regions is used here as opposite to non-informative

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 134–142, 2011.

regions, where we are sure no polyp is inside and therefore, there will be no need to continue analyzing them [4]. These results can be used later to classify the informative regions into polyp- vs. non-polyp-containing candidates.

The structure of the paper is as it follows: in Section 2 we present the segmentation algorithm which we will compare our performance results with. In Section 3 we present our segmentation algorithm along with the model of polyp appearance in which it was inspired. In Section 4 we present our experimental setup and show our results. Finally in Section 5 we show the main conclusions that we extract from our work and present some future research lines.

2 Related Work

There are different approaches to polyp detection in colonoscopy video, which can be divided [5] according to the type of feature they are based on, namely shape, color or texture. Some of the include, like us, a *Region Segmentation* stage that also use shape and color cues to guide the process, such as the work of Hwang et al. [6],although they are in a more advanced stage where they classify the segmented regions.

In general segmentation, which is one of the most difficult and critical tasks in computer vision, can be viewed as a perceptual grouping problem in which the image is divided into homogeneous regions, which can represent different features in the images depending on the methodology adopted. Some simple ways of segmentation exist however they prove to be over simplified for semantic grouping of image regions in more complex scenarios, as they are more sensitive to noise and other artifacts [7]. More sophisticated methods of image segmentation can be mainly divided into two different categories: segmentation by fitting and segmentation by clustering [8]. In the former, the problem of segmentation is viewed as an assertion that the pixels in an image conform to a model while, in the latter, the pixels are grouped according to some criteria such as gray level, color or texture. In order to perform efficiently, segmentation by fitting methods need strong gradient differences pertaining to the objects in the images which have to be segmented, which is not our case. Given that we want to segment informative regions containing polyps from clinically uninteresting areas, methods that segment by clustering seem well suited for our scenarios. Because of this, we have chosen two methods from this group to carry out our research:

a) Normalized Cuts: The *normalized cuts* method [9] is a graph theoretic approach for solving the perceptual grouping problem in vision. In *normalized cuts*, all the sets of points lying in the feature space are represented as a weighted, undirected graph. The weight of each arc is assigned using a set of pre-defined criteria. These can be based on the spatial distance among the pixels, their brightness values, etc. Usually the easiest way to perform segmentation in graph theoretic algorithms is to disconnect the edges having small weights usually known as the *minimum cut* [10]. The problem with minimum cuts is that it typically results in over segmentation since the method basically finds local minima. Shi and Malik [9] proposed in 2000 a new approach that aims at extracting the

global impression of an image instead of focusing on its local features. In this approach (*normalized cuts*), the cut between two graphs is normalized by the volumes of the resulting graphs. In this case graphs are constructed by taking each pixel as a node and by defining the edge weight as a the product of feature similarity term and spatial proximity term [9].
b) Watersheds: Watershed transformation [11] is one of the clustering based methods used as a tool for image segmentation. Watersheds operate on intensity gradients to perceive an image as a combination of catchment basins in a hilly area (a hill corresponds to high gradient) simulating the formation of image regions with projected flow of water. After identification of an intensity valley in an image, region growing algorithms are used to combine all the pixels which have similar intensities. This procedure is particularly effective in images where we have strong boundaries of objects which have to be segmented. For images which are rich in texture or where the intensity gradient is not prominent, an over segmentation is usually obtained, making convenient a posterior region merging.

3 Our Proposed Methodology

3.1 A Model of Polyp Appearance

The lighting of the probe gives us hints about how polyps appear in colonoscopy images. As the light falls perpendicularly to the walls of the colon, it creates shadows around the surfaces and, when the light falls into a prominent surface (Figure 1 (a)), it creates a bright spot surrounded by darker areas, which are the shadows, generating edges and valleys in the intensity image (Figure 1 (b)).

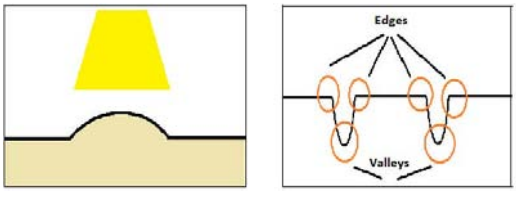


Fig. 1. Simulation of (a) an illuminated prominent surface (b) grey-scale profile

Even considering this model of prominent surface appearance under the lighting conditions of the colonoscope, there are some challenges to be overcome, namely: 1) non-uniform appearance of the polyps (flat or peduncular shapes); 2) the appearance of reflections in the image; and 3) the similarity between tissues inside and outside the polyp, also affected by the lighting conditions. Taking all this into consideration, we base our segmentation method on a model of polyp appearance that we can define as *a prominent shape enclosed in a region with presence of edges and valleys around its frontiers.*

3.2 Algorithm

The segmentation algorithm, which general scheme can be seen in Figure 2, consists of 4 different stages which will be described next.

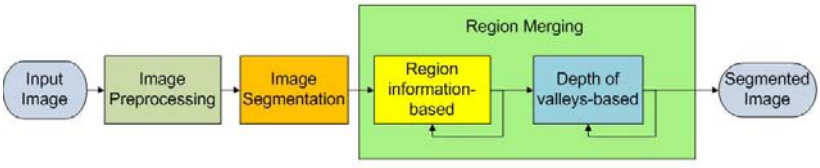
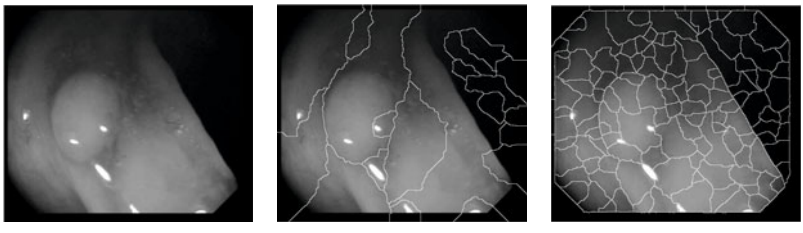


Fig. 2. Scheme of the segmentation algorithm

1. **Image Preprocessing:** Before applying any segmentation algorithm there are some operations that should be done: 1) converting the image into gray-scale, 2) de-interleaving (as our images come from a high definition interleaved video source), 3) correction of the reflections, and 4) inverting the grey-scale image.
2. **Image Segmentation:** We apply watersheds to the gradient image because the boundaries between the regions obtained in such way are closer to the boundaries that separate the different structures that appear in the image (Fig. 3 [4]).



(a) (b) (c)

Fig. 3. Use of gradient information: (a) original image (b) basic segmentation (c) Segmentation using gradient information

3. **Region Merging:**
   a) Region information-based: We first calculate the neighborhood map of the image and identify the frontier pixels between each pair of regions and then categorize the regions and frontiers, in terms of the amount of information that they contain [4]. For instance, a low information region will have a very high mean (or very low) grey level and very low standard deviation of this grey level. We will only merge regions with the same kind of information separated by weak frontiers. In this case, in order to consider a frontier as weak we propose a frontier weakness measure as defined in Equation 1. This measure combines the information of the the mean gradient of the frontier

pixels and the strength of the frontiers (measured as the frontiers that are kept after applying a two consecutive order increasing median filtering, which helps to remove the regions created by the veins). We merge and categorize regions until their number is stabilized or there are no weak frontiers left.

$$FrontierWeakness = \alpha * gradient + \beta * median \quad (1)$$

b) Depth of valleys-based: We define a *depth of valleys* measure that combines the information of the output of a ridges and valleys detector (see [12] for details) with the information that the morphological gradient provides. This gives information about the depth of the pixels in the valley with higher values for the pixels that constitute the frontier of the region (which will have both high 'valleyness' and gradient values and smaller from the inner pixels, as can be seen in Figure 4). Using this information we can continue merging regions, keeping only those which frontiers are strong in terms of *depth of valleys*. We merge regions until there are no weak frontiers according to the depth of valleys threshold value or when the number of regions is stabilized.

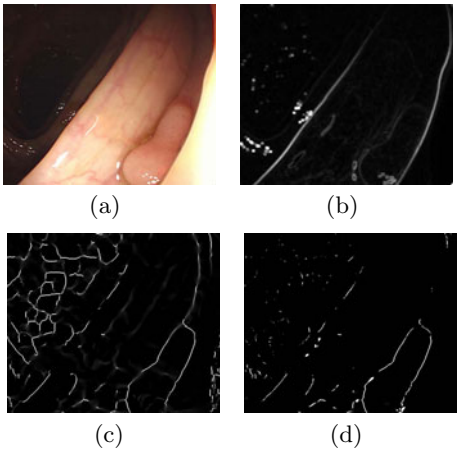


Fig. 4. Creation of the depth of valleys image: (a) Original image (b) Morphological Gradient image (c) Valleys image (d) Depth of valleys image

4 Results

4.1 Experimental Setup

In order to test the performance of our segmentation algorithm we have created a database which consists of 300 different studies of polyp appearance along with their corresponding polyp masks. We will evaluate the performance of our method by using two different measures: Annotated Area Covered (AAC) and Dice Similarity Coefficient (DICE) [7].

Table 1. Comparison between the results obtained by our method and normalized cuts with respect to the value of the depth of valleys

With Borders						
Measure / Method	Ours	NCuts	Ours	NCuts	Ours	NCuts
Threshold Value	0.6		0.7		0.8	
AAC	61.91%	63.66%	70.29%	69.06%	75.79%	70.86%
DICE	55.33%	44.97%	44.6%	37.75%	36.44%	34.01%
Without Borders						
Measure / Method	Ours	NCuts	Ours	NCuts	Ours	NCuts
Threshold Value	0.6		0.7		0.8	
AAC	60.71%	60.2%	70.29%	63.98%	74.32%	64.24%
DICE	55.68%	63.15%	48.01%	61.84%	45.01%	56.73%

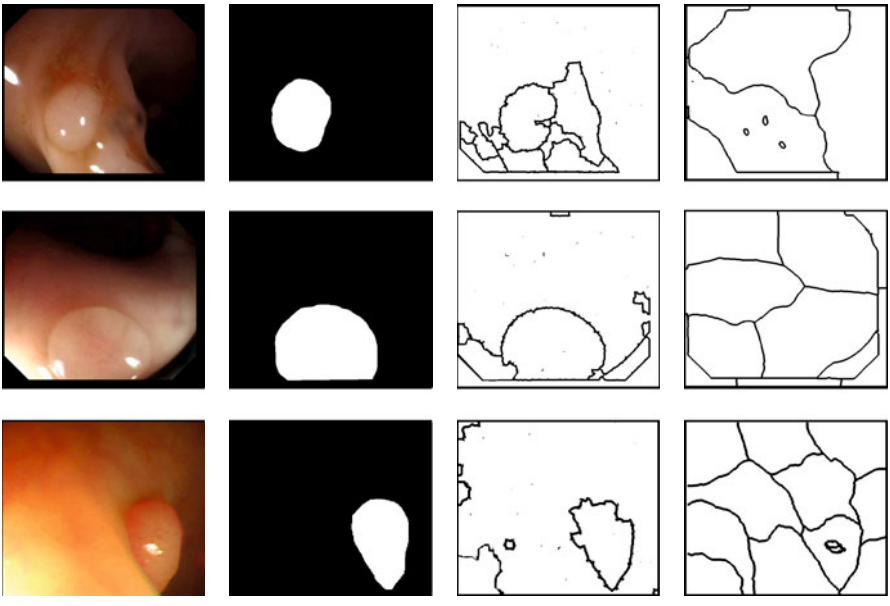
Both measures are complementary as the former calculates the amount of annotated polyp area while the latter complements it with the amount of non-polyp information that is kept in the region. We will compare our final segmentation results with the ones obtained using *normalized cuts*. To do so we will set the number of final regions that we have to obtain with *normalized cuts* at the minimum number of final regions that we have obtained with our method that gives us best results in terms of AAC and DICE. To obtain our best results we have run experiments in order to get the combination of parameter values (α, β, depth of valleys threshold) that gives good results for every image.

We also have to consider that our images have black borders around them. Our region segmentation method consider their presence and use the results of non-informative region identification (which borders of the image are part of) to avoid further processing of this areas. In order to test the effect that these borders have in the segmentation results, we have also created a database that eliminates the borders of the images. It has to be mentioned that in order to make the new image suitable to be processed some non-borders parts of the image should also be eliminated which causes the loss of some information. We will compare the results achieved by using these two versions of the images.

4.2 Experimental Results

In Table 1 we show results for polyp region detection comparing the performance of our method, with the performance achieved by *normalized cuts* (with the same number of final regions). Using the whole image we get better results than *normalized cuts* in terms of AAC and DICE. This means that our regions which contain polyps have more polyp information than the ones that *normalized cuts* and it is closer to the real polyp region.

The elimination of the borders in the image, in terms of AAC, has almost no effect for our method but it has more incidence for *normalized cuts*. DICE results improve for both methods by using the image without borders (better as the threshold value increases), although *normalized cuts* results are better. But, as it can be seen in Figure 5, *normalized cuts* non-polyp regions tend



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

Fig. 5. Comparison of segmentation results: (a) Original images (b) Polyp masks (c) Our method's output (d) Normalized cuts' output

to be larger than our non-polyp regions (in our case we know that the larger region corresponds always to the background). In our case, we could apply a size threshold value to discard some of them without having chance of losing the polyp region while in *normalized cuts* this would not happen.

In Figure 5 we can see examples of each method's output. It can be seen that the images segmented with our method (see Figure 5(c)) fit better the polyp mask (that was segmented by experts). Third row's results are obtained using the image without borders. We plan to improve our DICE results by merging some small regions that appear inside the polyp region and, after this is achieved, our overall region segmentation results by discarding some of the smallest regions in order to provide as result a very low number of relevant regions.

5 Conclusions and Future Work

In this paper, in the context of region segmentation in colonoscopy images, we present our novel segmentation approach. Our objective is to provide a low number of regions, one of them containing a polyp. Our algorithm also indicates the degree of information of the final regions. Our approach consists of applying a basic segmentation algorithm (such as watersheds) and then applying a region merging algorithm that takes into account our model of polyp appearance. This model states that a polyp is a prominent shape enclosed in a region with presence

of edges and valleys around its frontiers. In order to test the performance of our method we rely on a database of more than 300 studies where the experts have manually segmented the polyps in the images. We compare our method with one state-of-the-art technique (*normalized cuts*) and quantify the accuracy of our segmented regions using two complementary measures. Our method outperforms normalized cuts in the accuracy of polyp region detection and also offers a good non-informative region characterization.

Our future work, in terms of *Region Segmentation*, will be focused on reducing even more the number of final regions and once we have achieved a better performance (in terms of both AAC and DICE) we plan to start with the *region description* and *region classification* stages.

Acknowledgements

The authors would like to thank Farhan Riaz for his helpful comments and suggestions. This work was supported in part by a research grant from Universitat Autonòma de Barcelona 471-01-3/08, by the Spanish Government through the founded project "COLON-QA" (TIN2009-10435) and by research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

References

1. Tresca, A.: The Stages of Colon and Rectal Cancer. New York Times (About.com), p. 1 (2010)
2. Hassinger, J.P., Holubar, S.D., et al.: Effectiveness of a Multimedia-Based Educational Intervention for Improving Colon Cancer Literacy in Screening Colonoscopy Patients. Diseases of the Colon & Rectum 53(9), 1301 (2010)
3. Bernal, J., Sánchez, J., Vilariño, F.: Current challenges on polyp detection in colonoscopy videos: From region segmentation to region classification. a pattern recognition-based approach. In: Proceedings of the 2nd International Workshop on Medical Image Analysis and Description for Diagnosis Systems - MIAD 2011, Rome, Italy (January 2011) (in press)
4. Bernal, J., Sánchez, J., Vilariño, F.: Reduction of Pattern Search Area in Colonoscopy Images by Merging Non-Informative Regions. In: Proceedings of the XXVIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica, Madrid, Spain (November 2010) (in press)
5. Ameling, S., Wirth, S., Paulus, D.: Methods for Polyp Detection in Colonoscopy Videos: A Review. Inst. für Computervisualistik (2009)
6. Hwang, S., Oh, J., Tavanapong, W., Wong, J., De Groen, P.: Automatic polyp region segmentation for colonoscopy images using watershed algorithm and ellipse segmentation. Progress in biomedical optics and imaging 8(33) (2007)
7. Riaz, F., Ribeiro, M.D., Coimbra, M.T.: Quantitative comparison of segmentation methods for in-body images. In: Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, EMBC 2009, pp. 5785–5788 (2009)

8. Forsyth, D.A., Ponce, J.: Computer vision: a modern approach. Prentice Hall Professional Technical Reference (2002)
9. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2002)
10. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(8), 790–799 (2002)
11. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(6), 583–598 (1991)
12. López, A.M., Lumbreras, F., et al.: Evaluation of methods for ridge and valley detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(4), 327–335 (1999)

Interactive Labeling of WCE Images

Michal Drozdzal[1,2], Santi Seguí[1,2], Carolina Malagelada[3], Fernando Azpiroz[3], Jordi Vitrià[1,2], and Petia Radeva[1,2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain
[2] Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Barcelona, Spain
[3] Hospital de Vall d'Hebron, Barcelona, Spain

Abstract. A high quality labeled training set is necessary for any supervised machine learning algorithm. Labeling of the data can be a very expensive process, specially while dealing with data of high variability and complexity. A good example of such data are the videos from Wireless Capsule Endoscopy. Building a representative WCE data set means many videos to be labeled by an expert. The problem that occurs is the data diversity, in the space of the features, from different WCE studies. That means that when new data arrives it is highly probable that it will not be represented in the training set, thus getting a high probability of performing an error when applying machine learning schemes. In this paper an interactive labeling scheme that allows reducing expert effort in the labeling process is presented. It is shown that the number of human interventions can be significantly reduced. The proposed system allows the annotation of informative/non-informative frames of the WCE video with less than 100 clicks.

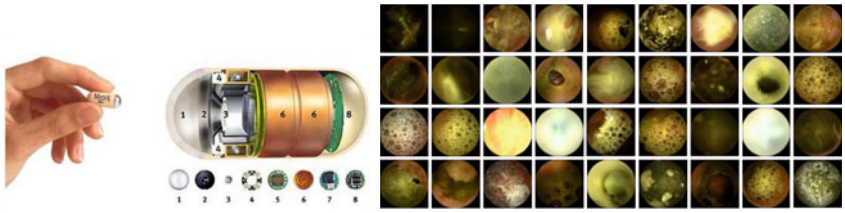
Keywords: WCE, interactive labeling, online learning, LSH.

1 Introduction

For many machine learning applications, the compilation of a complete training set for building a realistic model of a given class of samples is not an easy task. The motto "*there is no data like more data*" suggests that the best strategy for building this training set is to collect as much data as possible. But in some cases, when dealing with problems of high complexity and variability, the size of this data set can grow very rapidly, making the learning process tedious and time consuming. Time is expended in two different processes: 1) the labeling process, which generally needs human intervention, and 2) the training process, which in some cases exponentially increments computational resources as more data is obtained.

If we consider training as a sequential (in time) process, the training problem can be partially overcome by the integration of online learning methodologies and an "intelligent" reduction of the samples to be added into the training set. Given a training set at a given time, an "intelligent" system should add to the training set only those data samples that were not represented, or underrepresented, in the previous version of the set. In other words, the training set

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 143–150, 2011.



(a) (b)

Fig. 1. (a) The wireless video capsule; (b) Non informative frames

should be enlarged by those new data samples that enrich the representability of the classification models while avoiding unnecessary sample redundance.

But even in the former situation, the labeling process is still a problem. To overcome this problem, we should use a labeling process which minimizes the number of human interventions on the new samples. One possible strategy is to embed a model of the data into a sequential labeling process, for proposing to the human operator a given label for every sample. Then, the human operator faces two possible decisions: to accept the model proposal or to change the label of the sample. In practice, these two choices have a non symmetric cost for the human operator: accepting the model proposal can be efficiently implemented with a low cognitive load for the operator, while changing a label has a larger cost. This effort can be measured by the number of "clicks" the operator performs.

Wireless Capsule Endoscopy (WCE) image analysis (see Fig. 1(a)) is a clear scenario where these problems arise. The WCE contains a camera and a full electronic set which allows the radio frequency emission of a video movie in real time. This video, showing the whole trip of the pill along the intestinal tract, is stored into an external device which is carried by the patient. These videos can have duration from 1h to 8h, what means that the capsule captures a total of 7.200 to 60.000 images. WCE videos have been used in the framework of computer-aided systems to differentiate diverse parts of the intestinal tract[4], to measure several intestinal disfunctions [11] and to detect different organic lesions (such as polyps [6], bleeding [5] or general pathologies [2]). In most of these applications, machine learning plays a central role to define robust methods for frame classification and pattern detection.

A common stage in all these research lines is the discrimination of informative frames from non-informative frames. Non-informative frames are defined as frames where the field of view is occluded. Mainly, the occlusion is caused by the presence of intestinal content, such as food in digestion, intestinal juices or bubbles (see Fig. 1(b)). The ability of finding non-informative frames is important since: 1) generally, it helps to reduce time of video analysis, and 2) since the majority of non-informative frames are frames with intestinal content which information can be used as an indicator for intestinal disfunctions [8].

The main strategy in the search of non-informative frames is the application of machine learning techniques in order to build a two-class classifier. Generally, non-informative frames are characterized by their color information [1]. Robust

classifiers can be built when the training set is representative of the data population. The problem that occurs when dealing with WCE videos is the high color variability of non-informative frames in different videos. Therefore, it is probable that, as new videos become available, a new video with significantly different intestinal content color distribution will be added to the training set. A naive approach for the labeling process of this video could mean the manual annotation of up to 50.000 video frames.

The main contribution of this paper is a method for interactive labeling that is designed to optimally reduce the number of human interventions during the labeling process of a new video. The process is applied to the labeling of non-informative frames for WCE image analysis. In the active learning algorithm the key idea is that machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which is learns [10,7]. However, the goal of our system is basically to enlarge the training set while minimizing human intervention, not to minimize the overall classification error. This is done by finding an optimal classifier adaptation scheme for non-represented frames in the original training set.

The rest of the paper is organized as follows: Section 2 introduces the interactive labeling method, Section 3 describes the implemented interactive labeling system for WCE images. In Section 4, we present experimental results, and finally, in Section 5 we expose some conclusions and remarks on future research.

2 Interactive Labeling

Our purpose is to design an interactive labeling system that should allow, in an efficient way, 1) to detect frames that are not represented in the training set, 2) to obtain statistics of informative and non-informative frames that are not represented in the training set, and 3) to be able to iteratively increase the representability of the training set in an "intelligent" way by reducing significantly the number of clicks relates to manual labeling.

To this aim, we propose the following algorithm for interactive labeling of a set of new images optimizing the user feedback (see Alg. 1).

The critical step in order to minimize the number of clicks is to choose a good criterion for Step 6, since it represents the main strategy of choosing the order of presentation of the samples to be labeled by the user.

To this end, we studied three different sorting policies for the elements of N. These policies are based on the following criteria: 1) to choose those elements that are far from the training data L and far from the boundary defined by M_i, 2) to choose those elements that belong to the most dense regions of N, and 3) to choose the elements in a random way.

More specifically, we define them in the following way:

Criterion 1 (C1)*: Distance of data to the classifier boundary and training data.* In this criterion two measurements are combined: 1)The data are sorted from the farthest to the nearest distance with respect to the classifier boundary. This scheme assumes that the classifier, while proposing labels, will commit

Algorithm 1. Interactive labeling algorithm.

1: Let be L a set of labeled data, M_1 a discriminative model trained on this set, U a set of all unlabeled data samples from a new video and C a criterion to select the most informative frames from U;
2: Select the subset of samples $N = \{x_j^N\}$ from U such that they are considered as under-represented by the labeled set L.
3: Evaluate the subset N with M_1, assigning a label l_j to every data sample x_j from N.
4: i=1;
5: while there are elements in N do
6: Evaluate the elements of N with respect to the criterion C and get the set of n most informative samples $I \subset N$ (with the purpose of minimizing the expected number of expert clicks).
7: Delete the elements of I from N.
8: Present the samples from I to the user with their associated label.
9: Get the user feedback (nothing for samples with correct labels, one click for each wrongly classified sample).
10: Update L by adding the elements of I and its user-corrected label.
11: Perform an online training step for M_i by adding the elements of I to the model, getting M_{i+1}.
12: Evaluate the elements of N with M_{i+1}, assigning a label l_j to every data sample x_j from N.
13: $i = i + 1$;
14: end while

errors with higher probability for the samples that are far from the boundary than for the data that are relatively close to boundary. 2) The data are sorted from the farthest to the nearest with respect to the training data. This scheme assumes that the classifier, while proposing labels, will commit errors with higher probability for the samples that are far from the training set than for the data that are relatively close to the known data. A final sorting is performed in the data by adding the ranking indices of the two previously described schemes.

Criterion 2 (C2)*: Data density.* Each sample is sorted decreasingly with respect to a data density measure in its environments. This scheme assumes that the classifier should learn more quickly if we first provide samples from the zones with higher density. Data density can easily be computed as the mean distance to the k-nearest neighbors of the sample.

Criterion 3 (C3)*: Random order.* The order of presentation of the samples randomly determined.

3 The Interactive Labeling System

The goal of the interactive labeling system is two-fold: 1) to detect, for each new video, the set of frames that are not represented in the training set, and 2) to label those frames with minimal user effort. To this end, we propose a system design with two main components(see Fig. 3):

1. A data density estimation method that allows fast local estimation of the density and distance of a data sample to other examples, e.g. from the training set (see Step 3 of the algorithm for interactive labeling).
2. An online discriminative classifier which allows to sequentially update the classification model M_i of thousands of samples (see Step 2, 10 and 11 of the algorithm for interactive labeling).

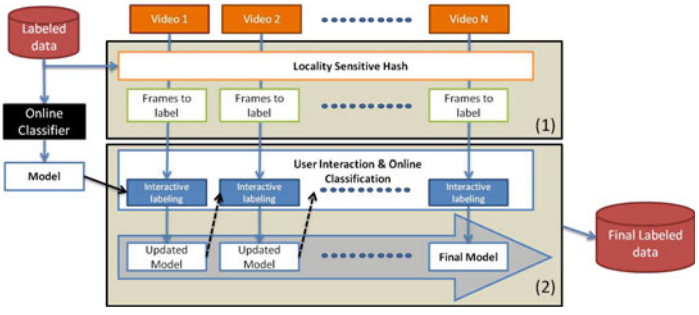


Fig. 2. The interactive labeling system architecture with its two main components: 1) Detection of frames not represented in the training set and 2) Labeling of frames and model enlarging using online classifier method

3.1 Fast Density Estimation

As previously commented, the local density of a data sample x_i with respect to a data set C can be easily estimated by computing the mean distance from x_i to its k-nearest neighbors in C. The simplest solution to this problem is to compute the distance from the sample x_i to every sample in C, keeping track of the "k-best so far". Note that this algorithm has a running time of $O(nd)$ where n is the cardinality of C and d is the dimensionality of samples.

Because of excessive computational complexity of this method for large data sets, we need a flexible method that allows from one side, effective measurements of characteristics of large data and, from the other side, introducing new unseen data into the training set for enlarging the data representation. An example of such flexible method is Locality Sensitive Hashing[3]. LSH allows to quickly find a similar sample in a large data set. The basic idea of the method is to insert similar samples into a bucket of a hash table. As each hash table is created using random projections over the space, several tables can be used to ensure an optimal result [3]. Another advantage of LSH is the ability to measure the density of data in given space vicinity by analyzing the number of samples inside the buckets (Fig. 3.1). In order to evaluate if the new sample improves the representation of the data set, the space density of the training set is estimated. If the new sample is in a dense part of the space then the sample is considered redundant and thus it is not considered in order to improve the training set. Otherwise, the sample is used to enlarge the training set.

The density D of the sample x is estimated according to the formula:

$$D(x, T_r) = \sum_{i=1}^{M} ||B_i|| \tag{1}$$

where M is the number of hash tables, $||B_i||$ is the number of elements in the bucket where the new element x is assigned and T_r represents the training set.

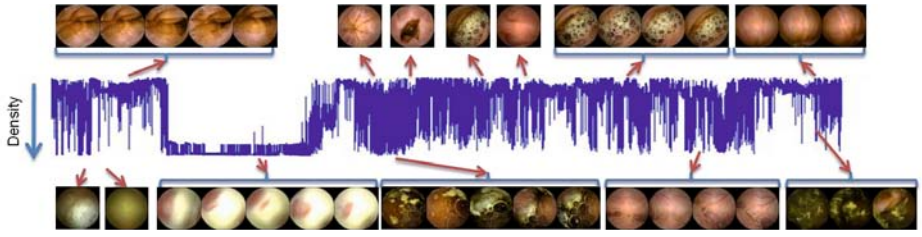


Fig. 3. Example of training set density estimation for a test video using LSH. The images show the zones of high and low density with respect to given labeled set L.

The subset of not-represented samples in the training set $N = \{x_1^*, ..., x_m^*\}$ from new unlabeled data U is defined as:

$$N = \{\forall x \in U : D(x, T_r) < T\} \tag{2}$$

where T represents a fixed threshold.

Note that (2) expresses the condition that the new samples fall in buckets with low density of the training set. That is if the new sample is in a dense part of the space then the sample is not considered to improve the training set. Otherwise, the sample is used to enlarge the training set.

3.2 Online Classifier

Taking into account that our classifier must be retrained with thousands of images/feature vectors of up to 256 components, using an online classifier is a must. Online classifiers are able to update the model in a sequential way, so the classifier, if needed, can constantly learn form new data, improving the quality of label proposal process. In order to optimize the learning process, the data are sorted according to the previously described criteria. A kernel-based online Perceptron classifier [9] is used because of its simplicity and efficiency. As previously mentioned, the main information used to detect non-informative frames is the color. In order to reduce the dimensionality of the data, each image represented as 24 million colors is quantized into 256 colors. As a result, each frame is represented by 256 color histogram. The score, for a given sample, takes this form:

$$S(x) = \sum_{j=1}^{K} \alpha_j K(v_i, x) \tag{3}$$

where $\langle (v_1, \alpha_1), ..., (v_k, \alpha_k) \rangle$ are the set of training vectors with the corresponding real estimated weights $(\alpha_1, ..., \alpha_k)$ by the learning algorithm when minimizing the cumulative hinge-loss suffered over a sequence of examples and $K()$ is a kernel function (in our case, we apply Radial Basis Function).

4 Results

For our experiments, we considered a set of 40 videos obtained using the WCE device. 10 videos were used to built the initial classification model M_1, and the

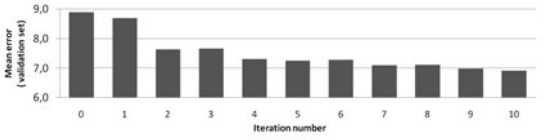


Fig. 4. Mean error on validation set of 20 WCE videos

other 10 to evaluate the proposed interactive labeling system. In the test, the 10 videos were sequentially processed. If needed, at each iteration, the training set could be increased by a new set of frames that improves the data representation. Additionally the validation set of 20 videos were used in order to evaluate the error of the final informative/non-informative frames classifier.

In the experiments we show that: 1) the proposed system reduces the effort needed for data labeling, 2) the first criterion *Criterion 1 (C1): Distance of data to the classifier boundary and training data* gives the best results, 3) the global performance of informative/non-informative frames classifier is improving while enlarging training set, and 4) the LSH optimizes the computation process.

Table 4, shows that all three proposed schemes reduce the number of clicks. Even random order improves a lot with respect to the naive approach. This phenomenon can be explained by the fact that the frame color in a certain video are well correlated. From the results it can be concluded that the criterion 1 looks to be the best sorting criterion for interactive labeling. Intuitively, the samples that are far from the boundary are classified with high confidence. However, when dealing with the frames that are not similar to the ones in the training set (there are far from the one in the training set), the confidence gives uncertainty measure. Therefore, when introducing examples, where the classifier performs an error (user switches the label) to the model it is highly probable that the boundary will change using small amount of data.

Introducing new data into the training set improves the final classifier performance and reduces the error by 2% after 10 iterations of the algorithm (where each iteration is new video introduced in the training set) (Fig. 4). Furthermore, the LSH in average reduces the number of frames to check by more than 80%. This means that tested videos have about 20% of the frames that are "strange". While inserting new frames into the classifier model, it can be seen, that at each

Table 1. Results

Video	#frames	#strange frames	#clicks Criterion_1	Criterion_2	Criterion_3
Video1	35847	4687	103	103	147
Video2	51906	10145	211	213	316
Video3	52777	5771	270	270	376
Video4	56423	13022	86	90	151
Video5	55156	7599	68	68	131
Video6	33590	17160	381	389	617
Video7	17141	1072	8	8	39
Video8	26661	5437	88	97	151
Video9	14767	1006	28	28	76
Video10	22740	1993	63	63	110
Average clicks per video	-	-	1.5%	1.5%	2.9%

iteration some data that are not represented in the training set are being found. The conclusion that can be drawn is that in order to create a good training set for informative/non-informative frame classification, the number of 20 WCE videos is not enough.

5 Conclusions

In this paper a system that minimizes the user effort during the process of constructing a good representation of the WCE data is presented. The methodology is based on two steps: 1) the detection of frames that enrich the training set representation and thus should be labeled, and 2) the interactive labeling system that allows to reduce the user effort, in the labeling process, using an online classifier which sequentially learns and improves the model for the label proposals. The detection of frames that enlarge the data representation has been performed using LSH. The LSH method allows a fast processing for getting efficient results for data density estimation. Three different sorting polices are defined and evaluated for the online classification: 1) *Distance of data to the classier boundary and training data*, 2) *Data density*, and 3) *Random order*. It is shown that by using adapted sorting criteria for the data we can improve the label proposal process and in this way reduce the expert efforts. Finally, we have observed that enlarging the initial training set with the non represented frames from unlabeled videos we achieve an improvement of classification performance.

References

1. Bashar, M., et al.: Automatic detection of informative frames from wireless capsule endoscopy images. Medical Image Analysis 14(3), 449–470 (2010)
2. Coimbra, M., Cunha, J.: MPEG-7 visual descriptors: Contributions for automated feature extraction in capsule endoscopy. IEEE TCSVT 16(5), 628–637 (2006)
3. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proc. of the 25th ICVLDB, VLDB 1999, pp. 518–529 (1999)
4. Igual, L., et al.: Automatic discrimination of duodenum in wireless capsule video endoscopy. In: IFMBE Proceedings, vol. 22, pp. 1536–1539 (2008)
5. Jung, Y., et al.: Active blood detection in a high resolution capsule endoscopy using color spectrum transformation. In: ICBEI, pp. 859–862 (2008)
6. Kang, J., Doraiswami, R.: Real-time image processing system for endoscopic applications. In: IEEE CCECE 2003, vol. 3, pp. 1469–1472 (2003)
7. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. Int. J. Comput. Vision 88, 169–188 (2010)
8. Malagelada, C., et al.: New insight into intestinal motor function via noninvasive endoluminal image analysis. Gastroenterology 135(4), 1155–1162 (2008)
9. Orabona, F., Keshet, J., Caputo, B.: The projectron: a bounded kernel-based perceptron. In: Proc. of the 25th ICML 2008, pp. 720–727 (2008)
10. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
11. Vilarino, F., et al.: Intestinal motility assessment with video capsule endoscopy: Automatic annotation of phasic intestinal contractions. IEEE TMI 29(2), 246–259 (2010)

Automatic and Semi-automatic Analysis of the Extension of Myocardial Infarction in an Experimental Murine Model

Tiago Esteves[1,2], Mariana Valente[2], Diana S. Nascimento[2], Perpétua Pinto-do-Ó[2], and Pedro Quelhas[1,2]

[1] Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[2] INEB - Instituto de Engenharia Biomédica, Rua do Campo Alegre, 823, 4150-180 Porto, Portugal
meb09026@fe.up.pt

Abstract. Rodent models of myocardial infarction (MI) have been extensively used in biomedical research towards the implementation of novel regenerative therapies. Permanent ligation of the left anterior descending (LAD) coronary artery is a commonly used method for inducing MI both in rat and mouse. Post-mortem evaluation of the heart, particularly the MI extension assessment performed on histological sections, is a critical parameter for this experimental setting. MI extension, which is defined as the percentage of the left ventricle affected by the coronary occlusion, has to be estimated by identifying the infarcted- and the normal-tissue in each section. However, because it is a manual procedure it is time-consuming, arduous and prone to bias. Herein, we introduce semi-automatic and automatic approaches to perform segmentation which is then used to obtain the infarct extension measurement. Experimental validation is performed comparing the proposed approaches with manual annotation and a total error not exceeding 8% is reported in all cases.

Keywords: Infarct extension evaluation, image segmentation, region growing, otsu, k-means, meanshift, watershed.

1 Introduction

Acute myocardial infarction is a major public health problem, resulting mainly from the occlusion of coronary arteries, due to the build-up of arteriosclerotic plaques, and the establishment of tissue ischemia eventually leading to end-stage heart failure. Permanent ligation of the left anterior descending (LAD) coronary artery in animal models, including the rat and the mouse, is a commonly used method for reproducing several of the human-associated pathological events. This surgical procedure also allows the implementation of pre-clinical models of disease which are a pre-requisite for testing cell/drug-therapies before proceeding into clinical trials [1]. The tissue extension of the induced myocardial infarction, which is defined as the percentage of the left ventricle affected by

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 151–158, 2011.

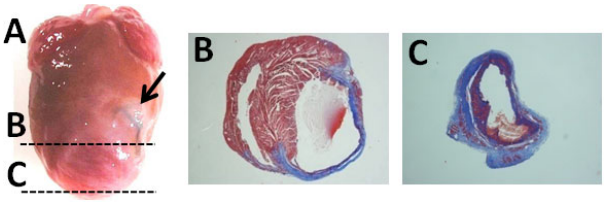


Fig. 1. Experimental myocardial-infarction mouse model. A - Macroscopic view of 21 days post-infarction heart; black arrow indicates the anatomical location of the LAD coronary artery ligation. B and C - Histological cross-sections of apical and mid region of LV stained with Masson Trichrome. Apex and free LV wall are fully compromised by ischemia, wich is illustrated by the collagen deposition (blue region) replacing the viable myocardium tissue (red region).

coronary occlusion, is a critical parameter to evaluate the effect of any applied therapy at the experimental setting. This is calculated as the average value of infarct extension over all cross-sections of the dissected heart stained with Masson's Trichrome, a histological stain that enables the identification of collagen deposition, a hallmark of established infarction [1,2]. To determine the infarct extension it is necessary to indentify the infarcted-tissue (blue area) and the normal-tissue (red area) in each section (Figure 1). Currently these tasks are performed manually by the biologists, which is a time-consuming and arduous endeavor. The latter is a driving force to the development of approaches to aid the analysis of the experimental MI extension. Our approaches entail the segmentation of the cross sections of the heart, which can be performed by means of automated image processing techniques.

The multiple techniques that may be applied to the segmentation of animal tissue can be discriminated in two major classes: automatic and semi-automatic techniques. In the former case the user needs to define initial parameters for each image in order to start the segmentation. Thus, automatic segmentation requires only the validation of the initial parameters and then the algorithms segment all the images in study without further user intervention.

Region growing is a semi-automatic technique that can be used to segment the cross sections of the heart. Alattar et al. describe the use of this technique in segmentation of the left ventricle in cardiac MRI (magnetic resonance imaging) scans [3]. This technique exploits spatial context by grouping pixels or subregions into larger regions. Homogeneity is the main criterion for merging the regions. However, the selection of similarity criteria used depends on the problem under consideration and also on the type of image data available [4,5].

Regarding automatic segmentation there are techniques such as thresholding, region based segmentation and cluster based segmentation that can also be used in tissue segmentation [4,6]. Sharma et al. introduce the segmentation of CT (computurized tomography) abdomen images using a threshold segmentation technique to separate different regions in the images [6]. In a thresholding technique a single value (threshold) is used to create a binary partition of the image intensities. All intensities greater than the threshold are grouped together into one class

and those below the threshold are grouped into a separate class [4,7]. Watershed is also a method applied in medical images segmentation. It is a region based segmentation which involves the concept of topography and hydrography. Hamarneh et al. present MR (magnetic resonance) cardiac images segmented with watershed transform [8]. Watershed can be described as a flooding simulation. Watersheds, or crest lines, are built when the water rise from different minima. All pixels associated with the same catchment basin are assigned to the same label [8,9]. For image segmentation, the watershed is usually, but not always, applied to a gradient image. Since real digitized images present many regional minima in their gradients, this typically results in an excessive number of catchment basins (oversegmentation) [5,9]. Ahmed et al. describe the segmentation of MR brain images using k-means clustering algorithm [10]. K-means segments the entire image into several clusters according to some measure of dissimilarity [8,10]. Mean-shift technique has also been used in segmentation of MR brain images [11]. The mean-shift algorithm is a clustering technique which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters, requiring only the definition of the radius of the kernel used [9].

We use these techniques to (1) segment all histology processed cross-sections of the excised mouse-hearts, (2) calculate the infarct extension and finalize by (3) comparing the results with manual annotation.

This paper is organized as follows: Section 2 introduces the methodology and describes automatic and semi-automatic techniques used in segmentation of the heart, Section 3 defines how to measure the infarct extension Section 4 presents the results obtained and finally the conclusion is presented in Section 5.

2 Methodology

To obtain the infarct extension it is necessary to segment the different tissues in each cross section of the heart. This can be performed with semi-automatic and full automatic techniques. Within the existing semi-automatic methods for image segmentation we had chosen to use region growing due to its speed and ease of interaction. Otsu thresholding technique, watershed segmentation, k-means and mean-shift clustering are the fully automatic techniques that we selected to segment the cross sections of the heart.

In order to improve the segmentation process we applied noise reduction. For this task we tested the method BM3D [12] and the Gaussian filter [5]. The results showed that there were no significant differences in the final segmentation between images filtered with either methods. This lead us to choose the Gaussian filter since BM3D filtering is considerably slower. Noise reduction is applied to all images prior to segmentation.

2.1 Semi-automatic Tissue Segmentation

Region growing exploits spatial context by grouping pixels or sub-regions into larger regions according to some criterion. The average gray level information is the criterion chosen for merging the neighboring regions in our work. Regions

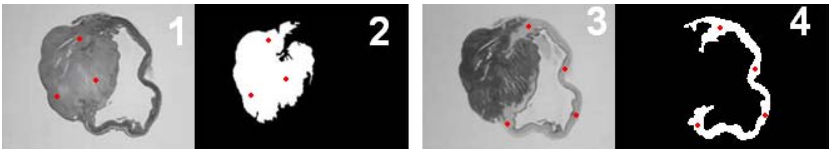


Fig. 2. Segmentation of normal tissue using the Red channel (1) and infarcted tissue using the Blue channel (3) in a cross section of the heart by region growing technique. The results of the segmentation process are binary images (2 and 4). The red points indicate the initial positions of the region growing process.

are merged if they satisfy the chosen criterion and no merging occurs when the criterion is not met [5,4]. The user needs to specify the initial points to begin the segmentation process. For the task of segmenting the normal and the infarcted heart tissue it is necessary to define the initial points for the process of region growing in each of the tissue-conditions. To segment the normal-tissue we used the gray level information present in the Red channel. For the segmentation of the infarcted-tissue we used the gray level information from the Blue channel. The result is a set of binary images, one for each tissue condition (Figure 2). Results are improved using morphological operations, for example to fill small holes inside the segmentation results.

Given the segmentation areas we can then calculate the infarct extension.

2.2 Automatic Tissue Segmentation

To automatically segment the different tissue-conditions in each of the heart cross-sections we use otsu thresholding, watershed segmentation, k-means and mean-shift clustering. All these image segmentation techniques allow the partition of the image in regions which we can associate to the distinct tissue-conditions by analyzing their color.

Otsu thresholding technique selects an adequate threshold of gray level for extracting objects from their background. This threshold is used to convert an intensity image to a binary one. All intensities greater than the threshold are grouped together into one class and those below the threshold are grouped into a separate class [13].

Using this technique, with different channels of the RGB image, we can obtain segmentations of the normal and infarcted-tissue. High values of image intensities in the Red channel relate to normal-tissue. The Blue channel has high image intensity values in infarcted areas. Based on these relationships between the Red and Blue color channels and the tissue properties we decided to subtract the Blue channel to the Red channel for the segmentation of the normal-tissue. We subtract the Red channel to the Blue channel for the segmentation of the infarcted-tissue (Figure 3 (a)).

Watershed technique is based on immersion simulation. The input image is considered as a topographic surface which is flooded by water starting from regional minima. Watershed lines are formed on the meeting points of water

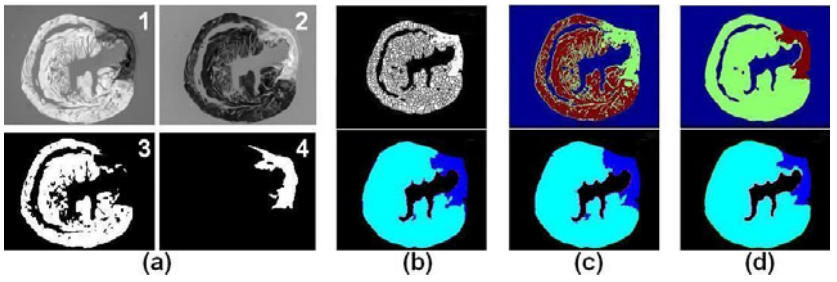


Fig. 3. Segmentations of a heart cross section and identification of the normal and infarted tissue: (a) Combination of the channels (1 and 2) and respective otsu thresholding results (3 and 4), (b) Watershed segmentation result, (c) K-means and (d) Mean-shift clustering results (top) and respectively identification of the regions (down)

coming from distinct minima. All pixels associated with the same catchment basin are assigned to the same label [14,15].

For the application of this technique we use the same image channel combination as for otsu thresholding. Performing watershed segmentation originates an oversegmentation of the tissue since it has many regional minima. However, by comparing the color intensities in each region we are able to decide if each region is from normal-tissue or from infarcted-tissue. Using also the otsu thresholding technique that allows to easily obtain the full tissue segmentation we focus our analysis only on the tissue region. The resulting tissue areas are coherent with normal/infarct tissue-areas (Figure 3 (b)).

K-means clustering technique assigns cluster labels to data points from the entire image [8]. For this technique we use the information of the three channels, selecting three clusters which will correspond to the background, normal- and infarcted-tissue. After obtaining the segmentation result we identify each segmented cluster from its average color intensity. To improve the segmentation we fill the holes using morphological operations (Figure 3 (c)).

Mean-shift clustering technique does not require prior knowledge of the number of clusters and only needs the definition of the radius of the kernel used. As in the previous technique we decided to obtain at most three clusters. If we obtain more than three clusters we iteratively increase the radius of the kernel used (Figure 3 (d)).

In this case we base our segmentation on the Red and Blue channels since they lead to better results than the use of all channels.

Following the segmentation results of full automatic and semi-automatic techniques we can measure the infarct extension.

3 Infarct Extension Evaluation

To better understand the calculation of the infarct extension we must analyze the different regions in the heart. In Figure 4 (A) we can observe the heart

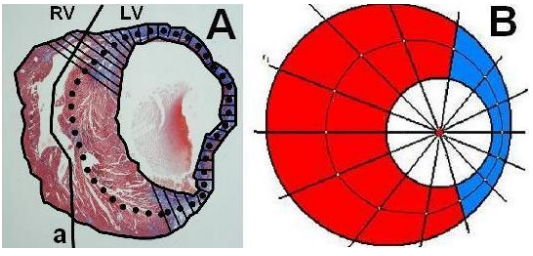


Fig. 4. Image of a heart cross section. A - The heart is bounded by the outside black continuous line which includes the left ventricle (LV) and right ventricle (RV) separated by line **a**. The interior black continuous line is identifying the lumen of the left ventricle and the region marked by lines shows the tissue with infarct. The dotted line is the midline between the inside and outside black continuous lines. B - Scheme of a cross section of the heart.

bounded by the exterior black continuous line. It is formed by right and left ventricles. The first consists only in normal-tissue. The left ventricle includes the infarcted-tissue, which is represented by the shaded region, lumen, which is bounded by the interior black continuous line and also normal-tissue. The infarct extension is usually calculated by two different methods:

Area measurement - Infarct extension is calculated by dividing the infarct area by the area of the heart tissue [1,2] (Figure 4 (A)). This is trivial based on the segmentation results obtained.
Midline length measurement - Infarct extension is calculated by dividing the midline infarct length by the length of midline [2] (Figure 4 (A)). Figure 4 (B) shows a scheme that represents a cross-section of the heart. To perform the midline measurement we first automatically find the midline by tracing lines from the centre of the lumen to the outside of the tissue. The midline is given by the middle distance between tissue borders. The points of the middle line where there is infarcted-tissue in bigger percentage than the normal-tissue (in the radial direction) are considered infarcted points. Secondly we divide the length of infarct midline by the length of the midline. To obtain the lumen of the heart we get a segmentation of all the heart tissue by otsu thresholding technique, which is trivial and we identify the biggest hole inside that segmentation.

The infarct extension is defined as the mean value of infarct extension in all the cross-sections of the heart.

Infarct extension is evaluated in the heart cross-sections considering or not the right ventricle [1,2] (Figure 4 (A)). However, it is not easy to find a robust way to remove the RV as it is variable in morphology and most biologists vary in their assessment of where the RV ends and the LV begins. As such, we perform both the analysis on the full heart, including the RV, and also obtain results on the LV only by removing the RV through image editing tools (manually).

Table 1. Results of infarct extension measurement in mice's hearts. The results are the average value of infarct extension obtained in transverse sections of each mouse heart. In the manual analysis the results only consider the left ventricle.

Infarct extension						
Midline length measurement						
Heart	Manual	Region growing	Otsu	Watershed	K-means	Mean-shift
#1 with RV	-	39%	38%	39%	37%	34%
#1 without RV	43%	38%	40%	41%	38%	36%
#2 with RV	-	47%	47%	48%	46%	47%
#2 without RV	52%	48%	48%	49%	47%	49%
Area measurement						
Heart	Manual	Region growing	Otsu	Watershed	K-means	Mean-shift
#1 with RV	-	29%	18%	17%	17%	14%
#1 without RV	22%	25%	21%	21%	20%	17%
#2 with RV	-	34%	29%	31%	28%	33%
#2 without RV	36%	38%	32%	36%	32%	36%

4 Results

The infarct extension was calculated manually and automatically in two independent hearts. The calculation was performed both on the whole cross-section tissue and also without considering the right heart ventricle for comparison. To automatically segment the tissue without taking into account the right ventricle we manually remove this region before the segmentation process. Table 1 shows the results for the infarct extension evaluation using our approaches and the manual annotation. The results are the average value of the infarct extension over all cross-sections of each independent heart.

Differences between the proposed approaches and manual annotation are never greater than 8% in the case of the evaluation considering the right ventricle. Removing the right ventricle the differences are never greater than 7%. The differences among the proposed approaches in mice's hearts considering the right ventricle tissue are at most 15% and without this are never greater than 8%.

5 Conclusion

The proposed approach enabled the full and semi-automatic calculation of infarct extension. The results obtained using our approaches were in close agreement with the manual annotation with differences never higher than 8%. The segmentation allowed an analysis of the infarct extension in a fraction of the manual method measure time.

Within the automatic segmentation approaches, the watershed technique produced better results, with the differences never above 5% (reduced to 3% by removing the right ventricle). The differences from the semi-automatic approach used were at most 7% considering the right ventricle (5% without this one). Although the differences were slightly higher in the semi-automatic approach, the

biologists prefer the possibility to control the segmentation results in relation to fully automatic approaches.

Future research will focus on integrating automatic image segmentation methods with anatomical models. This will enable the automatic segmentation and measurement of only the left ventricle of the heart, leading to better results.

Acknowledgments

These studies were supported by a FCT Grant Project PTDC/SAU-OSM/68473/2006. PQ and PPÓ are, respectively, *Ciência2008* and *Ciência2007* awardees; DSN is recipient of SFRH/BPD/42254/2007 by FCT (Portuguese government Fellowships).

References

1. Degabriele, N., Griesenbach, U., Sato, K.: Critical appraisal of the mouse model of myocardial infarction. Experimental Physiology 89(4), 497–505 (2004)
2. Takagawa, J., Zhang, Y., Wong, M.: Myocardial infarct size measurement in the mouse chronic infarction model: comparison of area- and length-based approaches. Journal of Applied Physiology 102, 2104–2111 (2007)
3. Alattar, M., Osman, N., Fahmy, A.: Myocardial segmentation using constrained multi-seeded region growing. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, pp. 89–98. Springer, Heidelberg (2010)
4. Wu, Q., Merchant, F., Castleman, K.: Microscope Image Processing, ch. 7. Elsevier, Amsterdam (1996)
5. Gonzalez, R., Woods, R., Eddins, S.: Digital Image Processing Using MATLAB, ch. 9. Pearson Education, London (2004)
6. Sharma, N., Aggarwal, L.: Automated medical image segmentation techniques. Journal of Medical Physics 35(1), 3–14 (2010)
7. Pham, D., Xu, C.: A survey of current methods in medical image segmentation. Annual Review of Biomedical Engineering 2, 315–338 (1998)
8. Hamarneh, G., Li, X.: Watershed segmentation using prior shape and appearance knowledge. Image and Vision Comp. 27(1-2), 59–68 (2009)
9. Khadir, S., Ahamed, R.: Moving toward region-based image segmentation techniques: A study. Journal of Theoretical and Applied Information Technology 5(1), 1–7 (2009)
10. Ahmed, M., Mohamad, D.: Segmentation of brain mr images for tumor extraction by combining kmeans clustering and perona-malik anisotropic diffusion model. International Journal of Image Processing 2(1), 1–8 (2010)
11. Mayer, A., Greenspan, H.: An adaptive mean-shift framework for MRI brain segmentation. IEEE Trans. on Medical Imaging 28(8), 1–12 (2009)
12. Dabov, K., Foi, A., Katkovnik, V.: Image denoising by sparse 3d transform-domain collaborative filtering. IEEE Trans. on Image Processing 16(8), 1–16 (2007)
13. Otsu, N.: A threshold selection method from gray level histograms. IEEE Trans. Systems, Man and Cybernetics 9(1), 62–66 (1979)
14. Marcuzzo, M., Quelhas, P., Campilho, A., Mendonça, A.M., Campilho, A.: Automated arabidopsis plant root cell segmentation based on svm classification and region merging. Computers in Biology and Medicine 39(9), 1–9 (2009)
15. Bleau, A., Leon, J.: Watershed-based segmentation and region merging. Computer Vision and Image Understanding 77(3), 317–370 (2000)

Non-rigid Multi-modal Registration of Coronary Arteries Using SIFTflow

Carlo Gatta[1,2], Simone Balocco[1,2], Victoria Martin-Yuste[3], Ruben Leta[4], and Petia Radeva[2]

[1] Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain
[2] Centre de Visió per Computador, Bellaterra, Spain
[3] Institut Clinic del Torax. Hospital Clinic Barcelona. Spain
[4] Cardiology Service and Institute of Cardiology, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Abstract. The fusion of clinically relevant information coming from different image modalities is an important topic in medical imaging. In particular, different cardiac imaging modalities provides complementary information for the physician: Computer Tomography Angiography (CTA) provides reliable pre-operative information on arteries geometry, even in the presence of chronic total occlusions, while X-Ray Angiography (XRA) allows intra-operative high resolution projections of a specific artery. The non-rigid registration of arteries between these two modalities is a difficult task. In this paper we propose the use of SIFTflow, in registering CTA and XRA images. At the best of our knowledge, this paper proposed SIFTflow as a XRay-CTA registration method for the first time in the literature. To highlight the arteries, so to guide the registration process, the well known Vesselness method has been employed. Results confirm that, to the aim of registration, the arteries must be highlighted and background objects removed as much as possible. Moreover, the comparison with the well known Free Form Deformation technique, suggests that SIFTflow has a great potential in the registration of multimodal medical images.

1 Introduction

Chronic total occlusions (CTO) are obstructions of native coronary arteries with the presence of Thrombolysis In Myocardial Infarction (TIMI) flow grade 0 within the occluded segment, *i.e.* no blood flow, with an estimated occlusion duration of more than 3 months. Re-canalization of a CTO still remains a challenge for invasive cardiologists, due to the fact that the obstructed artery is invisible to X-Ray imaging and thus the navigation of the catheter in the vessel is a delicate and potentially dangerous process.

We suggest one methodology to help the re-canalization: guide the interventionist by means of a proper visualization of coronary arteries from CTA volumes. This is because the occluded artery is visible in the CTA, so that, with an appropriate registration, the cardiologist can actually see the invisible part of the artery by fusing data coming from the pre-operative CTA. Nonetheless,

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 159–166, 2011.

this possibility poses a series of severe and challenging problems: (i) The X-Ray images present a very low signal to noise ratio; (ii) the exact identification of the artery boundaries in angiography sequences, even by an expert medical doctor, could be difficult; (iii) the presence of an heterogeneous background, like spine, diaphragm, bones, stents, catheter guide, etc., makes difficult the automatic segmentation of arteries in X-Ray images; (iv) regarding the multi-modal registration, the CTA volumes do not contain elements that appears in the X-Ray images, thus the registration is actually between two different "objects"; (v) images captured in the two different modalities may represent the coronary tree in different phases of the cardiac cycle and, normally, present an important non-rigid deformation, that is different in both modalities. Considering all of these problems, it is necessary to *segment the arteries* both in the X-Ray and CTA volumes automatically, prior to their *non-rigid registration*. Some recent attempts considered only rigid XRay to CT affine registration, with partial while encouraging results [6,8].

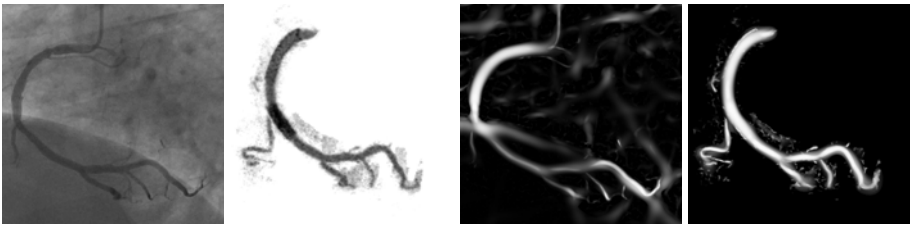
In this paper we show that some form of vessel detection/segmentation is required to obtain a good non-rigid registration between X-Ray and CTA images. Moreover, we propose the use of the SIFTflow [5], a methodology that is typically used for natural images, in the medical image context. The motivation to use SIFTflow for this specific problem, comes from the fact that it provides the required abstraction by describing the image locally using the SIFT descriptor and performing the registration in a way designed for "same class of objects" and not for the exact same object instance. In order to evaluate the approach, the performance of SIFTflow is compared with the well-known Free Form Deformation (FFD) [7], which is currently the state-of-the-art method in medical image registration. The use of SIFTflow in a specific area of medical imaging has been recently introduced in [2], while its use for multimodal registration is, at the best of our knowledge, totally novel. FFD has been successfully used for non-rigid registration purposes, so that it is a good base-line approach for the problem we are facing in this paper. Moreover, it can handle non iso-volumetric transformations, so it fits well with the case of multimodal artery non-rigid registration. In this paper we do not present a comparison to diffeo-morphic methods, since we do not need an invertible transformation; moreover the computational cost of diffeo-morphic methods is significantly higher than FFD and SIFTflow.

The long-term goal of our project is to provide a virtual visualization of the obstructed segment by means of fusing data coming from pre-operative CT and intra-operative X-Ray sequences. The results depicted in this paper provided an experimental proof of the potential of SIFTflow with proper pre-processing as a basic foundation of more sophisticated future methods.

2 Method

The proposed method is based on the proper combination of (i) a vessel segmentation/enhancement algorithm together with (ii) a method for multi-modal non-rigid registration.

To extract the coronary artery in CTA we use the method proposed in [9], as it has been proven to be competitive with other state-of-the-art methods and it is fully automatic. In our experiments, the method performed sufficiently well on low quality CTAs. Using the primary (CRA/CAU) and secondary (LAO/RAO) angles of the angiographic C-ARM, we use the segmented coronary artery to obtain a simulated 2D image, by projecting the 3D data following the C-ARM geometry, as extensively described in [3]. To enhance vessels in X-Ray images, we used the well-known Vesselness method [4]. Figure 1 shows an X-Ray image (a), a simulated X-Ray image using the segmented 3D CTA coronary artery (b), and the respective Vesselness maps (c-d). It is interesting to note that the Vesselness



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

Fig. 1. An X-Ray image (a), a simulated X-Ray image using the segmented 3D CTA coronary artery (b), and the respective Vesselness maps (c-d)

method removes undesired background structures, as the spine bone and non-tubular annoying structures, as well as low spatial frequencies intensity changes. Nonetheless, Vesselness enhances also the catheter, that is in fact a tubular-like structure, and the diaphragm border. These two objects are not visible in the CTO projection, so that they can negatively affect the registration process.

2.1 Registration Methods

In this paper we propose the use of SIFTflow for artery multi-modal registration and compare it to the well known FFD algorithm. To unify the description of both methods, we'll use the following notation. Being $\Phi \subset \mathbb{N}^2$ the lattice support of the image, let define $\boldsymbol{p}^{(M)} \in \Phi$ a generic point of the moving image M, and $\boldsymbol{w}(\boldsymbol{p}^{(M)}) \in \mathbb{R}^2$ as the displacement vector estimated by a non-rigid registration algorithm, that maps the point $\boldsymbol{p}^{(M)}$ to the corresponding point on the static image S, $\boldsymbol{p}^{(S)} = \boldsymbol{p}^{(M)} + \boldsymbol{w}(\boldsymbol{p}^{(M)})$. For the sake of compactness, in the remaining part of the paper we will use $\boldsymbol{w}(\boldsymbol{p}) \triangleq \boldsymbol{w}(\boldsymbol{p}^{(M)})$ considering in an implicit way that the displacement is defined only from the moving M to the static S image. The displacement vector can also be seen in terms of its two orthogonal components as $\boldsymbol{w}(\boldsymbol{p}) = (u(\boldsymbol{p}), v(\boldsymbol{p}))$. With $\boldsymbol{w}$ we denote the non-rigid transformation field, $\forall \boldsymbol{p}$; the symbol $\varepsilon(\boldsymbol{p}) \subset \Phi$ indicates a region centered in a generic point $\boldsymbol{p}$.

Free Form Deformation: Free Form Deformation [7] (FFD) is a non-rigid registration methodology that deforms the moving image M in order to match

the static image S, modeling the non-rigid transformation field $\boldsymbol{w}$ using splines. A regular grid of splines is defined and eq. (1) is minimized over the splines parameters, using the Normalized Mutual Information (NMI) as a similarity measure between images.

$$\mathcal{C}(\boldsymbol{w}) = -\mathrm{NMI}(S(\boldsymbol{p}), M(\boldsymbol{p}+\boldsymbol{w}(\boldsymbol{p}))) + \lambda \sum_{\boldsymbol{p}} \left(\frac{\partial^2 \boldsymbol{w}(\boldsymbol{p})}{\partial x^2} \right)^2 + \left(\frac{\partial^2 \boldsymbol{w}(\boldsymbol{p})}{\partial y^2} \right)^2 \quad (1)$$

The grid is then iteratively refined to allow more local deformations. Eq (1) is composed of two terms: the former evaluates the goodness of the deformation by computing the NMI between the static and deformed moving image; the latter is a smoothness term that forces the second order derivative of the displacement field $\boldsymbol{w}$ to be small. The parameter λ provides a way to balance the contribution of both terms.

SIFTflow: SIFTflow has been introduced first in [5] as a method to register images that represent different instances of the same object class (*i.e.* cars), rather than different images of the same object. This has been obtained by using a dense SIFT descriptor to characterize every image point $\boldsymbol{p}$, denoted with $s(\boldsymbol{p})$. The non-rigid registration process is driven by the minimization of the following equation:

$$\begin{aligned} E(\boldsymbol{w}) = \sum_{\boldsymbol{p}} \| s_S(\boldsymbol{p}) - s_M(\boldsymbol{p}+\boldsymbol{w}) \| + \frac{1}{\sigma^2} \sum_{\boldsymbol{p}} \left(u^2(\boldsymbol{p}) + v^2(\boldsymbol{p}) \right) + \\ \sum_{(\boldsymbol{p},\boldsymbol{q})\in\varepsilon} \min\left(\alpha|u(\boldsymbol{p}) - u(\boldsymbol{q})|, d\right) + \min\left(\alpha|v(\boldsymbol{p}) - v(\boldsymbol{q})|, d\right). \end{aligned} \quad (2)$$

The first term accounts for the dissimilarity in terms of SIFT descriptor. Since the definition of $\boldsymbol{w}$ is not regularized, and the space search regions is large (40×40 pixels), the second term provides a first order regularization on $\boldsymbol{w}$; finally, the last term promotes the smoothness of $\boldsymbol{w}$. Parameters σ, α and d are used to provide the desired balance between the terms.

3 Experimental Section

3.1 Material

To evaluate the proposed method, we used a set of 6 angiographic images from 4 patients. Images of Right Coronary Artery (RCA) have been acquired with a Philips Xcelera equipment, with pixel size of 0.33×0.33 mm. Regarding the CTA data, a set of 4 volumes has been acquired using a Toshiba Aquilion, with voxel size of $0.43\times0.43\times0.5$ mm. The coronary tree has been automatically segmented using the algorithm in [9], and then the RCA has been manually isolated from the aorta and other arteries. Finally, we collected only RCA images since the CTO is a pathology that mainly affects the RCA [1]. It is worth to note that the

actual image resolution is higher since the image on the intensifier is a zoomed version of the coronary tree, due to the perspective effect. To provide measures of the actual vessels, the XRay image resolution has been estimated by relating the physical size of the catheter (a *6 French* in all experiments) to its size in millimeters. This resulted in an actual resolution of 0.237×0.237 mm. From here on, all the measures in millimeters refer to real-world dimensions using the proportion of 0.237 mm per pixels.

3.2 Algorithms Setting

We are using three algorithms: a vessel enhancement method (Vesselness), FFD and SF. Regarding Vesselness, we set the scales in octaves, *i.e.* $s = 2^S$, for $S = \{1, 2, 3, 4\}$; this allows to highlight vessels with a caliber from 0.47 to a maximum of 7.58 mm. The parameters that weights the "tube-likeness" and the second order "structureness" have been set respectively to $\beta = 0.75$ and $c = 0.33$. The FFD has two main parameters, the λ (see eq. (1)) and the initial grid. In the experiments of this paper, we set $\lambda = 0$ and initial grid spacing of 128×128. This parametrization allows the FFD to handle the big deformations present in the images, and still provides a sufficient smoothing due to the use of splines. SIFTflow has different parameters that controls the smoothness of the solution and the magnitude of the displacements. In this paper we set $\alpha = 2$, $\sigma = 14$ and $d = 40$. This parametrization has been experimental derived starting from the standard settings in [5].

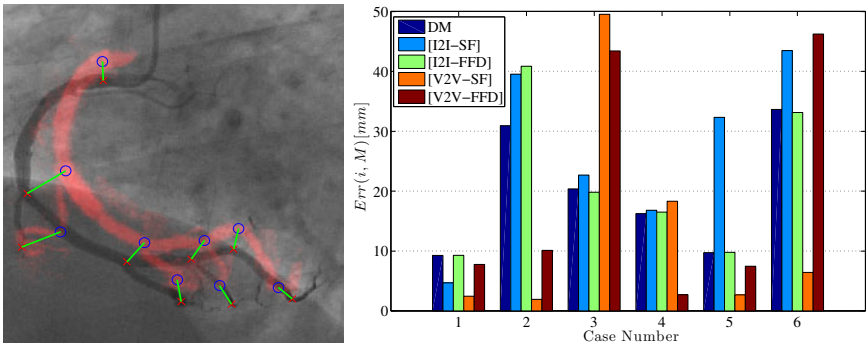
3.3 Results

To evaluate the automatic non-rigid registration, we collected the ground truth transformation by manually setting the correspondence between CTA and XRay images, as depicted in Fig. 2 (a). Due to the complexity of the images, a reliable ground truth can be obtained only by correspondences of clear landmarks as bifurcations, starting and ending points, and high curvature locations. For a given image i, we collected a number L of landmarks l, which ground truth translation is defined as $\boldsymbol{w}^{(i)}(l)$.

In our experiments, we performed the registration using four different pipelines: (1) FFD is applied on the images without any pre-processing (named [I2I-FFD]) and (2) with the Vesselness method (named [V2V-FFD]); (3) SF is applied on images without any pre-processing (named [I2I-SF]) and (4) with the Vesselness method (named [V2V-SF]). For all the cases, and for each method M, the displacement field is compared to the manually annotated landmarks translation as follows:

$$Err(i, M) = \frac{1}{L} \sum_{l} \|\boldsymbol{w}^{(i)}(l) - \boldsymbol{w}_M^{(i)}(l)\|.$$

The resulting average registration error is to be considered an upper bound of the actual error if considering the whole vessel instead of few landmarks. Figure 2(b)



(a) (b)

Fig. 2. (a) Ground-truth manually-annotated landmark registration pairs and (b) The registration error $Err(i, M)$ for all cases and methods

summarizes the results of the 6 cases we analyzed. The "DM" bars represent the average magnitude of the displacement between annotated landmark in XRA and CTA images; the rest of bars represent the average registration error of all the compared methods. As a general trend, the direct [I2I-FFD] registration is totally ineffective, despite the FFD can handle different modalities due to the use of the Mutual Information; nonetheless the background in XRA images does not allow FFD to find an acceptable solution, even at large grid spacings. The direct [I2I-SF] registration fails to converge to an optimal solution, and performs well in only one case (number 1); in some cases SF is prone to produce non-realistic results that totally disrupt the image content (see e.g. case # 5). When using vesselness as a pre-processing step, FFD improves its performance, getting the best result for case # 4 and slightly improving for case # 5. The [V2V-SF] method presents the best performance, being the best on 4 cases over 6.

Figure 3 show all the results of the experiment; the red overlay represents the artery obtained from the CTA projection. The first column shows the XRA and CTA without any registration. Cases # 3 and 6 present very challenging examples: in case # 3 the deformation of the artery in the XRA is so important that the two images represent two very different vessel morphologies. In case # 6 the CTA artery does not present a branch, that is visible in the XRA image; moreover the artery has few bifurcations, so that there is poor structure to help the registration.

4 Conclusion

In this paper we have shown preliminary results on non-rigid registration of coronary artery between X-Ray and CTA modalities. As expected, the image-to-image registration is a difficult task even for a mutual information-based algorithm, and a pre-processing step is necessary. In this paper we investigated the use of the well known Vesselness method, with encouraging results; nonetheless,

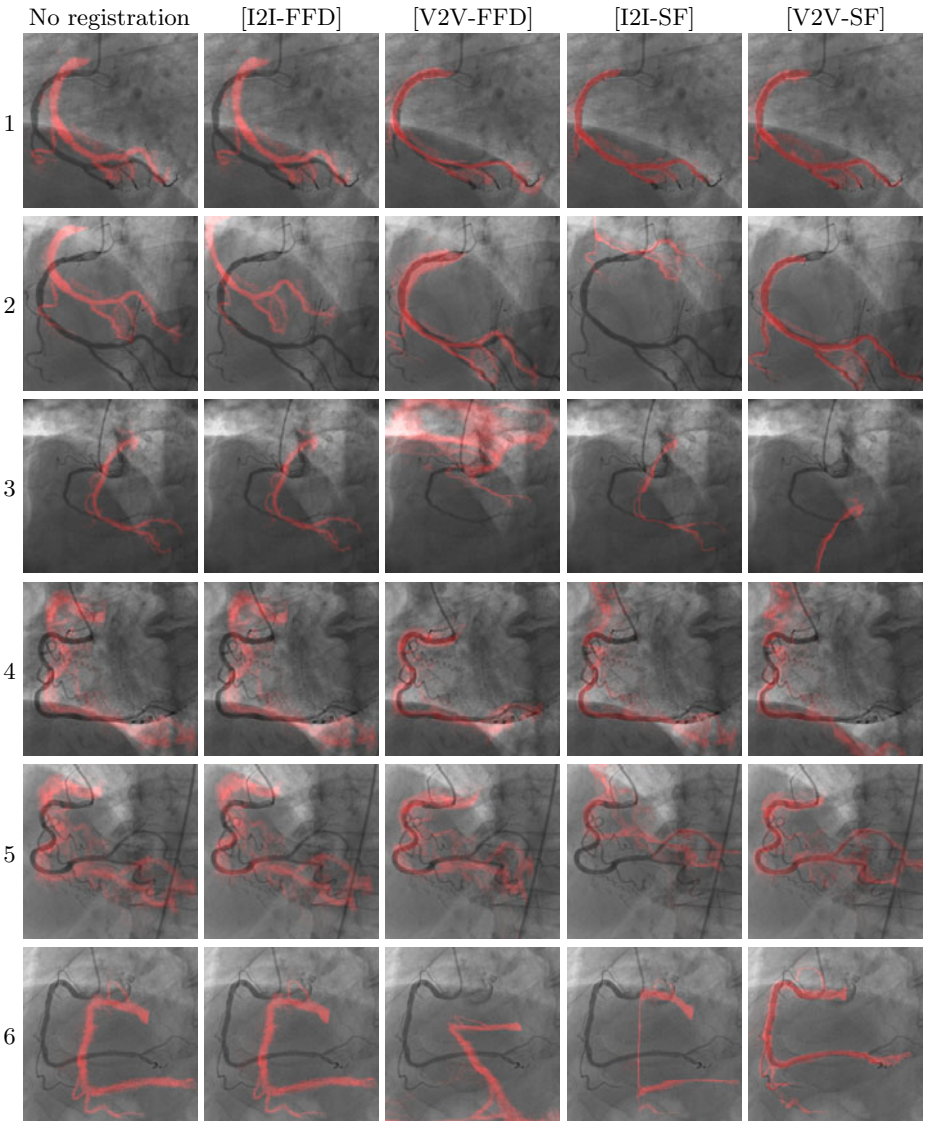


Fig. 3. Results of non-rigid registration with different pre-processing methods. The over-imposed red artery is the 2D CTA projection deformed according to the mentioned method.

a higher level of abstraction should be necessary to remove background annoying objects. In some sense, SIFTflow provides the required abstraction by describing the image locally using the SIFT descriptor and performing the registration in a way designed for "same class of objects" and not for the exact same object instance. The analysis proposed in this paper does not want to be exhaustive and

serves as a proof of concept for future research in segmentation and non-rigid registration of multi-modal cardiac imaging. Future works will be devoted in further regularization of the SIFTflow result, using a-priori knowledge of arterial motion; extend the SIFTflow functional to be able to handle partial occlusions, thus dealing with the CTO problem, and compare with other state-of-the-art registration algorithms.

Acknowledgments

This work has been supported in part by the projects: La Marató de TV3 082131, TIN2009-14404-C02, and CONSOLIDER-INGENIO CSD 2007-00018. The work of C. Gatta is supported by a Beatriu de Pinos Fellowship.

References

1. Cohen, H., Williams, D., Holmes, D.J., Selzer, F., Kip, K., Johnston, J., Holubkov, R., Kelsey, S., Detre, K.: Impact of age on procedural and 1-year outcome in percutaneous transluminal coronary angioplasty: a report from the nhlbi dynamic registry. Am. Heart J. 146, 513–519 (2003)
2. Drozdzal, M., Igual, L., Vitria, J., Malagelada, C., Azpiroz, F., Radeva, P.I.: Aligning endoluminal scene sequences in wireless capsule endoscopy. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis MMBIA 2010, pp. 117–124 (2010)
3. Dumay, A., Reiber, J., Gerbrands, J.: Determination of optimal angiographic viewing angles: basic principles and evaluation study. IEEE Transactions on Medical Imaging 13(1), 13–24 (1994)
4. Frangi, A., Niessen, W., Vincken, K., Viergever, M.: Multiscale vessel enhancement filtering. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 130–137. Springer, Heidelberg (1998)
5. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT flow: Dense correspondence across different scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 28–42. Springer, Heidelberg (2008)
6. Metz, C., Schaap, M., Klein, S., Neefjes, L.A., Capuano, E., Schultz, C., van Geuns, R.J., Serruys, P.W., van Walsum, T., Niessen, W.J.: Patient specific 4D coronary models from ECG-gated CTA data for intra-operative dynamic alignment of CTA with X-ray images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 369–376. Springer, Heidelberg (2009)
7. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast mr images. IEEE Transactions on Medical Imaging 18, 712–721 (1999)
8. Ruijters, D., ter Haar Romeny, B.M., Suetens, P.: Vesselness-based 2d-3d registration of the coronary arteries. Int. J. Comput. Assist. Radiol. Surg. 4(4), 391–397 (2009)
9. Wang, C., Smedby, Ö.: Coronary artery segmentation and skeletonization based on competing fuzzy connectedness tree. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 311–318. Springer, Heidelberg (2007)

Diffuse Liver Disease Classification from Ultrasound Surface Characterization, Clinical and Laboratorial Data*

Ricardo Ribeiro[1,2,**], Rui Marinho[3], José Velosa[3], Fernando Ramalho[3], and J. Miguel Sanches[1]

[1] Institute for Systems and Robotics / Instituto Superior Técnico
[2] Escola Superior de Tecnologia da Saúde de Lisboa
[3] Liver Unit, Department of Gastroenterology and Hepatology, Hospital de Santa Maria, Medical School of Lisbon
Lisbon, Portugal
ricardo.ribeiro@estesl.ipl.pt

Abstract. In this work liver contour is semi-automatically segmented and quantified in order to help the identification and diagnosis of diffuse liver disease. The features extracted from the liver contour are jointly used with clinical and laboratorial data in the staging process. The classification results of a *support vector machine*, a *Bayesian* and a *k-nearest neighbor* classifier are compared. A population of 88 patients at five different stages of diffuse liver disease and a leave-one-out cross-validation strategy are used in the classification process. The best results are obtained using the *k-nearest neighbor* classifier, with an overall accuracy of 80.68%. The good performance of the proposed method shows a reliable indicator that can improve the information in the staging of diffuse liver disease.

Keywords: Liver Cirrhosis, Contour Detection, Ultrasound, Classification.

1 Introduction

Staging of liver disease is needed because it is progressive, most of the time asymptomatic and potentially fatal. An accurate characterization of this disease is difficult but crucial to prevent its evolution and avoid irreversible pathologies such as the *hepatocellular carcinoma*.

Fatty liver infiltration (*steatosis*) is the earliest stage of the liver disease. It is asymptomatic and the progress of the hepatic injury to other conditions, more severe, is common. e.g., fibrosis. Pathologically, fibrosis appears during the course of organ damage and its progression rate strongly depends on the cause of liver disease, such as *chronic hepatitis* [1].

* This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds.
** Corresponding author.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 167–175, 2011.

Cirrhosis is the end-stage of every chronic liver disease. It is characterized by an asymptomatic stage, known as *compensated cirrhosis*, followed by a rapidly progressive phase where liver dysfunction occurs, called *decompensated cirrhosis*. The most severe evolution condition of the cirrhosis is the *hepatocelllular carcinoma* (HCC), also called, primary liver cancer [1].

Liver biopsy is the most accurate method for diagnosis. Due to its highly invasive nature, medical image modalities have been valuable alternative methods to detect and quantify this disease [1]. The non-ionizing and non-invasive nature of ultrasound (US) imaging and its widespread presence at almost all medical facilities makes it the preferred method for assessing diffuse liver diseases such as cirrhosis.

Using US, cirrhosis is suggested by liver surface nodularity, portal vein mean flow velocity and the enlargement of the caudate lobe [1]. The study in [2] refer that nodular liver surface is a reliable sign in the detection of liver cirrhosis and can have a diagnostic accuracy of 70% or more. The authors in [3] showed that the observed liver contour irregularities directly correlated with the gross appearance of the cirrhotic liver as seen at laparoscopy. Liver surface nodularity in US sign can be appreciated when ascites is present or when a high-frequency transducer (7.5 - 12 MHz) is used [3]. In [2] the results, using a low-frequency transducer (3.5 -5 MHz), also showed that liver surface is a significantly parameter associated with the histopathological diagnosis of liver cirrhosis.

Nevertheless, as reported by [4], the validity of the different methods to detect changes in the liver surface are very subjective, since the segmentation and contour of such surface is operator-dependent. These fact leads to a subjective and non reproducible method to study the liver surface and consequently to a poor aid of an accurate liver diagnosis.

In this sense, it is proposed a semi-automatic method for the liver surface detection, based on an image processing procedure that decomposes the US images of the liver parenchyma into two fields: the *speckle* image containing textural information and the *de-speckled* image containing intensity and anatomical information of the liver. Features extracted from the liver contour detected in the *de-speckled* field, as well as clinical and laboratorial features, are used to train supervised classifiers to detect the disease.

Diffuse liver disease stages are considered and several classifiers are used to assess the discriminative power of the selected features: (i) the *support vector machine* (SVM), (ii) the *Bayesian classifier* and (iii) the *k-nearest neighbor* (kNN).

Several *figures of merit* (FOM) were computed to assess and compare the performance of each classifier.

This paper is organized as follows. Section 2 formulates the problem and describes the pre-processing procedures, the extraction and selection of features and classifiers. Section 3 presents the experimental tests, by reporting and comparing the classification results obtained with the features extracted from the liver contour, with the clinical and laboratorial features and with the total set of features. Section 4 concludes the paper.

2 Problem Formulation

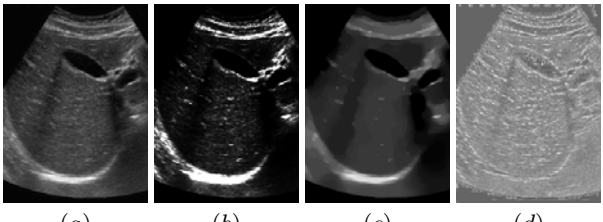
In the practice of US the perceived liver capsule and the adjacent overlying membranous structures (peritoneum, transverse fascia, pre-peritoneal fat) are not always clear and irregularities due to subfascial or sub-peritoneal pathology may be falsely described as abnormalities of the liver surface [3].

The decomposition procedure described in [5] to separate the textural and intensity information within US images is here adopted. In this approach an estimation of the *radio frequency* (RF) raw data is firstly done based on physical considerations about the data generation process, namely, by taking into account the dynamic range compression performed by the US equipment over the signal generated by the US probe. The observation model, in this approach, also considers the brightness and contrast parameters tuned by the medical doctor during the exam which changes from patient to patient.

The estimated RF image is decomposed in *de-speckled* and *speckle* fields according to the following model [6]

$$y(i,j) = x(i,j)\eta(i,j), \tag{1}$$

where $\eta(i,j)$ are considered independent and identically distributed (i.i.d.) random variables with *Rayleigh* distribution. This image describes the noise and textural information and is called *speckle* field. In this model, the noise is multiplicative in the sense that its variance, observed in the original image, depends on the underlying signal, $x(i,j)$. Fig.1 illustrates an example of the decomposition methodology.



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

Fig. 1. Decomposition procedure of US liver parenchyma. a) Observed *B-mode* US image. Estimated b) envelope RF image, c) *de-despeckled* and d) *speckle* image fields.

In the estimation of both images, RF envelope and *de-speckled* information, the use of total variation techniques allows the preservation of major transitions, as seen in the case of liver capsule and overlying structures.

Using the *de-despeckled* image, the liver surface contour is obtain using a snake technique, proposed by [7], which computes one iteration of the energy-minimization of active contour models. To initialize the snake, the operator needs to select four points of the liver surface.

Based on the detected **liver contour**, the following features were extracted:

1. root mean square of the different angles produced by the points that characterize the contour, rms_α, where the first point was assumed as the reference point,
2. root mean square of the variation of the points of the contour in the y axis, rms_y,
3. the mean and variance of the referred angles, m_α and v_α,
4. the variance of the y axis coordinates at each point, v_y, and
5. the correlation coefficient of the y axis coordinates, R.

Besides image based features, several other clinical data and biochemical tests are useful for evaluating and managing patients with hepatic dysfunction. The clinical study of the disease, conducted in [1], reported the following meaningful **clinical information** to be used:

1. Cause of disease (*diagnose*), which include none (0), alcohol (1), hepatitis B (2), hepatitis C (3), alcoholic hepatitis B (4) and C (5) and others (6), and the following binary indicators:
2. Tumor (T),
3. Ascites (A) which is the presence of free fluid within the peritoneal cavity, encephalopathy (*Ence*),
4. Gastro-Intestinal bleeding (GIB), infection (Inf) and alcoholic habits (Alc).

The **laboratorial features** related with the liver function [1] are: i) total bilirubin (Bil), ii) prothrombin time (INR), iii) albumin (Al), iv) creatinine ($Crea$), v) aspartate transaminase (AST), vi) alanine transaminase (ALT), vii) gamma glutamyl transpeptidase (gGT), viii) glycemia (Gly), ix) sodium (Na) and x) lactate dehydrogenase (LDH).

All these features, organized in a 23 length vector, are used in a forward selection method with the criterion of 1 - Nearest Neighbor *leave-one-out cross-validation* (LOOCV) performance in order to select the most significant features and increase the discriminative power of the classifier. Three different classifiers were implement and tested: i) the SVM, ii) *Bayesian classifier* and iii) kNN. A short description of each one is provided.

The aim of SVM is to find a decision plane that has a maximum distance (margin) from the nearest training pattern [8]. Given the training data $\{(x_i, \omega_i)|\omega_i = 1 \; or \; -1, \; i = 1, ..., N\}$ for a two-class classification (where x_i is the input feature; ω_i is the class label and N is the number of training sample), the SVM maps the features to a higher-dimensional space. Then, SVM finds a hyperplane to separate the two classes with the decision boundary set by the support vectors [8]. In this paper, a multiclass SVM classifier was adopted, using a Gaussian radial-basis kernel function and a polynomial kernel.

In the Bayes classifier the feature set, X, is assumed multivariate normal distributed [9] with means, μ_τ and covariances matrices, Σ_τ, according to each class. The linear discriminant functions are

$$g_\tau(X) = -\frac{1}{2}(X - \mu_\tau)^T \Sigma_\tau^{-1}(X - \mu_\tau) - \frac{1}{2} log\,|\Sigma_\tau| \qquad (2)$$

where $\tau \in \{N, CHC, CC, DC, HCC\}$ and the *a priori* probability, $P(\omega_\tau)$, of the classes were calculated based on their frequencies.

The non-parametric kNN classifier is also tested in this paper. It classifies a test sample to a class according to the majority of the training neighbors in the feature space by using the minimum Euclidean distance criterion [10]. All classifiers were implemented using the algorithm proposed by [11].

3 Experimental Results

Eighty eight data samples were obtained from 88 patients. The patients were selected from the Department of Gastroenterology and Hepatology of the Santa Maria Hospital, in Lisbon, with known diagnosis. The samples were labeled in five classes; *Normal*, ω_N, *Chronic Hepatitis without cirrhosis*, ω_{CHC}, *Compensated Cirrhosis*, ω_{CC}, *Decompensated Cirrhosis*, ω_{DC}, and *Hepatocelular Carcinoma*, ω_{HCC}. Among them, 36 belong to ω_N, 7 to ω_{CHC}, 8 to ω_{CC}, 32 to ω_{DC} and 5 patients to ω_{HCC}.

From figure 2 we can appreciate the results obtained from the *de-speckled* and contour steps. To standardize the proceedings, and as reported in the literature, we focused the study in the anterior liver surface, using a low-frequency transducer. The results showed that in the *de-speckled* US image the liver boundary was clearly defined in the cases reported (for example, normal liver (first row), cirrhotic liver without ascites (second row) and cirrhotic liver with ascites (third row)). The detected contour was plotted in the original US image, so that the physician could evaluate the liver contour.

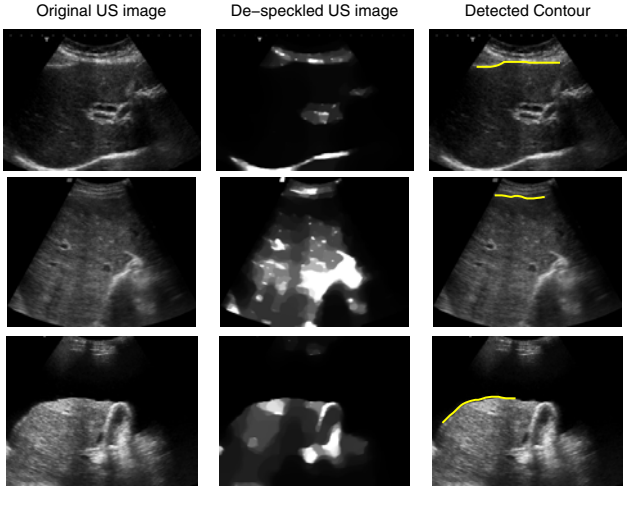


Fig. 2. Method used to detect the anterior liver surface contour. First row corresponds to a normal liver; second row to a compensated cirrhotic liver and the last row to a decompensated cirrhotic liver.

According to the criterions established for feature selection the results, reported in Table 1, showed 5 optimal features using only the contour features (optimal contour set), 8 optimal features using the clinical and laboratorial features (optimal clinical set) and 10 optimal features using the set of features (optimal feature set). In this last result, it is important to emphasize that the feature set selected is composed of five clinical features (A, $Ence$, T, $diagnose$, GIB), four laboratory features (AST, INR, Na, $Crea$) and one US contour feature (R). Thus, the combination of features from different sources (US image, clinical, etc) integrates, in a quantitative and objective way, the information used in medical practice. This result is in accordance with [3,12].

Table 1. Feature selection results from the evaluation of the feature set, using only the contour features, only the clinical and laboratorial features and all features

Feature Selection results for	
Contour features	m_α, R, rms_α, v_y, v_α.
Clinical/Lab. features	*AST*, *A*, *Ence*, *T*, *diagnose*, *INR*, *LDH*, *GIB*
All features	*AST*, *A*, *Ence*, *T*, *diagnose*, *INR*, *R*, *Na*, *GIB*, *Crea*.

The classification technique significantly affects the final diagnosis [10]. Using the LOOCV method, the same data set was tested with different types of classifiers, namely a kNN, a *Bayes* classifier and a SVM classifier with polynomial (SVM_P) and radial-basis (SVM_R) kernels.

To determine the optimal parameters for the classifiers the following procedures were done. The kNN algorithm was implemented for values of k=1,2,3,5,7 and 9. The SVM_P was trained with a degree range of $[1 : 5]$ and the SVM_R was implemented with a radius close to 1 ([0.1,0.2,...,2]). The best performance of the kNN classifier was achieved with $k = 1$ for the optimal contour, clinical and feature set, which resulted in an error rate of 40.90%, 21.59% and 19.32%, respectively.

Considering the SVM classifiers, using the proposed sets, the best result of the SVM_P corresponds to a degree of 1, attaining an error rate of 45.45% for the optimal contour set, 25.0% for the optimal clinical set and 23.86% for the feature set. With the SVM_R the best performance for the optimal contour set was obtained with a radius of 1 showing an error of 48.86%, an error rate of 21.59% for the optimal clinical set with a radius of 1.9, and a radius of 1.8, for the case of the optimal feature set, with an error of 27.27%.

In the case of the *Bayesian* classifier, for each class, the mean and covariance were estimated using the LOOCV method.

Table 2 resumes the classification results obtained using the optimal contour and clinical set. The best overall accuracy, using the optimal contour set, of 59.10% was achieved with the kNN classifier, followed by the SVM_P, 54.55%, the SVM_R, 51.14% and the *Bayesian* classifier, with the worst result of the tested classifiers, achieving an overall accuracy of 31.82%. The diagnostic yield

Table 2. Overall and individual class accuracies (%) obtained with different classifiers, using the optimal contour and clinical set

	ω_N	ω_{CHC}	ω_{CC}	ω_{DC}	ω_{HCC}	Overall
Optimal Contour Set						
Bayes	27.78	28.57	0.00	37.50	80.00	31.82
kNN (k=1)	88.89	28.57	0.00	56.25	0.00	59.10
SVM_P	83.33	0.00	0.00	56.25	0.00	54.55
SVM_R	69.44	0.00	0.00	62.50	0.00	51.14
Optimal Clinical Set						
Bayes	0.00	28.57	0.00	0.00	100.00	7.95
kNN (k=1)	94.44	14.29	37.50	87.50	60.00	78.41
SVM_P	91.67	0.00	0.00	90.63	40.00	72.73
SVM_R	94.44	0.00	37.50	93.75	40.00	78.41

Table 3. Overall and individual class accuracies (%) obtained with different classifiers, using the optimal feature set

	ω_N	ω_{CHC}	ω_{CC}	ω_{DC}	ω_{HCC}	Overall
Bayes	0.00	28.57	12.50	0.0	100.00	9.10
kNN (k=1)	91.67	71.43	25.00	87.50	60.00	80.68
SVM_P	91.67	28.57	0.00	87.50	80.00	76.14
SVM_R	88.88	0.00	0.00	93.75	40.00	72.73

was improved from 59.10% to 78.41% when using the optimal clinical set, for kNN and SVM_R classifier. The accuracy results were greatly improved with this feature set for the individual class classification. By means of SVM_R it was obtained an accuracy of 94.44%, 93.75% and 37.50% for ω_N, ω_{DC} and ω_{CC}, respectively. For ω_{HCC} and ω_{CHC}, the best results were 100.0% and 28.27%, respectively, with the Bayes classifier.

Combining features further improves the classifiers performance, as summarized in Table 3. With the optimal feature set, the best overall result was obtained with the kNN classifier. This result outperformed the best result obtained with the previous feature sets, which reinforce the idea of feature combination from different sources.

In terms of individual class accuracy, the best result for ω_N was obtained using the kNN and SVM_P classifiers, both with an accuracy of 91.67%. The best outcome in differentiating chronic hepatitis without cirrhosis samples (ω_{CHC}) and compensated cirrhosis (ω_{CC}), from the other classes was achieved by means of the kNN classifier with an accuracy of 71.43% and 25.00%, respectively. Regarding the classification of ω_{DC}, the best individual accuracy result was reached with the SVM_R classifier, yielding 93.75%. The detection of ω_{HCC} was 100.0% using the *Bayes* classifier.

4 Conclusions

In this work a semi-automatic detection of liver surface is proposed to help in the diagnosis of diffuse liver disease. The results shown in this paper suggest the usefulness of combining US liver contour features with laboratorial and clinical parameters for accurately identifying different stages of diffuse liver disease.

The pre-classification steps showed good results, since the *de-speckled* image field aided the detection of liver surface contour.

The optimal feature set outperformed the optimal contour and clinical set. In the classification procedure, using the optimal feature set, the kNN outperformed the rest of the classifiers in terms of overall accuracy. The low accuracy results in ω_{CC}, maybe due to the small sample size of the class. Another interesting result was the classification accuracy improvement in ω_N using the optimal clinical set. This finding demonstrated the problem of the semi-automatic detection of the contour, since it has an operator-dependent component, the initialization step.

Promising results were obtained, which showed the discriminant power of the features as well as of the classifier, specially in terms of individual class accuracy. These results promote the development of more robust classification techniques, particularly classification combiners.

In the future the authors intend to (i) expand the data set in order to obtain an equitable number of samples in each class, (ii) include other features to increase diagnostic accuracy, (iii) perform a more exhaustive analysis in terms of classifiers, such as using a combination of classifiers and (iv) use state-of-the-art automatic snakes, in order to create a fully automatic detection method.

References

1. Sherlock, S., Dooley, J.: Diseases of the liver and Biliary System, 11th edn. Blackwell Science Ltd., Malden (2002)
2. Gaiani, S., Gramantieri, L., Venturoli, N., Piscaglia, F., Siringo, S., D'Errico, A., Zironi, G., Grigioni, W., Bolondi, L.: What is the criterion for differentiating chronic hepatitis from compensated cirrhosis? a prospective study comparing ultrasonography and percutaneous liver biopsy. Journal of Hepatology 27(6), 979–985 (1997)
3. Simonovsky, V.: The diagnosis of cirrhosis by high resolution ultrasound of the liver surface. Br. J. Radiol. 72(853), 29–34 (1999)
4. Ladenheim, J.A., Luba, D.G., Yao, F., Gregory, P.B., Jeffrey, R.B., Garcia, G.: Limitations of liver surface US in the diagnosis of cirrhosis. Radiology 185(1), 21–23 (1992)
5. Seabra, J., Sanches, J.: Modeling log-compressed ultrasound images for radio frequency signal recovery. In: 30th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, EMBS 2008 (2008)
6. Seabra, J.C., Sanches, J.M.: On estimating de-speckled and speckle components from b-mode ultrasound images. In: Proceedings of the 2010 IEEE International Conference on Biomedical Imaging: from Nano to Macro, ISBI 2010, pp. 284–287. IEEE Press, Los Alamitos (2010)

7. Bregler, C., Slaney, M.: Snakes-A MatLab MEX file to demonstrate snake contour-following (1995)
8. Yeh, W., Jeng, Y., Li, C., Lee, P., Li, P.: Liver fibrosis grade classification with b-mode ultrasound. Ultrasound in Medicine & Biology 29, 1229–1235 (2003)
9. Mojsilovic, A., Markovic, S., Popovic, M.: Characterization of visually similar diffuse diseases from b-scan liver images with the nonseparable wavelet transform. In: International Conference on Image Processing, vol. 3, p. 547 (1997)
10. Kadah, Y., Farag, A., Zurada, J.M., Badawi, A.M., Youssef, A.M.: Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images. IEEE Trans. Med. Imaging 15, 466–478 (1996)
11. van der Heijden, F., Duin, R., de Ridder, D., Tax, D.M.J.: Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB, 1st edn. Wiley, Chichester (2004)
12. Berzigotti, A., Abraldes, J.G., Tandon, P., Erice, E., Gilabert, R., Garca-Pagan, J.C., Bosch, J.: Ultrasonographic evaluation of liver surface and transient elastography in clinically doubtful cirrhosis. Journal of Hepatology 52(6), 846–853 (2010)

Classification of Ultrasound Medical Images Using Distance Based Feature Selection and Fuzzy-SVM

Abu Sayeed Md. Sohail[1], Prabir Bhattacharya[2], Sudhir P. Mudur[3], and Srinivasan Krishnamurthy[4]

[1,3] Dept. of Computer Science, Concordia University, Montreal, Canada
[2] Dept. of Computer Science, University of Cincinnati, Ohio, USA
[4] Dept. of Obstetrics and Gynecology, Royal Victoria Hospital, Montreal, Canada
a_sohai@cse.concordia.ca,bhattapr@ucmail.uc.edu,mudur@cse.concordia.ca, srinivasan.krishnamurthy@muhc.mcgill.ca

Abstract. This paper presents a method of classifying ultrasound medical images towards dealing with two important aspects: (i) optimal feature subset selection for representing ultrasound medical images and (ii) improvement of classification accuracy by avoiding outliers. An objective function combining the concept of between-class distance and within-class divergence among the training dataset has been proposed as the evaluation criteria of feature selection. Searching for the optimal subset of features has been performed using Multi-Objective Genetic Algorithm (MOGA). Applying the proposed criteria, a subset of Grey Level Co-occurrence Matrix (GLCM) and Grey Level Run Length Matrix (GLRLM) based statistical texture descriptors have been identified that maximizes separability among the classes of the training dataset. To avoid the impact of noisy data during classification, Fuzzy Support Vector Machine (FSVM) has been adopted that reduces the effects of outliers by taking into account the level of significance of each training sample. The proposed approach of ultrasound medical image classification has been tested using a database of 679 ultrasound ovarian images and 89.60% average classification accuracy has been achieved.

Keywords: Medical Image Classification, Feature Selection, Fuzzy-SVM.

1 Introduction

As a low cost alternative to Magnetic Resonance Imaging (MRI), Ultrasound imaging has become quite popular during the last decade. Therefore, research on Computer Aided Diagnosis (CAD) using ultrasound medical images has gained significant momentum. So far, ultrasound breast and liver images have been mostly exercised in this regard and no proposition has been made for a CAD system using ultrasound ovarian images to provide decision support in the diagnosis of ovarian abnormalities. In consideration to feature extraction methods adopted, GLCM and GLRLM based statistical texture descriptors remain among

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 176–183, 2011.

the top choices. However, an approach of identifying and subsequently eliminating redundancy as well as irrelevance among the GLCM and GLRLM based statistical texture features in classifying ultrasound medical images could not be found in literature.

The process of medical image classification serves as a core component in developing any image analysis based computer-aided diagnosis system. Since the size of the training dataset for a CAD system needs to be sufficient to cover every possible variation, the dimension of the feature vectors cannot be allowed to grow arbitrarily large in designing an efficient image classification system. After a certain point, increase of the dimension ultimately causes deterioration of performance through introducing redundancy and irrelevance [1]. Besides, the number of operation required by a classifier is directly related to the dimension of the feature vector. For a K class classification problem where each feature vector consists of M attributes, a linear classifier requires to perform $O(KM)$ operations. For a quadratic classification technique, the required computational operation becomes $O(KM^2)$. Therefore, to achieve better classification performance as well to optimize computational cost, it is very important to keep the dimension of the feature vectors as low as possible while maximizing the separability among the classes of the training dataset.

2 Distance Based Criteria for Feature Selection

Our proposed multi-objective criteria for feature subset selection combines the concept of between-class distance and within-class divergence. Therefore, the ultimate objective becomes to select a subset of image features that (i) maximizes the distances among the classes, and (ii) minimizes the divergence within each class. Let T_S be a labelled training set with N_S samples. The classes ω_k are represented by subsets $T_k \subset T_S$, each class having N_k samples ($\sum N_k = N_S$). Measurement vectors in T_S (without reference to their class) are denoted by z_n. Measurement vectors in T_k (vectors coming from class ω_k are denoted by $z_{k,n}$. The sample mean of a class $\hat{\mu}_k(T_k)$ and that of the entire training set $\hat{\mu}(T_S)$ can be defined respectively as [1]: $\hat{\mu}_k(T_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} z_{k,n}$ and $\hat{\mu}(T_S) = \frac{1}{N_S} \sum_{n=1}^{N_S} z_n$. The matrix that describes the scattering of vectors from class ω_k is:

$$S_k(T_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} (z_{k,n} - \hat{\mu}_k)(z_{k,n} - \hat{\mu}_k)^T \tag{1}$$

Averaged over all classes, the scatter matrix describing the distribution of samples around its class center is given by:

$$S_w(T_S) = \frac{1}{N_S} \sum_{k=1}^{K} N_k S_k(T_k) = \frac{1}{N_S} \sum_{k=1}^{K} \sum_{n=1}^{N_k} (z_{k,n} - \hat{\mu}_k)(z_{k,n} - \hat{\mu}_k)^T \tag{2}$$

This matrix is the within-scatter matrix as it describes the average scattering within classes, and can be termed as "within-class divergence". Complementary to this is the between-scatter matrix S_b that describes the scattering of the

class-dependent sample means around the overall average, and can be termed as "between-class distance":

$$S_b(T_S) = \frac{1}{N_S} \sum_{n=1}^{K} N_k(\hat{\mu}_k - \mu)(\hat{\mu}_k - \mu)^T \tag{3}$$

Now, the minimal subset of features that maximizes S_b and minimizes S_w simultaneously should be the optimal subset of features. Using the weighted sum approach of solving multi-objective optimization problems, these two objectives could be converted to a scaler single objective function as follows:

$$F(T_S) = \alpha S_b(T_S) - \beta S_w(T_S) \tag{4}$$

Here, α and β are parameters used to specify the importance (weight) of each individual objective. The value of α is chosen as: $0 < \alpha < 1$, and the value of β is set with respect to α as: $\beta = 1 - \alpha$.

3 Feature Extraction from Ultrasound Medical Images

3.1 GLCM Based Statistical Texture Descriptors

The co-occurrence probabilities of GLCM provide a second-order method for generating texture features [2]. For extracting GLCM based texture features from ultrasound medical images, we obtained four co-occurrence matrices from each image using $\theta = \{0, 45, 90, 135\}$ degree and $d = 1$ pixel. After that, 19-statistical texture descriptors were calculated from each of these co-occurrence matrices as proposed in [2] and [3]. These are: Energy (F_1), Contrast (F_2), Correlation (F_3), Sum of Squares (F_4), Inverse Difference Moment (F_5), Sum Average (F_6), Sum Variance (F_7), Sum Entropy (F_8), Entropy (F_9), Difference Variance (F_{10}), Difference Entropy (F_{11}), Two Information Measures of Correlation (F_{12} and F_{13}), Maximal Correlation Coefficient (F_{14}), Autocorrelation (F_{15}), Dissimilarity (F_{16}), Cluster Shade (F_{17}), Cluster Performance (F_{18}), and Maximum Probability (F_{19}). Final values of these 19 descriptors were obtained by averaging each descriptor over the values calculated from the four matrices separately.

3.2 GLRLM Based Statistical Texture Descriptors

Let G be the number of grey levels, R be the longest run, and N be the number of pixels in the image. The GLRLM is a two dimensional matrix of $(G \times R)$ elements in which each element $p(i, j \mid \theta)$ gives the total number of occurrences of runs of length j of grey level i, in a given direction θ [4]. For extracting GLRLM based texture descriptors, ultrasound medical images were quantized using 64 grey levels. Then, four GLRLM matrices were calculated from each image using $\theta = \{0, 45, 90, 135\}$ degree. A total of 11 GLRLM based texture descriptors were calculated in our work from each of these 4 matrices. Five of these descriptors ($F_{20} - F_{24}$) were introduced in [4] and the other 6 descriptors were extended in [5] and [6]. These descriptors are: Short Runs Emphasis (F_{20}), Long Runs Emphasis (F_{21}), Grey Level Non-uniformity (F_{22}), Run Length Non-uniformity (F_{23}), Run

Percentage (F_{24}), Low Grey Level Runs Emphasis (F_{25}), High Grey Level Runs Emphasis (F_{26}), Short Run Low Grey-Level Emphasis (F_{27}), Short Run High Grey-Level Emphasis (F_{28}), Long Run Low Grey-Level Emphasis (F_{29}), Long Run High Grey-Level Emphasis (F_{30}). Final values of these 11 descriptors were obtained by averaging each descriptor over the values calculated from the four matrices separately.

4 Searching for Optimal Feature Subset Using GA

For selecting an optimal subset of texture descriptor from ultrasound medical images, 19 GLCM and 11 GLRLM based texture feature are extracted from the images of the training set. In order to reduce the impact of dynamic ranges of these two different groups of feature, they are first organized into a single feature vector and normalized as: $\hat{x}_k = (x_k - \mu_k)/\sigma_k$, where μ_k and σ_k are the mean and standard deviation of the k-th feature x_k. Initial Population size of Genetic Algorithm (GA) was set to 20. The chromosomes of the initial population were generated using the top 10 features returned by T-Statistics and Entropy based feature ranking method. Each individual chromosome represents a feature subset and was encoded as a M-bit binary vector as: $v = \{b_1, b_2, \cdots, b_M\}$, where M is the dimension of the feature vector. The value of b_i is either 0 or 1 depending on whether the corresponding feature has been selected or not. The fitness value for each chromosome is calculated according to the multi-objective criteria given by Eq. (4). Since, between-class separation is more important in image classification than that of within-class divergence, the parameters of the feature selection criteria, given by Eq. (4), were set to $\alpha = 0.70$ and $\beta = 0.30$ for emphasizing the between-class distance over within-class divergence. The parent selection process for offspring generation is implemented using the roulette wheel selection technique, where the probability of selecting a chromosome is calculated as: $P(C_i) = f(C_i) \Big/ \sum_{j=1}^{N} f(C_j)$; here, C_i is the chromosome to be considered for selection, $f(C_i)$ is the fitness value of C_i, and N is the population size. GA with traditional n-point crossover operation tends to explore mostly medium-sized subsets, while the edges of the non-dominated front are less well explored. To overcome this limitation, the Subset Size Oriented Common Features (SSOCF) crossover operator [7] has been adopted in our work which generates offspring populations with relatively even distribution. The mutation operation was modified to perform bit inversion mutation in both GLCM and GLRLM segment of the parent chromosome independently to generate an offspring. Here, the mutation probability P_m has a significant impact on the search ability as well as on the convergence speed of GA. Since the use of fixed P_m may lead towards premature convergence, the value of P_m was adjusted dynamically as [8]:

$$P_m = \begin{cases} P_{m0}/\log_2(g+1), & f \geq \bar{f} \\ P_{m0}, & f \leq \bar{f} \end{cases} \tag{5}$$

Here, P_{m0} is the initial mutation probability which is always between 0.1 and 0.25; f is the fitness of the parent chromosome, g is the current generation

number, and $\bar{f}$ is the average fitness of the current population. This dynamic adjustment process sets P_m to a higher value during the initial stage of evolution and thereby, strengthens the search ability towards finding the best individual solution. Probability of initial mutation was set to 0.15 and maximum no. of generation was limited to 200 to perform optimization using multi-objective GA. As Fig. 1 shows, the proposed multi-objective fitness criteria converges in both cases, with and without applying feature ranking during the initial population generation. However, incorporation of feature ranking during the initial population generation has significantly improved the convergence process. After convergence, the final optimal subset of texture descriptor consists of 14 features instead of 30 (described in Section 3), which are: $F = F_1, F_3, F_4, F_5, F_8, F_9, F_{10}, F_{16}, F_{18}, F_{24}, F_{27}, F_{28}, F_{29}, F_{30}$.

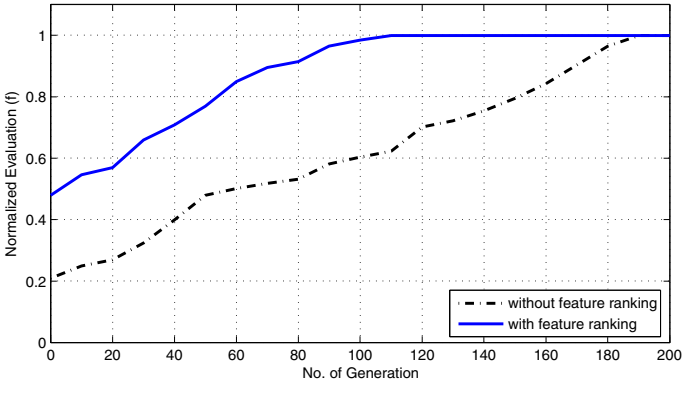


Fig. 1. Convergence of the proposed multi-objective feature selection criteria

5 Multi-class Classification Using Fuzzy-SVM

Let $S = \{(x_i, y_i)\}_{i=1}^{n}$ be a set of n training samples, where $x_i \in \Re^m$ is an m-dimensional sample in the input space, and $y_i \in \{-1, 1\}$ is the class label of x_i. SVM first maps the input data into a high-dimensional feature space through a mapping function $z = \phi(x)$ and finds the optimal separating hyperplane with the minimal classification errors. The hyperplane can be represented as: $w \cdot z + b = 0$, where w is the normal vector of the hyperplane, and b is the bias. The optimal hyperplane can be obtained by solving the following optimization problem [9]: Minimize $\frac{1}{2}\|w\|^2 + C\sum_{i}^{n}\xi_i$ Subject to $\gamma_i(w \cdot z_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \cdots, n$. Here, C is the regularization parameter that controls the tradeoff between margin maximization and classification error and ξ_i is called the slack variable that is related to classification errors in SVM. The optimization problem can be transformed into the following equivalent dual problem:

$$\text{Maximize } \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\gamma_i\gamma_j z_i \cdot z_j$$

$$\text{Subject to } \sum_{i=1}^{n} \gamma_i \alpha_i = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, \cdots, n \tag{6}$$

where α_i is the Lagrange multiplier. The decision function of SVM becomes:

$$f(x) = w \cdot z + b = \sum_{i=1}^{n} \alpha_i \gamma_i \phi(x_i) \cdot \phi(x) + b = \sum_{i=1}^{n} \alpha_i \gamma_i K(x, x_i) + b \tag{7}$$

where $K(x, x_i)$ is the kernel function in the input space that computes the inner product of two data points in the feature space. Research results have shown traditional SVM to be very sensitive to noises and outliners [10], [11]. FSVM is an extension of SVM that reduces to effects of outliers by taking into account the level of significance of various training samples. In FSVM, each training sample is assigned a fuzzy membership value $\{\mu_i\}_{i=1}^{n} \in [0, 1]$, which reflects the fidelity of the data; that is the level of confidence regarding the actual class information of the data. The higher the value of fuzzy membership, the more confident we are about its class label. Therefore, the training dataset of FSVM is represented as $S = \{(x_i, \mu_i, y_i)\}_{i=1}^{n}$. The optimization problem of the FSVM becomes [11]:

$$\begin{gathered} \text{Minimize } \tfrac{1}{2}\|w\|^2 + C \sum_{i}^{n} \mu_i \xi_i \\ \text{Subject to } \gamma_i (w \cdot z_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \cdots, n \end{gathered} \tag{8}$$

It can be noted that the error term ξ_i is scaled by the membership value μ_i. The fuzzy membership values are used to weigh the soft penalty term in the cost function of SVM. The weighted soft penalty term reflects the relative fidelity of the training samples during training. Important samples with larger membership values will have more impact in the FSVM training than those with smaller values. Similar to the conventional SVM, the optimization problem of FSVM can be transformed into its dual problem as follows:

$$\begin{gathered} \text{Maximize } \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \gamma_i \gamma_j K(x_i, x_j) \\ \text{Subject to } \sum_{i=1}^{n} \gamma_i \alpha_i = 0,\ \ \leq \alpha_i \leq \mu_i C,\ i = 1, \cdots, n \end{gathered} \tag{9}$$

Solving Eq. (9) will lead to a decision function similar to Eq. (7), but with different support vectors and corresponding weights α_i. Choosing appropriate fuzzy memberships for a given problem is very important for FSVM. To reduce the effect of outliers, the fuzzy membership function can be defined as a function of the distance between each data point and its corresponding class center [11]. Let X_+ and X_- denotes the mean of the training samples that belong to class $C^+(+1)$ and $C^-(-1)$ respectively. The radius of C^+ and C^- are given by: $r_+ = \max \|X_+ - X_i\| ; X_i \in C^+$ and $r_- = \max \|X_- - X_i\| ; X_i \in C^-$ respectively. The fuzzy membership function μ_i is defined as:

$$\mu_i = \begin{cases} 1 - \frac{\|X_+ - X_i\|}{(r_+ + \delta)} \text{ where, } X_i \in C^+ \\ 1 - \frac{\|X_- - X_i\|}{(r_- + \delta)} \text{ where, } X_i \in C^- \end{cases}$$

Here, $\delta > 0$ is a constant used to avoid any possible situation of $\mu_i = 0$. Multi-class classification is performed by arranging $\frac{q(q-1)}{2}$ binary FSVMs in "pair-wise" top down tree structured approach proposed in [12]. Here, q represents the number of classes in the training dataset, which is 4 for this particular problem.

Table 1. Comparison of classification performance achieved using different classifiers applying the proposed distance based feature selection method

	Simple C.	Endometrioma	Teratoma	Cancerous C.	Average
Fuzzy-SVM	92.58	89.41	87.84	88.57	89.60
SVM-RBF	91.41	87.65	85.81	85.71	87.64
SVM-Sigmoid	89.84	85.88	83.78	84.76	86.07
SVM-Polynomial	89.84	84.71	83.11	81.90	84.89
Fuzzy-kNN	91.41	88.82	87.16	86.67	88.51
Neural Network	86.72	81.76	79.73	79.05	81.82

Table 2. Comparison of classification performance without applying feature selection

	Simple C.	Endometrioma	Teratoma	Cancerous C.	Average
Fuzzy-SVM	88.67	85.29	81.76	83.81	84.88
SVM-RBF	86.33	83.53	80.41	81.90	83.04
SVM-Sigmoid	85.16	82.35	79.73	80.00	81.81
SVM-Polynomial	85.94	80.59	75.68	77.14	79.84
Fuzzy-kNN	87.50	85.88	81.08	82.86	84.33
Neural Network	80.08	76.47	75.00	74.29	76.46

6 Experimental Results

To investigate the performance of the proposed feature selection approach in ultrasound medical image classification, we conducted several experiments using a database of 679 ultrasound ovarian images of four categories: Simple Cyst (256 images), Endometrioma (170 images), Teratoma (148 images) and Cancerous Cyst (105). Feature vectors extracted from 300 images (100 simple cyst, 80 endometrioma, 70 teratoma and 50 cancerous cyst) of the database were used to train the classifiers applying "K-Fold Cross Validation" technique with $K = 5$. The choice of kernel function is among the most important customizations that can be made while adjusting an SVM classifier to a particular application domain. By performing experiments with FSVM using a range of Polynomial, Gaussian Radial Basis Function (RBF) and Sigmoid Kernels, we have found that Fuzzy-SVM with RBF kernel significantly outperforms the others, boosting the overall recognition accuracy. Average classification accuracy was adopted as the measure of classification performance in our experiments. For calculating the average classification accuracy, 20 test sets, each consists of randomly selected 30 test samples from 4 categories, were classified by the trained classifiers. The results of classification performance, both with and without feature selection, have been shown in Table 1 and Table 2.

7 Conclusion

We have presented a method of optimal feature subset selection using multi-objective GA. A weighted combination of between-class distance and within-class divergence has been introduced as the criteria of optimality in feature subset selection. Applying the proposed method, an optimal subset of GLCM and GLRLM based statistical texture descriptors has been identified for efficient representation and classification of ultrasound medical images. As the experimental results demonstrate, significant improvement can be achieved in classification performance (Table 1 and Table 2) when image representation is optimized using the best possible subset of image descriptors selected applying the proposed multi-objective separability criteria of feature selection. Our future plan is to adapt the proposed ultrasound image classification technique towards developing a Computer Aided Diagnosis system that will be able to provide decision support in the diagnosis of ovarian abnormalities.

References

1. Heijden, F., Duin, R., Ridder, D., Tax, D.M.J.: Classification, Parameter Estimation and State Estimation - An Engineering Approach Using MATLAB. John Wiley & Sons, Chichester (2004)
2. Haralick, R.M., Shanmugan, K., Dinstein, I.H.: Textural Features for Image Classification. IEEE Trans. Systems, Man and Cybernetics 3(6), 610–621 (1973)
3. Soh, L.-K., Tsatsoulis, C.: Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices. IEEE Trans. Geoscience Remote Sensing 37(2), 780–795 (1999)
4. Galloway, M.M.: Texture Analysis Using Gray Level Run Lengths. Computer Graphics Image Processing 4, 172–179 (1975)
5. Chu, A., Sehgal, C.M., Greenleaf, J.F.: Use of Gray Value Distribution of Run Lengths for Texture Analysis. Pattern Recognition Letters 11, 415–420 (1990)
6. Dasarathyand, B.R., Holder, E.B.: Image Characterizations Based on Joint Gray-Level Run-Length Distributions. Pattern Recognition Letters 12, 497–502 (1991)
7. Emmanouilidis, C., Hunter, A., MacIntyre, J.: A Multiobjective Evolutionary Setting for Feature Selection and a Commonality-Based Crossover Operator. In: Congress on Evolutionary Computing, San Diego, vol. 2, pp. 309–316 (2000)
8. Cai, Z., Xu, G.: Artificial intelligence: Principles and Applications. Tsinghua University Press, Beijing (2004)
9. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
10. Zhang, X.: Using Class-Center Vectors to Build Support Vector Machines. In: Neural Networks for Signal Processing IX, Madison WI, pp. 3–11 (August 1999)
11. Lin, C.F., Wang, S.D.: Fuzzy Support Vector Machines. IEEE Trans. Neural Networks 13(2), 464–471 (2002)
12. Platt, J., Cristianini, N., Shawe-Taylor, J.: Large Margin DAGs for Multiclass Classification. Advances in Neural Information Processing Systems 12, 547–553 (2000)

Ultrasound Plaque Enhanced Activity Index for Predicting Neurological Symptoms*

José Seabra[1], Luís Mendes Pedro[2],
José Fernandes e Fernandes[2], and João Sanches[1]

[1] Instituto de Sistemas e Robótica
Instituto Superior Técnico
1049-001 Lisboa, Portugal
[2] Instituto Cardiovascular de Lisboa
Faculdade de Medicina de Lisboa
1649-028 Lisboa, Portugal

Abstract. This paper aims at developing an ultrasound-based diagnostic measure which quantifies plaque activity, that is, the likelihood of the asymptomatic lesion to produce neurological symptoms. The method is rooted on the identification of an "active" plaque profile containing the most relevant ultrasound parameters associated with symptoms. This information is used to build an Enhanced Activity Index (EAI) which considers the conditional probabilities of each relevant feature belonging to either symptomatic or asymptomatic groups. This measure was evaluated on a longitudinal study of 112 asymptomatic plaques and shows high diagnostic power. In particular, EAI provides correct identification of all plaques that developed symptoms while giving a small number of false positives. Results suggest that EAI could have a significant impact on stroke prediction and treatment planning.

Keywords: Ultrasound, Carotid plaque, Enhanced Activity Index, Neurological symptoms.

1 Introduction

Carotid plaques are one of the commonest causes of neurological symptoms due to embolization of plaque components or flow reduction. Numerous studies report that plaque morphology, besides patient's clinical history and degree of stenosis, is an important ultrasound marker that positively correlates with symptoms [1,6]. However, such studies are focused on classifying plaques as symptomatic or asymptomatic and very few aim at identifying those stable lesions at high risk of becoming symptomatic. In fact, this information would be extremely useful for the physicians since they would be able to observe an asymptomatic lesion and quantitatively evaluate if such lesion is prone to developing neurological complications. As a consequence, the identification of a subset of "dangerous"

* This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 184–191, 2011.

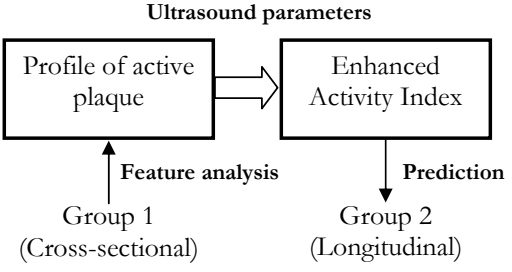


Fig. 1. Framework of proposed method

or "active" plaques, featuring high neurological risk would help in the indication of treatment. Needless to say, this decision has important clinical and economical consequences for all the parts involved in this process.

Major trials reported that the absolute benefit of surgical intervention based on the degree of stenosis alone as a decision making criterion is low in the asymptomatic disease and in symptomatic disease with moderate obstruction [2,4]. This clearly motivates the need for developing new strategies for plaque risk prediction. One such strategy [6] aims at combining quantitative (e.g. the degree of stenosis and histogram features) and qualitative information (e.g. textural appearance) obtained from ultrasound B-mode images. This study enabled to develop a diagnostic score, called Activity Index (AI), which could possibly correlate with clinical findings. Statistical analysis identified the most significant parameters as being the grey-scale median (GSM), percentage of echolucent (GSM<40) pixels (P40), surface disruption, severe stenosis, plaque heterogeneity and presence of juxta-luminal echolucent area. Hence, the AI consists of summing the scores for each significant variable. Results suggest that AI is an objective technique to assess plaque instability [6].

This paper uses ultrasound image processing as a first step for predicting the occurrence of plaque symptoms. In recent years, the importance of speckle in B-mode ultrasound images as well as its statistical modeling for tissue characterization has been reported [9]. This issue was also suggested in a recent work [7] where the application of a de-speckling algorithm was able to split the image in its noiseless and speckle components. These image sources were then used for extracting distinct echo-morphology and texture parameters, which contributed to a better analysis of the symptomatic plaque, and differentiation from the asymptomatic lesion [7].

Here, it is argued that an optimal method for identifying vulnerable lesions should include morphological and textural features, extracted from pixel intensity information, and information regarding plaque structure and appearance (e.g. stenosis, evidence of surface disruption and presence of echogenic cap) given by experienced physicians. The combination of this information is expected to produce a more comprehensive description of the profile of an active plaque, potentially providing the identification of lesions that would developed symptoms in the future. This paper proposes a diagnostic tool, named Enhanced Activity

Index (EAI), which uses an ultrasound feature set that positively correlates with symptoms. Here, the EAI technique is used to predict the occurrence of neurological complications in a longitudinal study conducted in asymptomatic subjects. Moreover, its diagnostic performance is compared to other well-established strategies for identifying plaques at high risk, including the degree of stenosis (DS) [4] and the AI [6].

2 Methods

This paper employs a two-step method outlined in Fig. 1. The first step consists of feature analysis which identifies an optimal feature set to discriminate between symptomatic and asymptomatic plaques. This step is based on ultrasound images of carotid plaques ($n = 221$, 70 symptomatic and 151 asymptomatic) acquired at a fixed time frame (cross-sectional study). Consequently, the ultrasound profile of the "active" plaque is used to compute the EAI in a longitudinal study conducted in asymptomatic subjects ($n = 112$).

2.1 Ultrasound Profile of "Active" Plaque

Prior to feature analysis, ultrasound images are processed according to [7]. Image processing includes normalization [3], estimation of envelope Radio-Frequency (RF) image and de-speckling. Moreover, each region of interest (ROI) containing the plaque is delineated by an experienced physician. This processing step provides different image sources from where features having different meanings can be extracted (Fig. 2):

- **Histogram features**, computed from pixel intensities in normalized image;
- **Rayleigh mixture models**, are used to describe echo-morphology in envelope RF images [8], consisting of a combination of individual Rayleigh distributions. The weights of each distribution and corresponding parameter are used as echo-morphology descriptors [8];
- **Rayleigh parameters** are theoretically obtained from the noiseless image;
- **Textural features** are obtained from grey level co-occurrence matrices (GLCM), Autoregressive models (ARM) and Wavelet models;
- **Morphological features** are given by the physician during consultation (e.g. evidence of plaque disruption, presence of fibrous cap, degree of stenosis and plaque echo-structure appearance).

A considerable amount of features ($l = 114$) were collected after ultrasound image processing. Naturally, not all the features are important to accurately characterize the plaque status, whether it is symptomatic or not. Hence, at this point an attempt is made to identify the most relevant ultrasound parameters for this particular problem. Hypothesis testing is a common method of drawing inferences about one or more populations based on statistical evidences from population samples (features). Here, we want to investigate if the statistical properties of a given feature significantly differ from the symptomatic to the

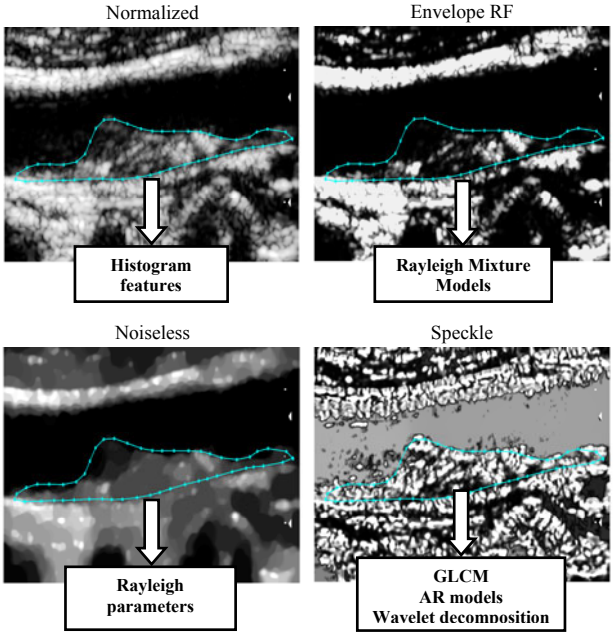


Fig. 2. Ultrasound image processing, resulting in normalized, envelope RF, noiseless and speckle images. Features are extracted from these distinct image sources.

asymptomatic group. Among different hypothesis tests, the application of the Mann-Whitney U-test yields a feature set that produces the most promising classification results with the AdaBoost classifier. This method performs a two-sided rank sum test of the null hypothesis that feature values in symptomatic and asymptomatic populations are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians. Moreover, the *p-value* is the probability of rejecting the null hypothesis assuming that the null hypothesis is true. Clinically significant features will have a *p-value* which is typically lower than 0.05 or 0.01. In this work, features were considered to be relevant for differentiating between symptomatic and asymptomatic groups when the $p\text{-}value < 0.05$.

Table 1 presents the parameters and corresponding sources and *p-values* of the so-called *optimal feature set*. A closer look at the 16-element feature set allows to verify that both subjective and image-based parameters are useful for plaque description. In particular, features from different image sources, namely the normalized image, the envelope RF image and speckle field are considered statistically relevant. This preliminary observation justifies the use of an ultrasound pre-processing set of operations since it enables to obtain useful parameters for plaque classification.

Table 1. Significant feature set obtained with MW-test

Feature	Source	$p - value$
evidence of plaque disruption	morphology	5.5×10^{-18}
presence of echogenic cap	morphology	0.001
degree of stenosis	morphology	2.9×10^{-13}
plaque echo-structure appearance	morphology	1.6×10^{-10}
mean	normalized histogram	0.001
skewness	normalized histogram	0.009
percentile 10	normalized histogram	0.022
percentile 50	normalized histogram	0.047
4^{th} Rayleigh parameter	RMM (envelope RF)	0.010
5^{th} Rayleigh parameter	RMM (envelope RF)	0.010
6^{th} Rayleigh parameter	RMM (envelope RF)	0.010
5^{th} mixture component	RMM (envelope RF)	0.004
6^{th} mixture component	RMM (envelope RF)	0.014
no. mixture components	RMM (envelope RF)	0.016
GLCM homogeneity	speckle	0.016
wavelet decomposition energy	speckle	0.004

2.2 Enhanced Activity Index

So far the profile of the active plaque has been established. A quantitative diagnostic measure - EAI - is now developed as follows (Fig. 3):

1. A statistical test allows to obtain a relevant feature set (Table 1) for separating plaques with and without symptoms in a cross-sectional study (Fig. 1);
2. Reference values are taken for each feature, f_i, and group (symptomatic, $\omega(S)$, and asymptomatic, $\omega(A)$), considering the mean ($\mu_i(S)$, $\mu_i(A)$) and variance ($\sigma_i^2(S)$, $\sigma_i^2(A)$);
3. The EAI* (re-named for convenience) is computed as:

$$\text{EAI}^* = \frac{\text{R}_S}{\text{R}_A}, \tag{1}$$

where

$$\text{R}_k = \sum_i p(f_i|\omega_k) \approx \mathcal{N}(\mu_i(k), \sigma_i^2(k)), \quad k = \{S, A\} \tag{2}$$

is the sum of the conditional probabilities of each feature belonging, respectively, to the symptomatic or asymptomatic group. For continuous variables, the conditional probabilities in (2) are computed assuming a normal distribution (Fig. 3) whereas the probability associated with each categorial parameter (morphological features, except for degree of stenosis) is given by the ratio between the number of plaques belonging to each class and having each categorial variable and the total number of plaques with each categorial variable. In (1), R_S and R_A represent the likelihoods of each plaque producing symptoms or stabilize, respectively. Hence, when $\text{EAI}^* = 1$, the result is inconclusive while for $\text{EAI}^* < 1$ the plaque will stay harmless with a significant probability which is higher as EAI* decreases. Contrarily, plaques showing a $\text{EAI}^* > 1$ are prone to produce symptoms, being more "dangerous" when EAI increases.

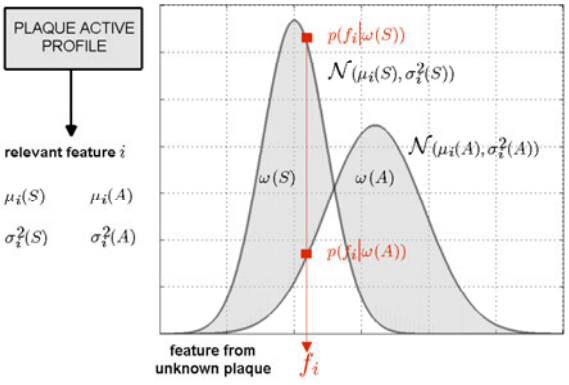


Fig. 3. Illustrative concept of conditional probabilities for a particular plaque feature f_i used to compute EAI

4. The EAI* is re-scaled using a sigmoid mapping function which places the EAI onto a 0-100 scale. This function is defined as:

$$\text{EAI} = \frac{100}{1 + \exp\left(1 - \text{EAI}^*\right)}. \tag{3}$$

This mapping technique is useful to make the predictive power of the proposed EAI method comparable to AI [6] and DS [4].

3 Experimental Results

We have provided a description of the ultrasound profile of the active plaque and consequently designed a score for predicting the occurrence of neurological complications. Here, the diagnostic power of EAI is evaluated on a group of 112 asymptomatic plaques, acquired from 112 patients. B-mode ultrasound images were collected from the ACSRS (Asymptomatic Carotid Stenosis and Risk Study) [5], consisting of a multicentre natural history study of patients with asymptomatic internal carotid diameter stenosis greater than 50%. The degree of stenosis was graded using multiple established ultrasonic duplex criteria. Patients were followed for possible occurrence of symptoms for a mean time interval of 37.1 weeks. At the end of the study, 13 out of 112 patients (11.6%) had developed symptoms.

To make this study more feasible, we compare it with other strategies of plaque risk prediction, including the AI and DS, using ROC (Receiver Operating Characteristic) curve analysis (Fig. 4). In a ROC curve, the TP rate (Sensitivity) is plotted as function of the FP rate (100-Specificity) for different cut-off points. Each point of the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. Moreover, the area under the ROC curve (ROC AUC) statistic is often used for model comparison. This measure indicates

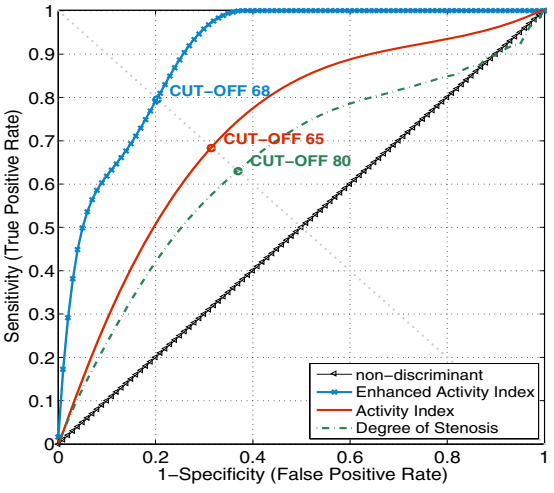


Fig. 4. ROC curves of degree of stenosis (DS), AI and EAI

Table 2. Confusion matrix with prediction outcome of DS, AI and EAI

		Ground truth					
		P			N		
		Sten	AI	EAI	Sten	AI	EAI
Predicted value	P'	12	11	**13**	68	49	**29**
	N'	1	2	**0**	31	50	**70**

that a predictive method is more accurate as higher is the ROC AUC. Moreover, the intercept of the ROC curve with the line at 90 degrees to the no-discrimination line is also considered as an heuristic method to investigate the cut-off providing the best discriminative power for each method (Fig. 4). The ROC AUCs are 0.6496 (64.96%), 0.7329 (73.29%) and 0.9057 (90.57%) for DS, AI and EAI, respectively. These results show that the EAI technique outperforms the other methods, which could be explained by the fact that it considers ultrasound parameters used in AI, besides the DS.

Moreover, the predictive analysis of EAI, AI and DS is evaluated from a different viewpoint, using a table of confusion (Table 2). The EAI method is able to identify the 13 patients that had developed symptoms by the end of the follow-up (longitudinal) study, whereas DS and AI methods were unable to identify, respectively 1 and 2 patients that developed neurological complications later. Moreover, as far as the false positive number is concerned, the EAI method yields 29 FP, which is significantly lower than other state-of-the-art methods. This means that if the decision of surgery for plaque removal was based in the former method only 29 patients were unnecessarily operated, suggesting that EAI is the most cost-effective method. Thus, the EAI technique demonstrates to provide the most accurate selection of patients at high risk within a population of asymptomatic subjects.

4 Conclusions

Carotid plaques are the commonest cause of neurological symptoms, however the early detection of plaques at high risk based on the degree of stenosis is neither optimal nor cost-effective. This paper proposes an Enhanced Activity Index technique which quantifies the likelihood of a stable plaque to becoming symptomatic. This method consists of a two-step method including the identification of an ultrasound profile of the active plaque and the computation of a risk score. The proposed prediction method provides an effective selection of a subgroup of plaques at high risk of developing symptoms.

References

1. Christodoulou, C., Pattichis, C., Pantziaris, M., Nicolaides, A.: Texture-based classification of atherosclerotic carotid plaques. IEEE Transactions on Medical Imaging 22(7) (2003)
2. ECST Collaborative Group: Randomised trial of endarterectomy for recently symptomatic carotid stenosis: Final results of the MRC ECST. Lancet 351, 1379–1387 (1998)
3. Elatrozy, T., Nicolaides, A., Tegos, T., Griffin, M.: The objective characterization of ultrasonic carotid plaque features. Eur. J. Vasc. Endovasc. Surg. 16, 223–230 (1998)
4. NASCET Collaborators: Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. New England Journal of Medicine 339(20), 1445–1453 (1998)
5. Nicolaides, A., Kakkos, S., Griffin, M., Sabetai, M., Dhanjil, S., et al.: Severity of Asymptomatic Carotid Stenosis and Risk of Ipsilateral Hemispheric Ischaemic Events: Results from the ACSRS Study. European Journal of Vascular & Endovascular Surgery 30(3), 275–284 (2005)
6. Pedro, L., Fernandes, J., Pedro, M., Gonçalves, I., Dias, N.: Ultrasonographic risk score of carotid plaques. European Journal of Vascular and Endovascular Surgery 24, 492–498 (2002)
7. Seabra, J., Pedro, L., Fernandes, J., Sanches, J.: Ultrasonographic Characterization and Identification of Symptomatic Carotid Plaques. In: Proceedings of IEEE International Conference on Engineering in Medicine and Biology. IEEE EMBS, Buenos Aires (2010)
8. Seabra, J., Sanches, J., Ciompi, F., Radeva, P.: Ultrasonographic plaque characterization using a rayleigh mixture model. In: Proceedings of IEEE International Symposium on Biomedical Imaging, pp. 1–4. IEEE Engineering in Medicine and Biology Society, Rotterdam (2010)
9. Thijssen, J.: Ultrasonic speckle formation, analysis and processing applied to tissue characterization. Pattern Recognition Letters 24(4-5) (February 2003)

On the Distribution of Dissimilarity Increments

Helena Aidos and Ana Fred

Instituto de Telecomunicações, Instituto Superior Técnico, Lisbon, Portugal
{haidos,afred}@lx.it.pt

Abstract. This paper proposes a statistical model for the dissimilarity changes (increments) between neighboring patterns which follow a 2-dimensional Gaussian distribution. We propose a novel clustering algorithm, using that statistical model, which automatically determines the appropriate number of clusters. We apply the algorithm to both synthetic and real data sets and compare it to a Gaussian mixture and to a previous algorithm which also used dissimilarity increments. Experimental results show that this new approach yields better results than the other two algorithms in most datasets.

Keywords: clustering, dissimilarity increments, likelihood ratio test, Gaussian mixture.

1 Introduction

Clustering techniques are used in various application areas, namely in exploratory data analysis and data mining [4]. Also known as unsupervised classification of patterns into groups (clusters), the aim is to find a data partition such that patterns belonging to the same cluster are somehow "more similar" than patterns belonging to distinct clusters. Clustering algorithms can be partitional or hierarchical, and can use a multitude of ways to measure the (dis)similarity of patterns [4,7].

Partitional methods assign each data pattern to exactly one cluster; the number of clusters, K, is usually small, and often set *a priori* by the user, as a design parameter. Otherwise, the choice of K may be addressed as a model selection problem. The most iconic partitional algorithm is also the most simple: K-means, using the centroid as cluster representative, attempts to minimize a mean-square error criterion based on the Euclidean distance as measure of pairwise dissimilarity [7]. Also, common methods to estimate probability density functions from data, such as Gaussian mixture decomposition algorithms [1], can also be used as clustering techniques.

Hierarchical methods, on the other hand, yield a set of nested partitions which is graphically represented by a dendrogram. A data partition is obtained by cutting the dendrogram at a certain level. Linkage algorithms, such as the single-link and the complete-link [4], are the most commonly used.

Fred and Leitão [3] have proposed a hierarchical clustering algorithm using the concept of *dissimilarity increments*. These increments, which are formally defined in Section 2, use three data patterns at a time, and therefore yield information that goes beyond pairwise dissimilarities. Fred and Leitão showed empirical evidence suggesting that dissimilarity increments vary smoothly within a cluster, and proposed an exponential distribution as statistical model governing the dissimilarity increments in each

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 192–199, 2011.

cluster [3]. They also noted that abrupt changes in the increments values means that the merging of two well separated clusters should not occur.

In this paper we propose a novel dissimilarity increments distribution (DID), supported on a theoretical-based analytical derivation for Gaussian data in $\mathbb{R}^2$. We use this distribution to construct a partitional clustering algorithm that uses a split&merge strategy, which iteratively accepts or rejects the merging of two clusters based on the distribution of their dissimilarity increments. We apply this algorithm to 6 synthetic data sets and 5 real data sets using as starting condition the clusters yielded by a Gaussian mixture algorithm proposed by Figueiredo and Jain [1], although any Gaussian mixture algorithm could be used instead.

This paper is structured as follows: Section 2 presents the derivation of the dissimilarity increments distribution (DID), and in Section 3 we propose a rewriting of the latter that depends on a single parameter: the expected value of increments. In Section 4, we show how to use this DID in a clustering algorithm. We present, in Section 5, the results of the proposed algorithm for 6 synthetic data sets with different characteristics (gaussian clusters, non-gaussian clusters, arbitrary shape clusters and densities) and 5 real data sets from the UCI Machine Learning Repository. These results are compared with the initial Gaussian mixture decomposition and with the hierarchical clustering algorithm proposed by Fred and Leitão in [3]. Conclusions are drawn in Section 6.

2 Dissimilarity Increments Distribution for 2D Gaussian Data

Consider a set of patterns, X. Given $\mathbf{x}_i$, an arbitrary element of X, and some dissimilarity measure between patterns, $d(\cdot,\cdot)$, let $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ be the triplet of nearest neighbors, where $\mathbf{x}_j$ is the nearest neighbor of $\mathbf{x}_i$ and $\mathbf{x}_k$ is the nearest neighbor of $\mathbf{x}_j$ different from $\mathbf{x}_i$. The dissimilarity increment [3] between the neighboring patterns is defined as

$$d_{inc}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = |d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k)| \,. \tag{1}$$

Assume that $X \in \mathbb{R}^2$, and that elements of X are independent and identically distributed, drawn from a normal distribution, $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. In this paper, the Euclidean distance is used as the dissimilarity measure, and our goal is to find the probability distribution of d_{inc}.

Let X^* be the result of an affine transformation of the patterns in X such that the transformed data has zero mean and the covariance matrix is a multiple of the identity matrix. Define $\mathbf{D}^* \equiv \mathbf{x}^* - \mathbf{y}^*$ as the dissimilarity in this space. Then the distribution of dissimilarities in this space becomes

$$(D^*)^2 \equiv \|\mathbf{x}^* - \mathbf{y}^*\|^2 = \sum_{i=1}^{2} \frac{(x_i^* - y_i^*)^2}{\Sigma_{ii}^*} \sim \chi^2(2),$$

where $\chi^2(2)$ is the chi-square distribution with 2 degrees of freedom, which is equivalent to an exponential distribution with parameter 1/2 [5].

Since the transformed data has circular symmetry, we have $\mathbf{D}^* = D^* \cos\alpha\, \mathbf{e}_1 + D^* \sin\alpha\, \mathbf{e}_2$, with $\alpha = \text{angle}(\mathbf{D}^*) \sim Unif([0, 2\pi[)$. Furthermore, $\mathbf{D} \equiv \mathbf{x} - \mathbf{y} = \sqrt{\Sigma^*_{11}}\, D^* \cos\alpha\, \mathbf{e}_1 + \sqrt{\Sigma^*_{22}}\, D^* \sin\alpha\, \mathbf{e}_2$, and

$$D^2 \equiv \|\mathbf{D}\|^2 = \underbrace{(\Sigma^*_{11} \cos^2\alpha + \Sigma^*_{22} \sin^2\alpha)}_{A(\alpha)^2} \underbrace{(D^*)^2}_{\|\mathbf{D}^*\|^2}, \tag{2}$$

where $A(\alpha)^2$ is called the expansion factor. Naturally this expansion factor will depend on the angle α. In practice it is hard to properly deal with this dependence. Therefore we will use the approximation that the expansion factor is constant and equal to the average value of the true expansion factor. We must find $\mathbb{E}[A(\alpha)^2]$, where $\alpha \sim Unif([0, 2\pi[)$ and $p_\alpha(\alpha) = \frac{1}{2\pi}$. After some computations, the expected value is given by

$$\mathbb{E}[A(\alpha)^2] = \int_0^{2\pi} p_\alpha(\alpha) A(\alpha)^2 \, \mathrm{d}\alpha = \frac{1}{2} \operatorname{tr}(\Sigma^*).$$

Under this approximation, the transformation equation (2) from the normalized space to the original space is $D^2 = \frac{1}{2} \operatorname{tr}(\Sigma^*)(D^*)^2$ and the probability density function of $D = d(\mathbf{x}, \mathbf{y})$ is (recall that $(D^*)^2 \sim \text{Exp}(1/2)$)

$$p_{_D}(z) = \frac{2z}{\operatorname{tr}(\Sigma^*)} \exp\left(-\frac{z^2}{\operatorname{tr}(\Sigma^*)}\right), \; z \in [0, \infty). \tag{3}$$

We can conclude that $D_1 = d(\mathbf{x}, \mathbf{y})$ and $D_2 = d(\mathbf{y}, \mathbf{z})$ follow the distribution in equation (3). The probability density function for $W = D_1 - D_2$ is given by the convolution

$$p_{_W}(w) = \int_{-\infty}^{\infty} \frac{4t(t+w)}{\operatorname{tr}(\Sigma^*)^2} \exp\left(-\frac{t^2 + (t+w)^2}{\operatorname{tr}(\Sigma^*)}\right) \mathbf{1}_{\{t \geq 0\}} \, \mathbf{1}_{\{t+w \geq 0\}} \, \mathrm{d}t. \tag{4}$$

Since we want to find the probability density function for the dissimilarity increments, we need to consider the probability density function of $|W| = d_{inc}$. Therefore, the probability density function for the dissimilarity increments is given by (derivations were omitted due to limited space)

$$p_{d_{inc}}(w; \Sigma^*) = \frac{w}{\operatorname{tr}(\Sigma^*)} \exp\left(-\frac{w^2}{\operatorname{tr}(\Sigma^*)}\right) + \frac{\sqrt{\pi}}{\sqrt{2}\left(\operatorname{tr}(\Sigma^*)\right)^{3/2}} (\operatorname{tr}(\Sigma^*) - w^2) \times$$
$$\times \exp\left(-\frac{w^2}{2\operatorname{tr}(\Sigma^*)}\right) \operatorname{erfc}\left(\frac{w}{\sqrt{2\operatorname{tr}(\Sigma^*)}}\right), \tag{5}$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function.

3 Empirical Estimation of DID

The DID, as per equation 5, requires explicit calculation of the covariance matrix, Σ^*, in the transformed normalized space. In the sequel we propose data model fitting by

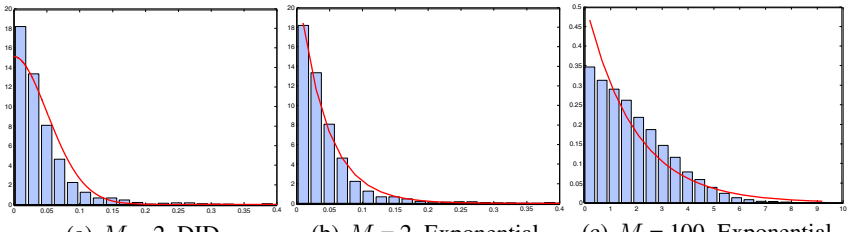
rewritting the distribution as a function of the mean value of the dissimilarity increments, $\lambda = \mathbb{E}[w]$. This is given by (after some calculation)

$$\lambda = \mathbb{E}[w] = \int_0^\infty w p_w(w)\, \mathrm{d}w = \frac{\sqrt{\pi}}{2} \left(\mathrm{tr}(\Sigma^*)\right)^{1/2} \left(2 - \sqrt{2}\right).$$

Hence, $\left(\mathrm{tr}(\Sigma^*)\right)^{1/2} = \frac{2\mathbb{E}(w)}{\sqrt{\pi}\left(2-\sqrt{2}\right)}$. Replacing in (5) we obtain an approximation for the dissimilarity increments distribution of a cluster that only depends of the mean of all the increments in that cluster:

$$p_{d_{inc}}(w;\lambda) = \frac{\pi\left(2-\sqrt{2}\right)^2}{4\lambda^2}\, w\, \exp\left(-\frac{\pi\left(2-\sqrt{2}\right)^2}{4\lambda^2}\, w^2\right) + \frac{\pi^2\left(2-\sqrt{2}\right)^3}{8\sqrt{2}\lambda^3} \times$$
$$\times \exp\left(-\frac{\pi\left(2-\sqrt{2}\right)^2}{8\lambda^2}\, w^2\right)\left(\frac{4\lambda^2}{\pi\left(2-\sqrt{2}\right)^2} - w^2\right) \mathrm{erfc}\left(\frac{\sqrt{\pi}\left(2-\sqrt{2}\right)}{2\sqrt{2}\lambda}\, w\right). \tag{6}$$

Figure 1 provides histograms for two data sets consisting of 1000 samples drawn from a Gaussian distribution in dimensions $M = 2$ and $M = 100$. As shown in figures 1(a) and 1(b), for $M = 2$ both the derived probability density function (6) and the exponential distribution suggested in [3] lead to a good fit to the histogram of the dissimilarity increments. However, as figure 1(c) shows, the latter provides a poor fit for high dimensions, such as $M = 100$, while the proposed distribution, even though derived for the 2D case, is much more adequate.



(a) M = 2, DID (b) M = 2, Exponential (c) M = 100, Exponential

Fig. 1. Histograms of the dissimilarity increments computed over M-dimensional Gaussian data and fitted distribution: (a) DID; (b) and (c) Exponential distribution

4 Clustering Using the Dissimilarity Increments Distribution

Let P^{init} be a partitioning of the data produced by a gaussian mixture decomposition. Then, each cluster in P^{init} follows a gaussian model; if $\mathbf{x}_i \in \mathbb{R}^2$, we are under the underlying hypothesis of the model presented in Section 2. One of the main difficulties of the gaussian mixture decomposition is the inability to identify arbitrarily shaped clusters. We propose to overcome this intrinsic difficulty by a process of merging clusters

using the previously derived model for dissimilarity increments. The decision to merge two clusters will depend on the dissimilarity increments distribution of each of the two clusters separately and of the two clusters combined.

We use the Mahalanobis distance [7] (which computes the distance between two gaussian distributions) to decide which clusters to test, by testing first clusters that are closer. The test we perform is a likelihood ratio test [6] consisting on the logarithm of the ratio between the likelihood of two separate clusters (two DID models) and the likelihood of the clusters together (single DID model). This likelihood ratio test is approximated by a chi-square distribution with one degree of freedom[1]. Therefore,

$$-2\log\left(\frac{p_{d_{inc}}(w;\lambda_1,\lambda_2)}{p_{d_{inc}}(w;\lambda_{12})}\right) \sim \chi^2(1). \tag{7}$$

Two clusters are merged if the p-value from the $\chi^2(1)$ distribution is less than a significance level α. This test is performed for all pairs of clusters until all the clusters that remain, when tested, are determined not to be merged. The overall procedure of this algorithm is summarized in algorithm 1.

Algorithm 1. GMDID

Input: 2-dimensional data, α
$P^{init} = \{C_1, \ldots, C_N\} \leftarrow$ data partition produced by a gaussian mixture decomposition
$D_{ij} \leftarrow$ Mahalanobis distance between clusters i and j
for all pairs (i, j) in ascending order of D_{ij} **do**
  $p_i \leftarrow$ DID for cluster i, $p_j \leftarrow$ DID for cluster j (eq. 6)
  $p_{ij} \leftarrow$ DID for cluster produced merging clusters i and j (eq. 6)
  p-value $\leftarrow$ Likelihood ratio test between $p_i p_j$ and p_{ij} (eq. 7)
  if p-value $< \alpha$ **then**
    merge clusters i and j
  else
    do not merge clusters i and j
  end if
end for
Return: $P = \{C_1, \ldots, C_K\} \leftarrow$ final data partition $K \leq N$

4.1 Graph-Based Dissimilarity Increments Distribution

In order to choose the parameter α, we propose to use the Graph-based Dissimilarity Increments Distribution index, hereafter designated by G-DID, which is a cluster validity index proposed by Fred and Jain [2] based on the minimum description length (MDL) of the graph-based representation of partition P. The selection among N partitions, produced by different values of α, using the G-DID index, is as follows

$$\text{Choose } P^i : i = \underset{j}{\operatorname{argmin}}\{\text{G-DID}(P^j)\}, \tag{8}$$

[1] We have two parameters in the numerator – the expected value of the increments for each of the two clusters separately – and one parameter in the denominator – the expected value of the increments for the two clusters combined.

where G-DID$(P) = -\log \hat{f}(P) + \frac{k_P}{2}\log(n)$ is the graph description length, and $\hat{f}(P)$ is the probability of partition P, with k_P clusters, according to a Probabilistic Attributed Graph model taking into account the dissimilarity increments distribution (see [2] for details). In [2], graph edge probability was estimated from an exponential model associated with each cluster, and the G-DID of the corresponding partition was used to select a design parameter of the hierarchical algorithm in [3]. For the selection of α for the GMDID algorithm describd above, we will use instead the DID model according to formula 6.

5 Experimental Results and Discussion

We can use any Gaussian mixture algorithm to obtain the conditions of the proposed clustering method; we chose the algorithm proposed by Figueiredo and Jain [1]. This method optimizes an expectation-maximization (EM) algorithm and selects the number of components in an unsupervised way, using the MDL criterion. With this Gaussian mixture algorithm we get a partition for the data set with as many clusters as gaussians.

To test the performance of the proposed method, we used 11 data sets: 6 synthetic data sets, and 5 real data sets from the UCI Machine Learning Repository[2]. The synthetic data sets were chosen to take into account a wide variety of situations: well-separated and touching clusters; gaussian and non-gaussian clusters; arbitrary shapes; and diverse cluster densities. These synthetic data sets are shown in figure 2. The *Wisconsin Breast-Cancer* data set consists of 683 patterns represented by nine features and has two clusters. The *House Votes* data set consists of votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the Congressional Quarterly Almanac. It is composed by two clusters and only the patterns without missing values were considered, for a total of 232 samples (125 democrats and 107 republicans). The *Iris* data set consists of three species of Iris plants (Setosa, Versicolor and Virginica). This data set is characterized by four features and 50 samples in each cluster. The *Log Yeast* and *Std Yeast* is composed of 384 samples (genes) over two cell cycles of yeast cell data. Both data sets are characterized by 17 features and consisting of five clusters corresponding to the five phases of the cell cycle.

We compared the proposed method (GMDID) to the Gaussian mixture (GM) algorithm [1] used to initialize the dissimilarity increments distribution, and to the method proposed by Fred and Leitão [3] based on Dissimilarity Increments (SL-AGLO). All these methods find the number of clusters automatically. GMDID and SL-AGLO have an additional parameter, so we compute partitions for several values of parameters. GMDID has a significance level α and we used 1%, 5%, 10% and 15% to decide whether two clusters should be merged or not. The SL-AGLO algorithm has an isolation parameter which is a threshold set in the tail of the exponential distribution of the dissimilarity increments of a cluster (with parameter the inverse of the mean of the increments of that cluster). We used values ranging from the mean of the exponential distribution to 10 times this mean for this threshold, and the choice of the best value was also done using G-DID.

[2] `http://archive.ics.uci.edu/ml`

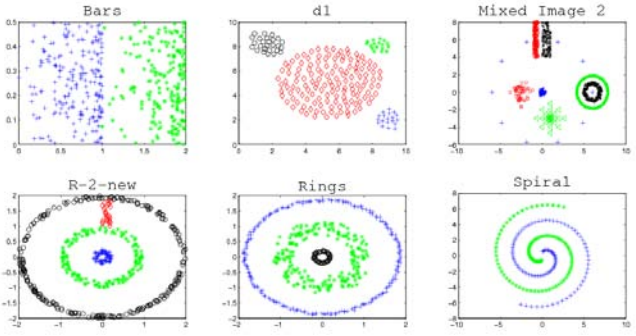


Fig. 2. Synthetic data sets

We assess the quality of each resulting partition P using the consistency index (CI), which is the percentage of correctly clustered patterns. Table 1 summarizes the results. The Gaussian mixture algorithm has problems finding the correct number of clusters, if the data sets are non-gaussians or can not be approximated by a single Gaussian. However, GMDID improved the results given by the Gaussian mixture, because it depends of the dissimilarity changes between neighboring patterns. If a cluster does not have a Gaussian behavior and the Gaussian mixture produces at least two gaussian components for that cluster, it may be possible that GMDID can find the cluster, for a certain statistical significance level, by merging those components together.

Table 1. Consistency values of the partitions found by the three algorithms. The values in parenthesis correspond to the number of clusters found by each algorithm. The first two columns correspond to the number of patterns (N) and the true number of clusters (Nc) of each data set.

	N	Nc	GM	GMDID	SL-AGLO
Bars	400	2	0.4375 (12)	**0.9525** (2)	0.9050 (4)
d1	200	4	**1.0000** (4)	**1.0000** (4)	**1.0000** (4)
Mixed Image 2	739	8	0.4709 (20)	**1.0000** (8)	0.9743 (10)
R-2-new	500	4	0.2880 (29)	**0.7360** (8)	0.6160 (3)
Rings	450	3	0.2444 (27)	**1.0000** (3)	**1.0000** (3)
Spiral	200	2	0.1400 (27)	**1.0000** (2)	**1.0000** (2)
Breast Cancer	683	2	0.5593 (5)	**0.7467** (3)	0.5783 (24)
House Votes	232	2	**0.8103** (2)	**0.8103** (2)	0.6810 (4)
Iris	150	3	**0.8000** (4)	0.6667 (2)	0.4800 (6)
Log yeast	384	5	0.3281 (10)	**0.3594** (4)	0.3229 (4)
Std yeast	384	5	0.4349 (2)	0.4349 (2)	**0.5391** (7)

Despite the fact that the dissimilarity increments distribution proposed here is for 2-dimensional Gaussian distributions, we noticed that when the algorithm is applied to real data sets (which have dimensions higher than two) the results are still slightly

better when compared to the Gaussian mixture and to SL-AGLO. In the future we will develop the dissimilarity increments distribution to a general M-dimensional data set, which should further improve these results.

6 Conclusions

We derived the probability density function for the dissimilarity increments under the assumption that the clusters follow a 2-dimensional Gaussian distribution. We applied this result by proposing a novel clustering approach based on that distribution. The application example presented here used the Gaussian mixture algorithm from Figueiredo and Jain [1], however any Gaussian mixture decomposition could be used. We showed that the proposed method is better or equally good when compared to the Gaussian mixture or to another method based also on dissimilarity increments.

The proposed method has a condition that the data should consist of 2-dimensional Gaussian clusters. However, we presented results for real data sets with dimension higher than two and the algorithm performs reasonably well compared to others. In future work we will extend the increment distribution to a generic dimension M.

Acknowledgments

We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250). This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), scholarship number SFRH/BD/39642/2007, and grant PTDC/EIA-CCO/103230/2008.

References

1. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
2. Fred, A., Jain, A.: Cluster validation using a probabilistic attributed graph. In: Proceedings of the 19th International Conference on Pattern Recognition, ICPR 2008 (2008)
3. Fred, A., Leitão, J.: A new cluster isolation criterion based on dissimilarity increments. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(8), 944–958 (2003)
4. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31(3), 264–323 (1999)
5. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, Applied Probability and Statistics, 2nd edn., vol. 1. John Wiley & Sons Ltd., Chichester (1994)
6. Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses, 3rd edn. Springer Texts in Statistics. Springer, Heidelberg (2005)
7. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 2nd edn. Elsevier Academic Press, Amsterdam (2003)

Unsupervised Joint Feature Discretization and Selection

Artur Ferreira[1,3] and Mário Figueiredo[2,3]

[1] Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal
[2] Instituto Superior Técnico, Lisboa, Portugal
[3] Instituto de Telecomunicações, Lisboa, Portugal
arturj@isel.pt, mario.figueiredo@lx.it.pt

Abstract. In many applications, we deal with high dimensional datasets with different types of data. For instance, in text classification and information retrieval problems, we have large collections of documents. Each text is usually represented by a bag-of-words or similar representation, with a large number of features (terms). Many of these features may be irrelevant (or even detrimental) for the learning tasks. This excessive number of features carries the problem of memory usage in order to represent and deal with these collections, clearly showing the need for adequate techniques for feature representation, reduction, and selection, to both improve the classification accuracy and the memory requirements. In this paper, we propose a combined unsupervised feature discretization and feature selection technique. The experimental results on standard datasets show the efficiency of the proposed techniques as well as improvement over previous similar techniques.

Keywords: Feature discretization, feature selection, feature reduction, Lloyd-Max algorithm, sparse data, text classification, support vector machines, naïve Bayes.

1 Introduction

Feature selection (FS) and *feature reduction* (FR) are central problems in machine learning and pattern recognition [1–3]. A typical dataset with numeric features uses floating point or integer representations and most FS/FR methods are applied directly on these representations. *Feature discretization* (FD) [4] has been proposed as a means of reducing the amount of memory required as well as the training time, leading to an improvement on the classification accuracy.

The need for FD techniques is most noticed with high dimensional datasets. For instance, in text classification problems, we have large collections of documents. For text, the *bag-of-words* (BoW) or similar representations are typically used. These representations have a large number of features, many of which may be irrelevant or misleading for the learning tasks. The FD, FS, and FR techniques used altogether reduce the memory requirements to represent these collections and improve the classification accuracy.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 200–207, 2011.

1.1 Our Contribution

In this work, we propose a new unsupervised technique for FD and FS. We perform scalar FD with the well-known Lloyd-Max algorithm [5], using a stopping criterion for bit allocation. The FS step applies a simple unsupervised criterion with indicators on the original (floating point or integer) feature and on the discretized feature. We compare our method against other techniques on standard datasets, using several classifiers.

The remaining text is organized as follows. Section 2 reviews some issues about document representation, feature discretization, reduction, and selection techniques. Section 3 presents both our methods for FD and FS. Section 4 reports experimental results and Section 5 presents concluding remarks and future work.

2 Background

This section briefly reviews some background concepts regarding document representation and feature discretization, reduction, and selection techniques.

2.1 Document Representation

Let $D = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_n, c_n)\}$ be a labeled dataset with training and test subsets, where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the i-th feature vector and c_i is its class label. For text classification problems, a BoW representation of a document consists in a high-dimensional vector containing some measure, like the *term-frequency* (TF) or the *term-frequency inverse-document-frequency* (TF-IDF) of a term [6, 7]. Each document is represented by a single vector, which is usually sparse, since many of its features are zero [6].

Let $\mathbf{X}$ be the $p \times n$ *term-document* (TD) matrix representing D; each column of $\mathbf{X}$ corresponds to a document, whereas each row corresponds to a term (e.g., a word); each column is the BoW representation of a document [6, 7]. Typically, TD matrices have high memory requirements, due to their dimensions and the use of floating point features.

2.2 Feature Discretization

FD has been used to reduce the amount of memory as well as to improve classification accuracy [4]. In the context of scalar FD [4], two techniques are commonly used: 1) *equal-interval binning* (EIB), *i.e.*, uniform quantization with a given number of bits for each feature; 2) *equal-frequency binning* (EFB), *i.e.*, non-uniform quantization yielding intervals such that for each feature the number of occurrences in each interval is the same. As a consequence, an EFB discretized feature has uniform distribution, thus maximum entropy. For this reason, this technique is also named *maximum entropy quantization*.

FD can be performed in supervised or unsupervised modes. The supervised mode uses the class labels when choosing discretization intervals and, in principle, may lead to better classifiers. However, in practice, it has been found that unsupervised FD methods tend to perform well in conjunction with several

classifiers; in particular, the EFB method in conjunction with *naïve Bayes* (NB) classifier produces excellent results [4]. For text classification problems, typically FS techniques have been applied to the (floating point) sparse BoW data [8–11]. In this paper, we consider solely unsupervised approaches for FD as well as the problem of text classification with discrete features.

2.3 Feature Selection Methods

This subsection reviews unsupervised and supervised FS methods that have been proved effective for several types of problems [8, 12, 13]. These methods adopt a filter [2] approach to the problem of FS, being applicable with any classifier. The unsupervised *term-variance* (TV) [12] method selects features (terms) X_i by their variance, given by

$$\mathrm{TV}_i = \mathrm{var}_i = \frac{1}{n}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 = \mathbf{E}[X_i^2] - \bar{X}_i{}^2, \tag{1}$$

where $\bar{X}_i$ is the average value of feature X_i, and n is the number of patterns.

The *minimum redundancy maximum relevancy* (mrMR) method [13] computes, in a supervised fashion, both the redundancy and the relevance of each feature. The redundancy is computed by the *mutual information* (MI) [14] between pairs of features, whereas relevance is measured by the MI between features and class label. The supervised *Fisher index* (FI) of each feature, on binary classification problems, is given by

$$\mathrm{FI}_i = \left|\mu_i^{(-1)} - \mu_i^{(+1)}\right| / \sqrt{\mathrm{var}_i^{(-1)} + \mathrm{var}_i^{(+1)}}, \tag{2}$$

where $\mu_i^{(\pm 1)}$ and $\mathrm{var}_i^{(\pm 1)}$, are the mean and variance of feature i, for the patterns of each class. The FI measures how well each feature separates the two (or more, since it can be generalized) classes. In order to perform FS based on any of these methods we select the m ($\leq p$) features with the largest rank.

3 Proposed Methods for Discretization and Selection

In this section, we present our proposals for unsupervised FD and unsupervised FS. For explanation purposes let us consider, like in subsection 2.1, a labeled dataset $D = \{(\mathbf{x}_1, c_1), ..., (\mathbf{x}_p, c_p)\}$, where $\mathbf{x}_i \in \mathbb{R}^n$ denotes the i-th feature vector and c_i is its class label.

3.1 Unsupervised Feature Discretization - FD Step

The proposed method for unsupervised FD named *FD step* performs scalar quantization of each feature with the well-known Lloyd-Max algorithm [5]. The algorithm runs for a given target distortion Δ in a *mean square error* (MSE) sense and a maximum number of bits q. The Lloyd-Max procedure is applied individually to each feature using the pair (Δ, q) as the stopping condition: the procedure stops when distortion Δ is achieved or when the maximum number of bits q per feature is reached.

3.2 Unsupervised Feature Selection - FS Step

The *unsupervised FS* (UFS) step uses a filter approach, on the discretized features, by computing the ranking criterion r_i for feature i given by

$$r_i = \text{var}(X_i)/b_i, \tag{3}$$

where $b_i \leq q$ is the number of bits allocated to feature i in the FD step, and $\text{var}(X_i)$ is the variance of the original (non-discretized) feature. We sort features in decreasing rank and keep only the first m ($\leq p$) features. The key idea of the FS step is: features with higher variance are more informative than features with lower variance; for a given feature variance, features quantized with a smaller number of bits are preferable because we can express all that variance (information) in a small number of bits, for the same target distortion Δ.

3.3 Choice of the Number of Features

In order to choose an adequate number of features, we propose to use a cumulative measure. Let $\{r_i,\ i = 1, ..., p\}$ be the values as given by (3) and $\{r_{(i)},\ i = 1, ..., p\}$ the same values after sorting in descending order. We propose choosing m as the lowest value that satisfies

$$\sum_{i=1}^{m} r_{(i)} / \sum_{i=1}^{p} r_{(i)} \geq L, \tag{4}$$

where L is some threshold (such as 0.95, for instance).

4 Experimental Evaluation

We have carried out the evaluation of our method with well-known datasets from the UCI Repository[1]. We consider datasets with BoW data as well as other types of data. We use linear *support vector machines* (SVM), *naïve Bayes* (NB), and *k-nearest neighbours* (KNN) classifiers provided by the PRTools toolbox [15].

4.1 Sparse and Non-sparse Datasets

Regarding sparse BoW datasets, we use: SpamBase, where the goal is to classify email messages as SPAM or non-SPAM; Example1[2]; and Dexter, where the task is learn to classify Reuters articles as being about "corporate acquisitions" or not. We also consider different non-text datasets: Ionosphere, WDBC, and Wine, as described in Table 1. In the case of Example1, each pattern is a 9947-dimensional BoW vector. The Dexter dataset has the same data as Example1 (with different train, test, and validation partitions) with 10053 additional distractor features (independent of the class), at random locations; it was created for the NIPS 2003

[1] `http://archive.ics.uci.edu/ml/datasets.html`
[2] `http://download.joachims.org/svm_light/examples`

Table 1. Dataset characteristics: two-class Example1, Dexter, SpamBase, Ionosphere, and WDBC; three-class Wine dataset. p, c, and n are the number of features, classes, and patterns respectively. c_a, c_b, c_c are the number of patterns per class. Mem. is the number of bytes needed to represent the floating point versions of the datasets (train, test, and validation partitions).

Dataset	p	c	**Subset**	n **Patt.**	$\mathbf{c_a, c_b, c_c}$	**Type of data**	**Mem.**
Example1	9947	2	Train	2000	(1000,1000,-)	Sparse BoW data	98.6 MB
			Test	600	(300,300,-)	subset of Reuters	
Dexter	20000	2	Train	300	(150,150,-)	Same data as Example1	198.3 MB
			Test	2000	(1000,1000,-)	with 10053 distractor	
			Valid.	300	(150,150,-)	features	
SpamBase	54	2	—	4601	(1813,2788,-)	Sparse BoW data - email	947.8 kB
Ionosphere	34	2	—	351	(225,126,-)	Radar data	45.5 kB
WDBC	30	2	—	569	(212,357)	Breast Cancer Diagnostic	65.1 kB
Wine	13	3	—	178	(59,71,48)	Wine from three cultivar	8.8 kB

FS challenge[3]. We evaluate on the validation set, since the labels for the test set are not publicly available; the results on the validation set correlate well with the results on the test set [3]. In the SpamBase dataset, we have used the first 54 features, which constitute a BoW. The Ionosphere, WDBC, and Wine datasets have non-sparse data, being quite different from BoW data.

4.2 Bit Allocation and Discretization Results

We compare the bit allocation results for each feature, using EFB and our FD step described in subsection 3.1. In both cases we consider $q = 16$. The EFB method stops allocating bits at iteration j, leading to $b_j, j \in \{1, \ldots, q\}$ bits for feature i, when the discretized feature entropy $H_j(X_i)$ exceeds 99% of the maximum possible entropy, that is, it stops at j bits, with $H_j(X_i) > 0.99 b_j$. Our FD step uses $\Delta = 0.01\text{var}(\mathbf{x}_i)$. Table 2 shows the FD results as well as the amount of memory needed to represent these datasets.

For sparse data and high dimensional datasets, such as Example1, Dexter, and SpamBase datasets, our FD step usually allocates much less bits than the EFB method. On non-sparse data, they tend to allocate comparable number of bits, depending on the statistic of each feature. The EFB method allocates less bits solely on the WDBC and Wine datasets. Notice that the FD step with $q = 16$ never uses more than 5 bits per feature.

4.3 Test Set Error Rate

We assess the results of SVM, NB, and KNN (K=3) classifiers on original and discrete features. The EFB and FD step discretization are carried out under the same conditions as in Table 2. Table 3 shows the test set error rates on Ionosphere, WDBC, and Wine datasets, without FS.

[3] `http://www.nipsfsc.ecs.soton.ac.uk`

Table 2. Minimum, maximum, average and total bits allocated using EFB and our FD step, up to q=16 bits. EFB stops when the discretized feature entropy exceeds 99% of its maximum value. FD step uses $\Delta = 0.01\text{var}(X_i)$. The row corresponding to the FD method that allocates less bits is underlined. Notice that Mem. has a much smaller value than the corresponding entry in Table 1.

Dataset		**EFB**					**FD step**				
Name	p	Min.	Max.	Avg.	Total	Mem.	Min.	Max.	Avg.	Total	Mem.
Example1	9947	2	8	7.99	18564	6033300	**2**	**5**	**2.19**	**5092**	**1654900**
Dexter	20000	2	16	15.98	82284	26742300	**2**	**4**	**2.12**	**10901**	**3542825**
SpamBase	54	2	16	15.74	834	479655	**2**	**4**	**3.23**	**171**	**98347**
Ionosphere	34	2	16	7.52	248	10881	**2**	**4**	**3.64**	**120**	**5265**
WDBC	30	**2**	**2**	**2.00**	**60**	**4268**	4	5	4.17	125	8891
Wine	13	**2**	**16**	**3.08**	**40**	**890**	4	4	4.00	52	1157

Table 3. Test set error rate (%) (average of 10 runs) for the Ionosphere, WDBC, and Wine datasets, using SVM, NB, and KNN (K=3) classifiers without FS. For each dataset and each classifier, the best result is underlined.

	Ionosphere			**WDBC**			**Wine**		
Representation	**SVM**	**NB**	**KNN**	**SVM**	**NB**	**KNN**	**SVM**	**NB**	**KNN**
Original (floating point)	**9.95**	**12.94**	**11.94**	4.80	**6.60**	8.40	1.51	1.24	12.62
EFB-discretized	16.92	30.35	17.41	4.20	6.80	6.00	1.86	1.95	11.37
FD step-discretized	13.43	18.91	16.42	**3.60**	6.80	**4.80**	**0.98**	**1.06**	**2.4**

On both WDBC and Wine datasets, our FD step leads to discretized features such that the test set error rate is lower or equal to that of EFB-discretized or floating point features. On Ionosphere dataset, both discretization techniques lead to worse results when compared to the original floating point features.

Fig. 1 shows the test set error rates (average of ten train/test partitions) for the SpamBase and WDBC datasets with FS methods (including our UFS step) on floating point, EFB-discretized, and FD step-discretized. The mrMR method attains better results with the FD-discretization. The FD step + UFS step attains good results for $m \geq 25$, better than the supervised mrMR method. Our FD step is adequate for discretization with the mrMR method. The KNN classifier attains adequate results on the WDBC dataset.

Table 4 shows the test set error rate of linear SVM and NB classifiers, on the SpamBase, WDBC, and Wine datasets, comparing our UFS method with the supervised FI method and the unsupervised TV method. For each FS algorithm, we apply our cumulative technique discussed in subsection 3.3 to compute an adequate number of reduced features by (4), with $L = 0.95$. We use the floating point, EFB-discretized, and FD step discretized data.

These results show that our FD method is adequate; it leads to an improvement on the classification accuracy, using the FS methods, as compared with the same methods applied on the original floating point data. The UFS step leads to adequate results. The proposed method for computing the number of features m with $L = 0.95$ also leads to good results.

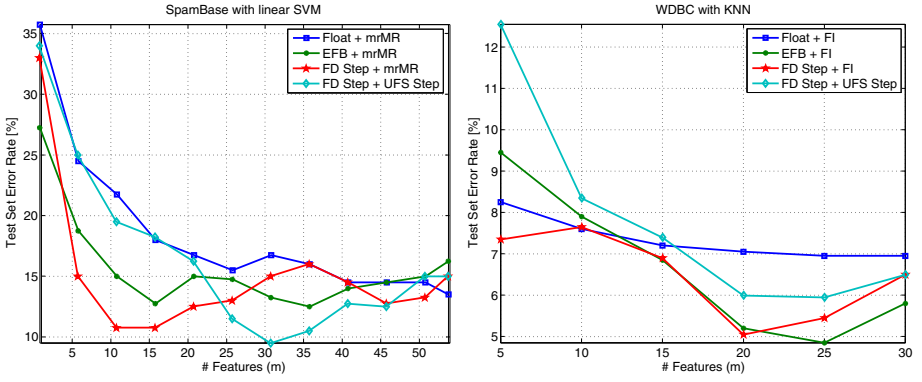


Fig. 1. Test set error rates (%) (average of ten runs) for SpamBase with linear SVM classifier and WDBC with KNN classifier. FS is performed by our FS step, mrMR (SpamBase), and FI (WDBC) on floating point, EFB-discretized, and FD Step data.

Table 4. Test Set Error rate (%) (average of ten runs) for the SpamBase, WDBC, and Wine datasets, for linear SVM and NB, with FS. The number of features m is given by (4), with $L = 0.95$. For each dataset and classifier, the best result is underlined.

Dataset name	**Linear SVM**						**Naïve Bayes (NB)**					
	Float		**EFB**		**FD step**		**Float**		**EFB**		**FD step**	
	FI	TV	FI	UFS	FI	UFS	FI	TV	FI	UFS	FI	UFS
SpamBase	**12.20**	**12.20**	18.30	19.10	12.30	14.70	15.60	16.00	**12.90**	13.30	14.30	16.00
WDBC	5.07	7.40	**3.80**	3.93	4.00	3.87	**6.27**	12.47	7.07	7.20	6.60	6.67
Wine	2.07	34.93	1.87	26.67	**1.20**	1.40	2.80	27.00	**2.13**	23.80	3.27	3.13

5 Conclusions

In this paper, we have proposed a new unsupervised technique for feature discretization based on the Lloyd-Max algorithm. The proposed algorithm allocates a variable number of bits for each feature attaining adequate representations for supervised and unsupervised feature selection methods, with several classifiers. We have also proposed an efficient unsupervised feature selection technique for discrete features, as well as a criterion to choose an adequate number of features.

The experimental results on standard datasets with sparse and non-sparse data indicate that our discretization method usually allocates a small number of bits per feature; this leads to efficient dataset representation and large memory savings. The joint use of feature selection method and the criterion to choose the number of features lead to classification results comparable or superior to other feature discretization techniques, using several classifiers with different types of (sparse and non-sparse) data. Future work will address joint feature quantization as well as the development of a supervised discretization procedure.

Acknowledgements

This work was partially supported by the Portuguese *Fundação para a Ciência e Tecnologia* (FCT), under grant SFRH/BD/45176/2008.

References

1. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, Heidelberg (2001)
2. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182 (2003)
3. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.): Feature Extraction, Foundations and Applications. Springer, Heidelberg (2006)
4. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Elsevier, Morgan Kauffmann (2005)
5. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. on Information Theory IT-28, 127–135 (1982)
6. Joachims, T.: Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers, Dordrecht (2001)
7. Manning, C., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
8. Forman, G.: An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
9. Yan, X.: A formal study of feature selection in text categorization. Journal of Communication and Computer 6(4) (April 2009)
10. Ferreira, A., Figueiredo, M.: Unsupervised feature selection for sparse data. In: 19th Europ. Symp. on Art. Neural Networks-ESANN 2011, Belgium (April 2011)
11. Ferreira, A., Figueiredo, M.: Feature transformation and reduction for text classification. In: 10th Int. Workshop PRIS 2010, Portugal (June 2010)
12. Liu, L., Kang, J., Yu, J., Wang, Z.: A comparative study on unsupervised feature selection methods for text clustering. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 597–601 (2005)
13. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
14. Cover, T., Thomas, J.: Elements of Information Theory. John Wiley, Chichester (1991)
15. Duin, R., Paclik, P., Juszczak, P., Pekalska, E., Ridder, D., Tax, D., Verzakov, S.: PRTools4.1, a Matlab Toolbox for Pattern Recognition. Technical report, Delft University of Technology (2007)

Probabilistic Ranking of Product Features from Customer Reviews

Lisette García-Moya, Henry Anaya-Sánchez,
Rafel Berlanga, and María José Aramburu

Universitat Jaume I, Spain
{lisette.garcia,henry.anaya,berlanga,aramburu}@uji.es

Abstract. In this paper, we propose a methodology for obtaining a probabilistic ranking of product features from a customer review collection. Our approach mainly relies on an entailment model between opinion and feature words, and suggest that in a probabilistic opinion model of words learned from an opinion corpus, feature words must be the most probable words generated from that model (even more than opinion words themselves). In this paper, we also devise a new model for ranking corpus-based opinion words. We have evaluated our approach on a set of customer reviews of five products obtaining encouraging results.

Keywords: product feature extraction, opinion mining, entailment model.

1 Introduction

The Web has become an excellent way of expressing opinions about products. Thus, the number of Web sites containing such opinions is huge and it is constantly growing. In recent years, opinion mining and sentiment analysis has been an important research area in Natural Language Processing [7]. Product featured extraction is a task of this area, and its goal is to discover which aspects of a specific product are the most liked or disliked by customers. A product has a set of components (or parts) and also a set of attributes (or properties). The word *features* is used to represent both components and attributes. For example, given the sentence, "The battery life of this camera is too short", the review is about the "battery life" feature and the opinion is negative.

This paper focuses on the feature extraction task. Specifically, given a set of customer reviews about a specific product, we address the problem of identifying all possible potential product features and ranking them according to their relevance. The basic idea of our ranking method is that if a potential feature is valid, it should be ranked high; otherwise it should be ranked low in the final result. We believe that ranking is also important for feature mining because ranking helps users to discover important features from the extracted hundreds of fine-grained potential features.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 208–215, 2011.

The remainder of the paper is organized as follows: Section 2 describes related work. In Section 3, we explain the proposed methodology. Section 4 presents and discusses the experimental results. Finally, in Section 5, we conclude with a summary and future research directions.

2 Related Work

Existing product feature extraction techniques can be broadly classified into two major approaches: supervised and unsupervised ones.

Supervised product feature extraction techniques require a set of preannotated review sentences as training examples. A supervised learning method is then applied to construct an extraction model, which is able to identify product features from new customer reviews. Different approaches such as Hidden Markov Models and Conditional Random Fields [12,13], Maximum Entropy [9], Class Association Rules and Naive Bayes Classifier [14] and other ML approaches have been employed for this task.

Although the supervised techniques can achieve reasonable effectiveness, preparing training examples is time consuming. In addition, the effectiveness of the supervised techniques greatly depends on the representativeness of the training examples. In contrast, unsupervised approaches automatically extract product features from customer reviews without involving training examples. According to our review of existing product feature extraction techniques, the unsupervised approaches seem to be more flexible than the supervised ones for environments in which various and frequently expanding products get discussed in customer reviews.

Hu and Liu's work [6,5] (PFE technique), uses association rule mining based on the Apriori algorithm [1] to extract frequent itemsets as explicit product features. However, this algorithm neglects the position of sentence words. In order to remove wrong frequent features, two types of pruning criteria were used: compactness and redundancy pruning. The technique is efficient and does not require the use of training examples or predefined sets of domain-independent extraction patterns. However, the design principle of PFE technique is not anchored in a semantic perspective. As a result, it is ineffective in excluding non-product features and opinion-irrelevant product features. Such limitations greatly limit its effectiveness. Details about these limitations are presented in [11]. To address these limitations, Wei et al. [11] proposed a semantic-based product feature extraction technique (SPE) that exploits a list of positive and negative adjectives defined in the General Inquirer [10] in order to recognize opinion words, and subsequently to extract product features expressed in customer reviews. Even when the SPE technique attains better results than previous approaches, both rely on mining frequent itemsets, with its commented limitations.

Qiu et al. [8] proposed a double propagation method, which exploits certain syntactic relations between opinion words and features, and propagates them iteratively. A dependency grammar was adopted to describe relations between opinion words and features themselves. The extraction rules are designed based

on these relations. This method works well for medium-size corpora. However, for large and small corpora, it can result in low precision and low recall [15]. To deal with these two problems, Zhang et al. [15] introduce *part-whole* and *no* patterns to increase the recall. Finally, feature ranking is applied to the extracted feature candidate in order to improve the precision of the top-ranked candidates.

3 Methodology

In this section we propose a new methodology to extract features from opinion reviews. It firstly extracts a set of potential features. Then, it defines a translation model based on words entailments. The purpose is to obtain a probabilistic ranking of these potential features. Finally, a new model for ranking corpus-based opinion words is proposed.

3.1 Extraction of Potential Features and Construction of a Basic Opinion Words List

Potential Features: In this work, we consider a set of *potential features* defined as the word sequences that satisfy the following rules:

1. Sequences of nouns and adjectives (e.g. "battery life", "lcd screen").
2. When gerund and participle occur between nouns, they are considered as part of the feature (e.g. "battery charging system").
3. Let PF_1 and PF_2 be potential features extracted by applying any of the previous rules. Let also $connector_1 = (of, from, at, in, on)$, and $connector_2 = (the, this, these, that, those)$. If the pattern $PF_1 connector_1 [connector_2] PF_2$ occurs, then the phrase formed by PF_2 concatenated with PF_1 is extracted as a potential feature. For example, "quality of photos" $\rightarrow$ "photo quality".

Opinion Words: We construct our own list of basic *opinion words*. An initial list was created by the intersection of adjectives from the list proposed by Eguchi and Lavrenko [3], the synsets of WordNet scored positive or negative in SentiWordNet [4] and the list of positive and negative words from the General Inquirer. Then, this initial list was extended with synonyms and antonyms from WordNet 3.0. Finally, the obtained list was manually checked, discarding those adjectives with context-dependent polarity. Additionally, some adverbs and verbs with context-independent polarity were added. The resulting list is formed by 1176 positive words and 1412 negative words.

3.2 Translation Model for Feature Ranking

In order to rank the set of potential features from customer reviews with vocabulary $V = \{w_1, \ldots, w_n\}$, we rely on the entailment relationship between words given by $\{p(w_i|w_j)\}_{w_i,w_j \in V}$, where $p(w_i|w_j)$ represents some posterior probability of w_i given w_j. In this work, we interpret $p(w_i|w_j)$ as the probability that $w_j \in V$ entails word $w_i \in V$.

In the context of customer reviews, opinion words usually express people sentiments about features, and therefore they can be seen as feature modifiers. Thus, in our work we consider that feature words can be successfully retrieved from the ranking given by the following conditional probability:

$$p(w_i|O) = \sum_{w \in V} p(w_i|w) \cdot p^*(w|O) \tag{1}$$

where $i \in \{1, \ldots, n\}$ and $\{p^*(w|O)\}_{w \in V}$ represents a basic language model of opinion words. The underlying idea is that in a probabilistic opinion model of words learned from an opinion corpus, feature words must be the most probable words generated from that model (even more than opinion words themselves), because of the entailment relationship between opinion and feature words. In this way, we regard that the probability of including a word w into the class of feature words F can be defined as:

$$p(F|w) \propto p(w|O). \tag{2}$$

Notice that if we estimate $\{p(w_i|w_j)\}_{w_i,w_j \in V}$ from customer reviews, the model $\{p(w_i|O)\}_{w_i \in V}$ can be seen as a corpus-based model of opinion words that is obtained by smoothing the basic model $\{p^*(w_i|O)\}_{w_i \in V}$ with the translation model $\{p(w_i|w_j)\}_{w_i,w_j \in V}$. Accordingly, we can also obtain a ranking of corpus-based opinion words by using Bayes formula on $\{p(w_i|O)\}_{w_i \in V}$. That is,

$$p(O|w_i) \propto \frac{p(w_i|O)}{p(w_i)}. \tag{3}$$

The same analysis can also be applied to features defined by multiword phrases (e.g. "battery life", "battery charging system" or "highest optical zoom picture"). Specifically, the probability of including a phrase $s = w_{i_1} \ldots w_{i_m}$ into the class of general features $\mathcal{F}$ can be defined as:

$$p(\mathcal{F}|s) \propto p(s|O) = p(w_{i_1} \ldots w_{i_m}|O). \tag{4}$$

3.3 Probability Density Estimation

The above probabilistic models for retrieving features and corpus-based opinion words depend on estimations for $\{p(w_i|w_j)\}_{w_i,w_j \in V}$, $\{p(w_i)\}_{w_i \in V}$ and the basic model of opinion words $\{p^*(w_i|O)\}_{w_i \in V}$.

For estimating $\{p(w_i|w_j)\}_{w_i,w_j \in V}$ we rely on a translation model like that presented in [2]. Thus, we firstly compute an initial word posterior probability conditioned on the vocabulary words defined as:

$$p_1(w_i|w_j) = \frac{p_1(w_i, w_j)}{p_1(w_j)} \tag{5}$$

where

$$p_1(w_i, w_j) \propto \sum_{v \in W} p(w_i|v) \cdot p(w_j|v) \cdot p(v), \tag{6}$$

$$p_1(w_j) = \sum_{w_i \in V} p_1(w_j, w_i), \tag{7}$$

and W is the set of all possible word windows of size k that can be formed in each sentence from the customer reviews. In the experiment carried out in this paper the best performance is achieved using $k = 5$. These probabilities are estimated using $p(w_i|v) = |v|_{w_i}/|v|$ and $p(v) = |W|^{-1}$, where $|v|_{w_i}$ is the number of times w_i occurs in window v, $|v|$ is the length of v, and $|W|$ is the cardinal of W.

For all $w_i, w_j \in V$, the probability $p_1(w_i|w_j)$ can be seen as the probability of translating w_j into w_i in one translation step. Then, we define $p(w_i|w_j)$ as a smoothed version of $p_1(w_i|w_j)$ obtained by generating random Markov chains between words. Specifically, we define $p(w_i|w_j)$ as:

$$p(w_i|w_j) = \Big((1-\alpha) \cdot (I - \alpha \cdot P_1)^{-1}\Big)_{i,j} \tag{8}$$

where I is the $n \times n$ identity matrix, P_1 is a $n \times n$ matrix whose element P_{ij} is defined as $p_1(w_i|w_j)$, and α is a probability value that allows the generation of arbitrary Markov chains between words. In the experiment carried out in this paper we use $\alpha = 0.99$, which allows the generation of large chains, and thus a great smoothing.

Thus, the overall model of words $\{p(w_i)\}_{w_i \in V}$ can be estimated from the linear equation system given by the n variables $\{p(w_i)\}_{w_i \in V}$, and $n + 1$ equations:

$$p(w_i) = \sum_{w_j \in V} p(w_i|w_j) \cdot p(w_j) \qquad (i \in \{1, \ldots, n\}) \tag{9}$$

$$\sum_{w_i \in V} p(w_i) = 1. \tag{10}$$

The basic model of opinion words considered in this work is estimated from the list of basic opinion words described in section 3.1. We consider $p^*(w_i|O)$ defined as the following non-smoothed model:

$$p^*(w_i|O) = \begin{cases} \frac{1}{|Q|} & \text{if } w_i \in Q \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where Q is the set of basic opinion words and $|Q|$ is the size of Q.

4 Experiments

In order to validate our methodology, we have conducted several experiments on the customer reviews from five products: Apex AD2600 Progressive-scan DVD player, Canon G3, Creative Labs Nomad Jukebox Zen Xtra 40GB, Nikon coolpix 4300 and Nokia 6610. The reviews were collected from Amazon.com and

Table 1. Summary of customer review data set

	Apex	Canon	Creative	Nikon	Nokia
Number of review sentences	738	600	1705	350	548

CNET.com.[1] Table 1 shows the number of review sentences for each product in the data set. Each review was annotated using the Stanford POS Tagger.[2]

Firstly, we propose to compare our ranking method with a version of the method proposed by Zhang et al. [15]. Zhang et al. considered that the importance of a feature is determined by its relevance and its frequency. In order to obtain the relevance score of a feature, they apply the HITS algorithm where *potential features* act as authorities and *feature indicators* act as hubs forming a directed bipartite graph. The basic idea is that if a potential feature has a high authority score, it must be a highly-relevant feature. If a feature indicator has a high hub score, it must be a good feature indicator. The final score function considering the feature frequency is:

$$score(s) = A(s) \cdot \log(freq(s)) \tag{12}$$

where $freq(s)$ is the frequency of the potential feature s, and $A(s)$ is the authority score of the potential feature s. In our case, an opinion word o co-occurring with any word $w_i \in s$ in the same window v is considered as a feature indicator of s. We are going to consider this method as our *baseline*.

Table 2. Precision at top N

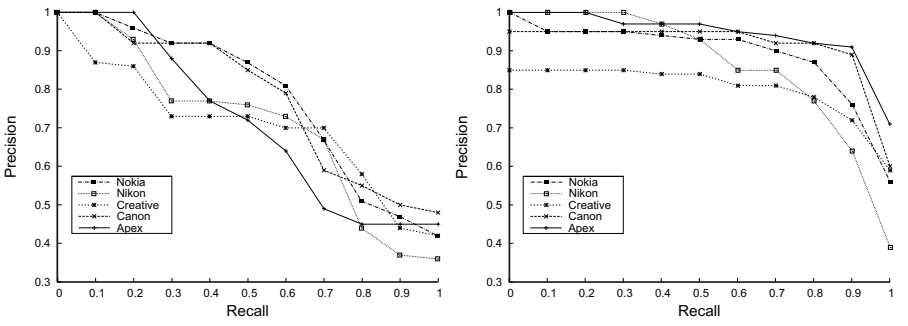
N	*Baseline*					*Our approach*				
	Apex	Canon	Creative	Nikon	Nokia	Apex	Canon	Creative	Nikon	Nokia
50	72.0	92.0	78.0	72.0	86.0	96.0	92.0	80.0	90.0	94.0
100	62.0	78.0	72.0	55.0	67.0	95.0	94.0	82.0	78.0	91.0
150	48.0	61.3	69.3	43.3	52.7	92.0	91.3	82.0	72.0	86.0
200	42.5	54.0	69.5	38.0	49.0	91.0	91.5	81.0	65.0	78.5
250	42.0	54.0	64.4	37.2	46.0	85.6	90.0	80.0	58.4	73.6

The performance of the methods is firstly evaluated in terms of the measure **precision@*N***, defined as the percentage of valid features that are among the top N features in a ranked list. In Table 2, we show the obtained results for each $N \in \{50, 100, 150, 200, 250\}$. As it can be seen, our method consistently outperforms the baseline for each value of N. Also, it can be appreciated that different from the baseline, the precision of our rankings do not decrease dramatically when N is greater than 100.

Secondly, we consider the 11-point interpolated average precision to evaluate the retrieval performance regarding the recall factor. Figure 1 compares the obtained curves. It can be seen that even considering the 11-point of recall scores,

[1] This customer review data set is available at `http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip`

[2] `http://nlp.stanford.edu/software/tagger.shtml`



(a) Baseline (b) Our approach

Fig. 1. 11-point Interpolated Recall-Precision curve for each product. The x-axis represents different recall levels while y-axis represents the interpolated precision at these levels.

our approach outperforms the baseline, while also maintains a good precision through out all recall values.

Finally, as it was explained in Section 3.2, it is possible to obtain a ranking of corpus-based opinion words (see Equation 3). Table 3 shows the first 15 opinion words obtained for some products together with their relevance value (i.e., $p(w_i|O)/p(w_i)$). In this table, we bold-faced those words that are not included in our basic opinion list. The obtained ranking corroborates the usefulness of the proposal for also retrieving corpus-based opinion words.

Table 3. Fragment of the ranking of corpus-based opinion words obtained for some products

Relevance	Creative	Relevance	Nokia	Relevance	Nikon
1.061	helpful	1.219	haggard	1.076	worse
1.057	**clock**	1.212	mad	1.070	**claim**
1.050	weighty	1.210	**gott**	1.069	kind
1.044	**biggie**	1.205	**junky**	1.063	**internal**
1.040	strange	1.202	**major**	1.059	**damage**
1.038	unlucky	1.197	bad	1.055	**refuse**
1.038	flashy	1.197	**duper**	1.052	**cover**
1.037	superfluous	1.197	**rad**	1.026	correct
1.037	evil	1.196	happy	1.023	**warranty**
1.036	smoothly	1.195	minus	1.022	**touchup**
1.036	user-friendly	1.193	negative	1.022	**alter**
1.036	**date**	1.190	brisk	1.022	redeye
1.036	**ounce**	1.189	significant	1.010	**cost**
1.036	ridiculous	1.188	**require**	1.009	outstanding
1.036	shoddy	1.188	**penny**	1.008	comfortable

5 Conclusions and Future Work

In this paper, a new methodology for obtaining a probabilistic ranking of product features from customer reviews has been proposed. The novelty of our approach relies on modeling feature words from a stochastic entailment model between opinion and feature words. In addition, a model for obtaining a ranking of corpus-based opinion words is also proposed. One strong point of our method is that

it does not depend on any natural language processing except for POS tagging. The experimental results obtained over a set of customer reviews of five products validate the usefulness of our proposal. Our future work is oriented to design a method to cut the ranked list in order to remove those spurious features.

Acknowledgments. This work has been partially funded by the Spanish Research Program project TIN2008-01825/TIN and the Fundacio Caixa Castelló project P1.1B2008-43. Lisette García-Moya has been supported by the PhD Fellowship Program of the Universitat Jaume I (PREDOC/2009/12).

References

1. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In: Proceedings of the International Conference on Very Large Data Base, vol. 487, p. 499 (1994)
2. Berger, A., Lafferty, J.: Information retrieval as statistical translation. In: Proceedings of the 22nd ACM SIGIR, Berkeley, CA, pp. 222–229 (1999)
3. Eguchi, K., Lavrenko, V.: Sentiment retrieval using generative models. In: Proceedings of the EMNLP 2006, pp. 345–354. Association for Computational Linguistics (2006)
4. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of the 5th LREC, pp. 417–422 (2006)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD, pp. 168–177 (2004)
6. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of AAAI 2004, pp. 755–760 (2004)
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Now Publishers Inc. (2008)
8. Qiu, G., Liu, B., Bu, J., Chen, C.: Expanding domain sentiment lexicon through double propagation. In: Proceedings of the IJCAI 2009, pp. 1199–1204 (2009)
9. Somprasertsri, G., Lalitrojwong, P.: Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features. In: Proceedings of the IEEE International Conference on Information Reuse and Integration, pp. 250–255 (2008); IEEE Systems, Man, and Cybernetics Society
10. Stone, P.J., Dunphy, D.C., Smith, M.S., Ogilvie, D.M.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge (1966)
11. Wei, C.P., Chen, Y.M., Yang, C.S., Yang, C.C.: Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. Information Systems and E-Business Management, 149–167 (2009)
12. Wong, T.L., Lam, W.: Hot Item Mining and Summarization from Multiple Auction Web Sites. In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 797–800. IEEE Computer Society, Washington, DC (2005)
13. Wong, T.L., Lam, W.: Learning to extract and summarize hot item features from multiple auction web sites. Knowl. Inf. Syst. 14(2), 143–160 (2008)
14. Yang, C.C., Wong, Y.C., Wei, C.P.: Classifying web review opinions for consumer product analysis. In: Proceedings of the 11th International Conference on Electronic Commerce, pp. 57–63. ACM, New York (2009)
15. Zhang, L., Liu, B., Lim, S.H., O'Brien-Strain, E.: Extracting and ranking product features in opinion documents. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1462–1470. Coling 2010 Organizing Committee, Beijing (2010)

Vocabulary Selection for Graph of Words Embedding

Jaume Gibert[1], Ernest Valveny[1], and Horst Bunke[2]

[1] Computer Vision Center, Universitat Autònoma de Barcelona
Edifici O Campus UAB, 08193 Bellaterra, Spain
{jgibert,ernest}@cvc.uab.es
[2] Institute for Computer Science and Applied Mathematics, University of Bern,
Neubrückstrasse 10, 3012 Bern, Switzerland
bunke@iam.unibe.ch

Abstract. The Graph of Words Embedding consists in mapping every graph in a given dataset to a feature vector by counting unary and binary relations between node attributes of the graph. It has been shown to perform well for graphs with discrete label alphabets. In this paper we extend the methodology to graphs with n-dimensional continuous attributes by selecting node representatives. We propose three different discretization procedures for the attribute space and experimentally evaluate the dependence on both the selector and the number of node representatives. In the context of graph classification, the experimental results reveal that on two out of three public databases the proposed extension achieves superior performance over a standard reference system.

1 Introduction

Classically, graph matching has been addressed by graph and subgraph isomorphism, which try to define one-to-one correspondences between nodes of graphs, by means of finding common super- and substructures among the involved graphs, or even by defining distances between instances of graphs. For the latter, graph edit distance is a commonly used approach since it allows one to define a measure of similarity between any kind of graphs. This is accomplished by computing the amount of distortion (in terms of edit operations) needed to transform one graph into the other. A good reference for graph matching can be found in [1], where the authors properly define a taxonomy of how to compare graph instances.

Recently, among the structural pattern recognition researchers, there is an increasing interest in bridging the gap between the strong representation that graphs provide and the large repository of algorithms originally developed for the processing and analysis of feature vectors. To this end, graph embedding into real vectors spaces is of great attraction, since once a graph is directly associated to a feature vector, any statistical pattern analysis technique becomes available. An important work concerning graph embeddings is the one described in [2]. It

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 216–223, 2011.

approaches the problems of graph clustering and graph visualization by extracting different features from an eigen-decomposition of the adjacency matrices of the graphs. Another important graph embedding procedure is explained in [3]. The nodes of the graph are embedded into a metric space and then the edges are interpreted as geodesics between points on a Riemannian manifold. The problem of matching nodes to nodes is viewed as the alignment of the point sets. Other approaches are based on random walks, and particularly, on quantum walks, in order to embed nodes into a vector space [4]. The embedding is based on the expected time for the walk to travel from one node to another, the so-called commute time. Due to its generality, the work in [5] is worth mentioning here. It classifies and clusters a set of graphs by associating to every input graph a vector whose components are edit distances to a set of graph prototypes. Finally, in [6], to solve the problem of molecules classification, the authors associate a feature vector to every molecule by counting unary and binary statistics in the molecule; these statistics indicate how many times every atomic element appears in the molecule, and how often there is a bond between two specific atoms. The good point of this embedding methodology is that it is both easy to compute and does not require costly operations.

In this work, instead of working with just molecules, we aim at generalizing the main embedding technique described in [6] to other graphs. The principal problem of this generalization is the fact that molecules are attributed graphs whose attributes are discrete. When dealing with more general graphs - for instance, when nodes are attributed with n-dimensional vectors - the continuity of the labels requires a pre-processing step, where some representatives of those labels have to be selected. This article is thus focused on how to choose the set of node attribute representatives (also known as *vocabulary*) when these labels are n-dimensional continuous attributes, and how the choice of the vocabulary affects the classification of graphs under this embedding.

First, we give a brief introduction to the embedding procedure we want to generalize. This is done in the next section. Then, in Section 3, we describe the techniques that we have used to select node representatives. In Section 4 and 5, experiments and their results are presented. Finally we draw some conclusions in Section 6.

2 Graph of Words Embedding

Although the embedding of graphs into vectors spaces provides a way to be able to apply statistical pattern analysis techniques to the domain of graphs, the existing methods still suffer from the main drawback that the classical techniques also did, this is, their computational cost. The Graph of Words Embedding tries to avoid these problems by just visiting nodes and edges instead of, for instance, travelling along paths in the graphs or computing, for every adjacency matrix, the eigen-decomposition. In this section we first briefly explain the motivation of the technique and then formally define the procedure.

2.1 Motivation

In image classification, a well-known image representation technique is the so-called bag of visual features, or just bag of words. It first selects a set of feature representatives, called words, from the whole set of training images and then characterizes each image by a histogram of appearing words extracted from the set of salient points in the image [7].

The graph of words embedding proceeds in an analogous way. The salient points in the images correspond to the nodes of the graphs and the visual descriptors are the node attributes. Then, one also selects representatives of the node attributes (words) and count how many times each representative appears in the graph. This leads to a histogram representation for every graph. Then, to take profit of the edges in the original graphs, one also counts the frequency of the relation between every pair of words. The resulting information is combined with the representative's histogram in a final vector.

2.2 Embedding Procedure

A graph is defined by the 4-tuple $g = (V, E, \mu, \nu)$, where V is the set of nodes, $E \subseteq V \times V$ is the set of edges, μ is the nodes labelling function, assigning labels to every node and, ν is the edges labelling function, assigning labels to every edge in the graph. In this work we just consider graphs whose nodes attributes are real vectors, this is $\mu : V \rightarrow \mathbb{R}^d$ and whose edges remain unattributed, this is, $\nu(e) = \varepsilon$ for all $e \in E$ (where ε is the null label).

Let $\mathcal{P}$ be the set of all nodes attributes in a given dataset of graphs $\mathcal{G} = \{g_1, \ldots, g_M\}$. From all points in $\mathcal{P}$ we derive n representatives, which we shall call *words*, in analogy to the bag of words procedure. Let this set of words be $\mathcal{V} = \{w_1, \ldots, w_n\}$ and be called *vocabulary*. Then, before assigning a vector to each graph, we first construct an intermediate graph that will allow us an easier embedding. This intermediate graph, called *graph of words* $g' = (V', E', \mu', \nu')$ of $g = (V, E, \mu, \nu) \in \mathcal{G}$ with respect to $\mathcal{V}$, is defined as:

- $V' = \mathcal{V}$
- E' is defined by: $(w, w') \in E' \Leftrightarrow$ there exists $(u, v) \in E$ such that $\lambda(u) = w$ and $\lambda(v) = w'$
- $\mu'(w) = |\{v \in V \,|\, w = \lambda(v)\}|$
- $\nu'(w, w') = |\{(u, v) \in E \mid \lambda(u) = w, \lambda(v) = w'\}|$

where λ is the node-to-word assignment function $\lambda(v) = \arg\min_{w_i \in \mathcal{V}} d(v, w_i)$, this is, the function that assigns a node to its closest word.

Once the graph of words is constructed, we easily convert the original graph into a vector by combining the node and edge information of the graph of words, this is, by keeping both the information of the appearing words and the relation between these words. We consider the histogram

$$\phi_a^{\mathcal{V}}(g) = (\mu'(w_1), \ldots, \mu'(w_n)). \tag{1}$$

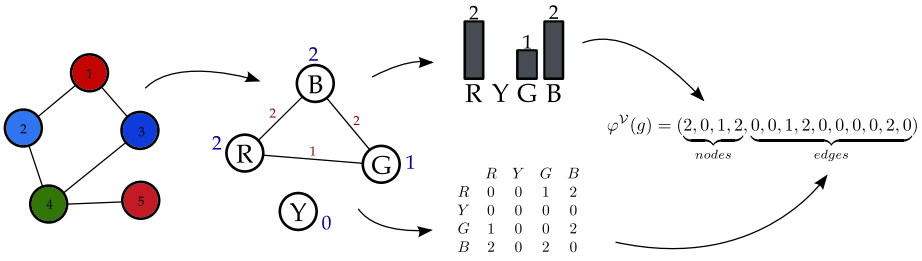


Fig. 1. Example of the graph of words embedding. The graph on the left is assigned to the vector on the right by considering the vocabulary $\mathcal{V} = \{R, Y, G, B\}$. Nodes 1 and 5 are assigned to the R word, 2 and 3 to the B word and 4 to the G word. Note that none is assigned to the Y word. The histogram of words is considered as well as the adjacency matrix. The resulting vector is the concatenation of both types of information.

and a flattened version of the adjacency matrix of the graph of words $A = (a_{ij})$, with $a_{ij} = \nu'(w_i, w_j)$:

$$\phi_b^{\mathcal{V}}(g) = (a_{11}, \ldots, a_{ij}, \ldots, a_{nn}), \;\; \forall\, i \leq j \tag{2}$$

The final graph of words embedding is the concatenation of both pieces of information, this is,

$$\varphi^{\mathcal{V}}(g) = (\phi_a^{\mathcal{V}}(g), \phi_b^{\mathcal{V}}(g)). \tag{3}$$

In Figure 1, there is an example of the graph of words procedure for a simple vocabulary of size equal to 4.

It is clear that the resulting vectors of the embedding heavily depend on the set of words $\mathcal{V}$; not only which words are there -this means, how the words have been selected- but also how many. To study these dependencies is one of the main objectives of this paper. The next section introduces different methods to select node attribute representatives.

3 Vocabulary Construction

The fact that we just consider graphs with n-dimensional points as node labels leaves us with some specific representative selection techniques. Three of them have been chosen and are described in the following subsections.

Grid selection. The set of all node attributes is $\mathcal{P} = \{p_1, \ldots, p_N\}$, where for all $i \in \{1, \ldots, N\}$, $p_i \in \mathbb{R}^d$. As a first and simple method to select words from $\mathcal{P}$, we have divided the parcel of $\mathbb{R}^d$ where all points in $\mathcal{P}$ lay on by using a regular grid. Then, each cell on the grid is represented by its centre point. The set of all cell centres will constitute our *grid vocabulary*.

Spanning prototypes. The regular grid approach might suffer from outliers or from a non homogeneous distribution of points over $\mathbb{R}^d$. To try to avoid that,

we have considered a method that tries to produce representatives by uniformly picking points all over the range where they actually live. Inspired in the spanning prototypes selector for the graph embedding in [5], we iteratively construct our vocabulary by first selecting the median vector of $\mathcal{P}$ and then adding at each iteration the point which is furthest away from the already selected ones.

kMeans algorithm. The last vocabulary selector we have considered is the very-well known kMeans algorithm [8]. It iteratively builds a set of k cluster centres by first initializing the k centres with some points, assigning each point in the set to the closest centre, and then recomputing the centres as the mean of the set of points assigned to the same cluster. The process finishes when there are no centre changes between one iteration and the next one.

In our experiments, in order to eliminate the randomness of the initialization, we have initialized the algorithm by using the spanning prototypes described before.

4 Experiments

Using the graph of words embedding we want to solve the problem of graph classification. We use a kNN classifier in conjunction with a χ^2 distance [9]. In this section we describe the databases that have been used and an independent reference system.

4.1 Databases

In order to ease visualization, we have only carried out experiments on databases whose attributes are (x, y) points in $\mathbb{R}^2$. We have chosen three different datasets from the IAM Graph Database Repository [10], describing both synthetic and real data.

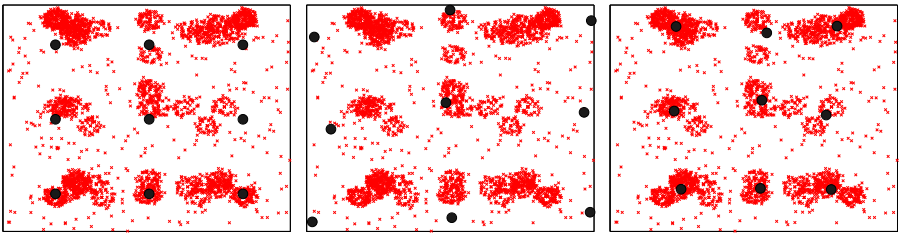
The *Letter Database* represents distorted letter drawings. From a manually constructed prototype of every of the 15 Roman alphabet letters that consist only on straight lines, different degrees of distortions have been applied. Each ending point of these lines is attributed with its (x, y) coordinates. We have been working on all levels of distortion (low, medium, high) but will only report those results for the low level of distortion.

The second graph dataset is the *GREC Database*, which represents architectural and electronic drawings under different levels of noise. In this database, intersections and corners constitute the set of nodes. These nodes are attributed with their position on the 2-dimensional plane.

Finally, the *Fingerprint Database* consists of graphs that are obtained from a set of fingerprint images by means of particular image processing operations. Ending point and bifurcations of the skeleton of the processed images constitute the (x, y) attributed nodes of the graphs.

4.2 Reference System: Graph Edit Distance

We have considered a reference method that is independent of the whole graph embedding process, and particularly, to the selection of any representative. In



(a) Regular grid (b) Spanning prototypes (c) kMeans centres

Fig. 2. Example of the vocabulary selectors for the Letter database. The small crosses are node attribute points and the bold dots are the corresponding vocabulary entries. In this experiment, the vocabulary size was chosen equal to 9.

[10], the results of a k-nearest neighbour classifier based on the graph edit distance are reported, and we use these results to compare the ones obtained with the described methodology.

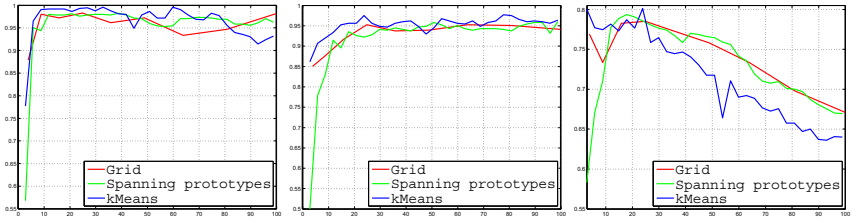
5 Results

Ideally, the better the words in the vocabulary adapt themselves to the implicit clusters of the graph nodes, the better the embedding of graphs by means of the graph of words technique will perform to solve the problem of graph classification. In Figure 2, we show an example of the vocabularies obtained from the three selectors of Section 3. Here the vocabulary size was defined equal to 9. For the Letter database, it can be seen that with the kMeans algorithm, the words better adapt to the clusters that are implicitly distributed over the range of node attributes than under the regular grid and the spanning prototype procedures.

As we already said, not only the selection of representatives for our vocabularies is an important issue to investigate, but also how many of these words are selected. In Figure 3, we show, for every dataset and every vocabulary selector, the accuracy rates of the classification on the validation set as a function of the number of selected words.

From this set of figures, we can see that the kMeans algorithm is in the Letter's and the GREC's case working better than the other two selectors, while for the Fingerprint database it is the other way around. This seems to be due to the distribution of the node attributes over the plane for the Fingerprint graphs. This distribution is almost a regular grid and this makes the kMeans algorithm to learn clusters that might not correspond to the actual ones.

An interesting fact to comment on here is the rapid decrease of the performance in the Fingerprint database when the number of words in the vocabulary is increasing. Obviously, this situation is not occurring for the Letter and GREC cases. A small vocabulary in the graph of words configuration is describing global relations among the nodes of the original graphs, while a larger one would be describing local characteristics. In case of the Fingerprint graphs, the local



(a) Letter (b) GREC (c) Fingerprint

Fig. 3. Accuracy rates on the validation set as a function of the number of selected words in the vocabulary. Every different color in each figure represents a different vocabulary selector.

Table 1. Results on the test set for the reference system (kNN based on graph edit distance) and the grid vocabulary, the spanning selector and kMeans algorithm. The accuracy rates (AR) are shown in %. The number of words in the vocabulary of the best configuration on the validation set is shown in the vocabulary size (VS) column. Bold face numbers indicate a higher performance over the other systems.

	Letter		GREC		Fingerprint	
	AR	VS	AR	VS	AR	VS
Grid selector	98	9	95.2	64	**78.5**	25
Spanning prototypes	97.8	27	96	99	77.6	3
kMeans	98.8	36	**97.5**	81	77.7	18
kNN classifier based on GED	**99.6**	-	95.5	-	76.6	-

connectivity between nodes is less important since the category of a fingerprint is determined by its general shape and, therefore, with less words the accuracy of the classifier is higher.

Finally, in Table 1, we show the results on the test set using the best configuration on the validation set. It does not seem to be a general rule by which we could say that one selector is better than another. However, the kMeans algorithm gets better results than the two other ones in two out of the three databases. It is also worth noticing that, for the GREC and Fingerprint datasets, the graph of words embedding is superior over the classification in the graph domain by means of the graph edit distance, while being computationally cheaper.

6 Conclusions

In this article we have extended the graph of words embedding approach to the case of graphs with n-dimensional vectors as nodes attributes and unlabelled edges. We have proposed three strategies for node representative selection and have experimentally evaluated their performance in the context of graph classification. Results have shown that the selector should, in general, adapt properly to the inherent clusters of the node attributes and it turns out that in two out

of three public databases the proposed method performs better than a standard reference classifier. In addition, the described methodology is computationally cheaper than such a reference system.

There are still several open problems to tackle under this framework. First of all, other space discretization techniques would certainly reveal more properties about the dependence of the graph of words technique to the vocabulary selection. Along this line, the methodology could also be extended to the use of fuzzy clustering techniques, where not just a discrete assignment of every node in the graph is described but a more flexible assignment is used, in terms of continuous degrees of membership.

The experiments that have been carried out in this article are restricted to a specific kind of graphs, whose node attributes are 2-dimensional vectors and edges remain unlabelled. The authors are currently working on the use of more general graphs with n-dimensional vectors as node attributes and labelled edges.

References

1. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. International Journal of Pattern Recognition and Artificial Intelligence 18(3), 265–298 (2004)
2. Luo, B., Wilson, R.C., Hancock, E.R.: Spectral embedding of graphs. Pattern Recognition 36(10), 2213–2230 (2003)
3. Robles-Kelly, A., Hancock, E.R.: A Riemannian approach to graph embedding. Pattern Recognition 40(3), 1042–1056 (2007)
4. Emms, D., Wilson, R.C., Hancock, E.R.: Graph Embedding using a Quasi-Quantum Analogue of the Hitting Times of Continuous Time Quantum Walks. Quantum Information and Computation 3-4(9), 231–254 (2009)
5. Riesen, K., Bunke, H.: Graph Classification and Clustering Based on Vector Space Embedding. World Scientific, Singapore (2010)
6. Gibert, J., Valveny, E., Bunke, H.: Graph of Words Embedding for Molecular Structure-Activity Relationship Analysis. In: Bloch, I., Cesar Jr., R.M. (eds.) CIARP 2010. LNCS, vol. 6419, pp. 30–37. Springer, Heidelberg (2010)
7. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision, pp. 1–22 (2004)
8. Jain, A., Murty, M., Flynn, P.: Data Clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
9. Cha, S., Srihari, S.N.: On measuring the distance between histograms. Pattern Recognition 35(6), 1355–1370 (2002)
10. Riesen, K., Bunke, H.: IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 287–297. Springer, Heidelberg (2008)

Feature Selection in Regression Tasks Using Conditional Mutual Information*

Pedro Latorre Carmona[1], José M. Sotoca[1], Filiberto Pla[1],
Frederick K.H. Phoa[2], and José Bioucas Dias[3]

[1] Dept. Lenguajes y Sistemas Informáticos, Jaume I University, Spain
[2] Institute of Statistical Science. Academia Sinica. R.O.C.
[3] Instituto de Telecomunicações and Instituto Superior Técnico, Technical University of Lisbon
{latorre,sotoca,pla}@uji.es,
fredphoa@stat.sinica.edu.tw, bioucas@lx.it.pt

Abstract. This paper presents a supervised feature selection method applied to regression problems. The selection method uses a *Dissimilarity matrix* originally developed for classification problems, whose applicability is extended here to regression and built using the conditional mutual information between features with respect to a continuous relevant variable that represents the *regression function*. Applying an agglomerative hierarchical clustering technique, the algorithm selects a subset of the original set of features. The proposed technique is compared with other three methods. Experiments on four data-sets of different nature are presented to show the *importance* of the features selected from the point of view of the regression estimation error (using Support Vector Regression) considering the Root Mean Squared Error ($RMSE$).

Keywords: Feature Selection, Regression, Information measures, Conditional Density Estimation.

1 Introduction

Feature selection aims at reducing the dimensionality of data. It consists of selecting the most relevant features (attributes) among the set of original ones [12]. This step is crucial for the design of regression and classification systems. In this framework, the term relevant is related to the impact of the variables on the prediction error of the variable to be regressed (*target variable*).The relevant criterion can be based on the performance of a specific predictor (wrapper method), or on some general relevance measure of the features for the prediction (filter method). Wrapper methods may have two drawbacks [19]: (a) they can be computationally very intensive; (b) their results may vary according to initial conditions or other chosen parameters. On the other hand, filter methods allow sorting features independently of the regressor. Eventually, embedded methods try to include the feature selection as a part of the regression training process. In order to tackle the combinatorial search problem to find an optimal subset of features, the

* This work was supported by the Spanish Ministry of Science and Innovation under the projects Consolider Ingenio $2010CSD2007-00018$, and EODIX $AYA2008-05965-C04-04/ESP$ and by *Fundació Caixa-Castelló* through the project $P11B2007-48$.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 224–231, 2011.

most popular variable selection methods seek to avoid having to perform an exhaustive search applying forward, backward or floating sequential schemes [9][16].

This paper presents a (filter type) feature clustering-based method aiming at finding a subset of features that minimizes the regression error. To do this, the conditional mutual information will be estimated to define a criterion of distance between features. This distance has already been used in [18] for feature selection in classification tasks. Thus, the main contribution of this paper is to establish a methodology to properly solve the estimation of this distance for regression problems where the relevant variable is continuous, through the assessment of the conditional mutual information.

The organization of the rest of this paper is as follows. Section 2 gives an overview of the theoretical foundations of this distance and solves its estimation in regression problems. Section 3 describes the experiments carried out, the databases used and the other feature selection methods in regression used in the comparison. Section 4 presents and discusses the regression results obtained. Finally, some concluding remarks are depicted in Section 5.

2 Feature Selection for Regression

2.1 The Minimal-Relevant-Redundancy Criterion Function

Sotoca *et al* show in [18] that if $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_m)$ is a subset of m random variables out of the original set of n random variables $\mathbf{X} = (X_1, \ldots, X_n)$, that is, $\widetilde{\mathbf{X}} \subset \mathbf{X}$, then, the decrease in mutual information about a relevant variable Y can be expressed as $I(\mathbf{X};Y) - I(\widetilde{\mathbf{X}};Y) = I(\mathbf{X};Y/\widetilde{\mathbf{X}})$. They also show that the decrease of mutual information of the original and the reduced sets with respect to the relevant variable Y is upper bounded by $I(\mathbf{X};Y) - I(\widetilde{\mathbf{X}};Y) \leq \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i;Y/\widetilde{X}_j)$, where $I(X_i;Y/\widetilde{X}_j)$ is the conditional mutual information between the feature X_i and the cluster representative $\widetilde{X}_j$ and expresses how much information variable X_i can predict about the relevant variable Y that $\widetilde{X}_j$ cannot. This bound can be interpreted as a *Minimal Relevant Redundancy - mRR* criterion, meaning that the selected features will tend to be as independent as possible with respect to the information content of the relevant variable Y they are attempting to predict.

One way to find a solution to the minimization problem is to approximate this bound by a clustering process, where $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_m)$ are the representative features of each cluster. To solve this problem, given two variables X_i and X_j, the following function satisfies the properties of a metric distance:

$$D(X_i, X_j) = I(X_i;Y/X_j) + I(X_j;Y/X_i) = H(Y/X_i) + H(Y/X_j) - H(Y/X_i, X_j) \tag{1}$$

The same metric will be considered here, using an agglomerative hierarchical approach based on a Ward's linkage method [20]. The number of groups is reduced at each iteration until m clusters are reached. For each resulting cluster, C_j, its representative feature $\widetilde{X}_j$ is chosen as the feature X_j with the highest value of the mutual information with respect to the continuous relevant variable Y, that is,

$$\widetilde{X}_j = \{X_j \in C_j \ ; \ I(X_j;Y) \geq I(X_i;Y); \quad \forall X_i \in C_j\} \tag{2}$$

2.2 Estimation of the Conditional Mutual Information for Regression Tasks

Given a set of N n-dimensional training samples $(\mathbf{x}_k, y_k)$, $k = 1, \ldots, N$ defined by the set of variables (features) $\mathbf{X} = (X_1, \ldots, X_n)$ where there is a dependency $y_k = f(\mathbf{x}_k)$, and for which a specific regressor can be applied, the conditional differential entropy $H(Y/\mathbf{X})$ can be written as [1]:

$$H(Y/\mathbf{X}) = -\int p(\mathbf{x}, y) \log p(y/\mathbf{x}) d\mathbf{x} dy \tag{3}$$

Considering that the joint probability distribution, $p(\mathbf{x}, y)$ can be approximated by the *empirical* distribution [15]: $p(\mathbf{x}, y) = \frac{1}{N} \cdot \sum_{k=1}^{N} \delta(\mathbf{x} - \mathbf{x}_k, y - y_k)$, where $\delta(\mathbf{x} - \mathbf{x}_k, y - y_k)$ is the Dirac delta function, and substituting $p(\mathbf{x}, y)$ into Eq. 3, we have:

$$H(Y/\mathbf{X}) = -\frac{1}{N} \cdot \sum_{k=1}^{N} \log p(y_k/\mathbf{x}_k) \tag{4}$$

Calculating Eq. 4 for one and for all pairs of two variables X_i, X_j, and substituting in Eq. 1, the *Dissimilarity matrix* of distances $D(X_i, X_j)$ can be obtained.

The assessment of $p(y/\mathbf{x})$ in Eq.4 is usually called Kernel Conditional Density Estimation (*KCDE*). This is a relatively recent active area of research that basically started with the works by Hyndman *et al* [8], among others. One way to obtain $p(y/\mathbf{x})$ is to use a (training) dataset $(\mathbf{x}_k, y_k)$ and a Nadaraya-Watson type kernel function estimator, as in [7], considering only the y_k training values that are paired with values $\mathbf{x}_k$:

$$\widehat{p}(y/\mathbf{x}) = \frac{\sum_k K_{h_1}(y - y_k) \cdot K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)}{\sum_k K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)} \tag{5}$$

where K_h is a compact symmetric probability distribution function, for instance, a gaussian kernel. In this case:

$$K_h(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} h^n |\Sigma|^{\frac{1}{2}}} \cdot \exp\left(-\frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2h^2}\right) \tag{6}$$

where Σ is a covariance matrix of a n-dimensional vector, $\mathbf{x}$. There are two bandwidths, h_1 for the y kernel and h_2 for the $\mathbf{x}$ kernel. The accuracy in the estimation of the conditional density functions is dependent on the assessment of the (h_1, h_2) parameters. The most common way to establish this accuracy would be the *Mean Integrated Square Error* [17] which can be defined (in this case) as:

$$MISE(h_1, h_2) = \int \left[p(y/\mathbf{x}) - \widehat{p}(y/\mathbf{x})\right]^2 dy p(\mathbf{x}) d\mathbf{x} \tag{7}$$

However, the following cross-validated log-likelihood defined in [7] will be used here because of its lower computational requirements:

$$L(h_1, h_2) = \frac{1}{N} \sum_k \log(\widehat{p}^{(-k)}(y_k/\mathbf{x}_k) \cdot \widehat{p}^{(-k)}(\mathbf{x}_k)) \tag{8}$$

where $\widehat{p}^{(-k)}$ means $\widehat{p}$ evaluated with $(\mathbf{x}_k, y_k)$ left out. $\widehat{p}(\mathbf{x})$ is the standard kernel density estimate over $\mathbf{x}$ using the bandwidth h_2 in Eq. 5. It can be shown that maximizing

the *KCDE* likelihood is equivalent to minimizing the *MISE* criterion. Substituting the Watson-Nadaraya type kernels into the $L(h_1, h_2)$ [7]:

$$L(h_1,h_2) = \frac{1}{N}\sum_k \log\left[\left(\frac{\sum_{j\neq k} K_{h_1}(y_k - y_j)K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{\sum_{j\neq k} K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}\right)\cdot\left(\sum_{j\neq k}\frac{K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{N-1}\right)\right]$$
$$= \frac{1}{N}\sum_k \log\left(\frac{\sum_{j\neq k} K_{h_1}(y_k - y_j)K_{h_2}(\|\mathbf{x}_k - \mathbf{x}_j\|)}{N-1}\right) \quad (9)$$

3 Experimental Validation

3.1 Methods

The proposed method, hereafter called MI_{Dist}, would consist of the steps described in Algorithm 1. Other three methods were used and compared to the method proposed here.

- Monteiro *et al* method [13] based on a Particle Swarm Optimization (*PSO*) strategy [10] (Particle-Swarms Feature Selection, $PS-FS$). It is a *Wrapper*-type method to make feature selection using an adaptation of an evolutionary computation technique developed by Kennedy and Eberhart [10]. For further details, see [13].
- Forward Stepwise Regression (FSR). Consider a linear regression model. The significance of each variable is determined from its t-statistics with the null hypothesis that the correlation between y and X_i is 0. The significance of factors is ranked using the p-values (of the t-statistics) and with this order a series of reduced linear models is built.
- Elastic Net (EN). It is a sparsity-based regularization scheme that simultaneously does regression and variable selection. It proposes the use of a penalty which is a weighted sum of the l_1-norm and the square of the l_2-norm of the coefficient vector formed by the weights of each variable. For further details, see [21].

3.2 Dataset Description

Four datasets were used to test the feature selection methods, divided into three groups:

- Hyperspectral datasets. Two hyperspectral datasets corresponding to a *Remote Sensing* campaign ($SEN2FLEX2005$, [14]) were used.
  1. *CASI-THERM*. It consists of the reflectance values of image pixels that were taken by the Compact Airborne Spectrographic Imager ($CASI$) sensor [14]. Corresponding thermal measurements for those pixels were also performed. The $CASI$ sensor images are formed by 144 bands between 370 and 1049nm.
  2. *CASI-AHS-CHLOR*. It consists of the reflectance values of image pixels that were taken by the $CASI$ and Airborne Hyper-spectral Scanner (AHS) [14] sensors. Corresponding chlorophyll measurements for those pixels were also performed. AHS images consist of 63 bands between 455 and 2492 nm.

1. **Kernel width estimation**. Obtain, for each pair and tuple $(Y; X_i)$ and $(Y; X_i, \widetilde{X}_j)$, the pair of parameters (h_1, h_2) that minimize $L(h_1, h_2)$ (Eq. 9).
2. **Kernel Density Estimation**. Obtain the Watson-Nadaraya type Kernel Density estimators $K_{h_1}(y - y_k)$ and $K_{h_2}(\|\mathbf{x} - \mathbf{x}_k\|)$ applying Eq. 6
3. **Assessment of the A-Posterior Probabilities**. Estimate $\widehat{p}(y/\mathbf{x})$ as $\widehat{p}(y/\mathbf{x}) \leftarrow \frac{\sum_k K_{h_1}(y-y_k) \cdot K_{h_2}(\|\mathbf{x}-\mathbf{x}_k\|)}{\sum_k K_{h_2}(\|\mathbf{x}-\mathbf{x}_k\|)}$.
4. **Estimation of the Conditional Entropies**. Obtain, for each variable $X = X_i$ and every possible combination $X = (X_i, \widetilde{X}_j)$ the Conditional Entropies: $H(Y/X) \leftarrow -\frac{1}{N} \cdot \sum_{k=1}^{N} \log p(y_k/\mathbf{x}_k)$.
5. **Construction of the** *Dissimilarity Matrix* $D(X_i, \widetilde{X}_j)$. Obtain $D(X_i, X_j) = I(X_i; Y/X_j) + I(X_j; Y/X_i) = H(Y/X_i) + H(Y/X_j) - H(Y/X_i, X_j)$
6. **Clustering**. Application of a Hierarchical Clustering strategy based on Ward's linkage method to find clusters in $D(X_i, \widetilde{X}_j)$. The number of clusters is determined by the number of variables to be selected.
7. **Representative selection**. For each cluster the variable with the highest value of the mutual information with respect to the continuous relevant variable Y is selected.

Algorithm 1. Selection of variables using Mutual Information based measures

- *Bank*32*NH*. It consists of 8192 cases, 4500 for training and 3692 for testing, with 32 continuous variables, corresponding to a simulation how bank-customers choose their banks. It can be found in the *DELVE* Data Repository.
- *Boston Housing*. Dataset created by D. Harrison et al [6]. It concerns the task of predicting housing values in areas of Boston. The whole dataset consists of 506 cases and 13 continuous variables. It can be found in the *UCI* Machine Learning Repository.

4 Results and Discussion

One (h_1, h_2) kernel width pair was obtained for each $(y_k, \mathbf{x}_k)$ pair, applying an Active Set method [5] for non-linear multi-variable optimization with constraints. The starting values were fixed at: $h_{1,0} = h_{2,0} = \frac{1}{2\log(N)}$, as in [11], and the lower and upper bounds at $[h_{i,m}, h_{i,M}] = [0.1 \cdot h_{i,0}, 10 \cdot h_{i,0}]$, $i = 1, 2$. For the assessment of $p(y/\mathbf{x})$ when $\mathbf{x} = (X_i, X_j)$, the covariance matrix considered was diagonal: $\Sigma = diag(\sigma_i^2, \sigma_j^2)$, where σ_i^2 and σ_j^2 are the variance value of variables i and j, respectively, for the training set. For the selection of the variable that represents a cluster, Eq.2 was applied.

Support Vector Regression (SVR) [2] with a radial (gaussian) basis function was used for regression, and the Root Mean Squared Error ($RMSE$) as the performance criterion. For each one of the four datasets, an exhaustive grid search using equally spaced steps in the logarithmic space of the *tuning* SVR parameters (C, ε, σ) was made to select and fix the best parameters for each one of these datasets. A 10-fold cross-validation strategy was used to obtain the $RMSE$ error on the *Boston Housing* dataset.

Figure 1 shows the *RMSE* error given by the *SVR* method for the four datasets and the first 20 variables selected (13 variables for *Boston Housing*) by each one of the variable selection methods tested. Tables 1 to 3 show the $RMSE$ Error over the first 5,

Table 1. $RMSE$ Error over the first 5 variables. (+ = positive), (− = negative).

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.9160**	3.1266	4.5636	3.3524	6.53 (-)	7.24 (+)
CASI-THERM	**3.3267**	3.3892	3.6425	3.4381	0.08 (-)	0.02 (-)
Bank32NH	0.0961	0.0953	**0.0950**	**0.0950**	0.48 (-)	1.03 (-)
Boston Housing	4.4270	**4.3702**	4.8016	4.6252	1.43 (-)	1.44 (-)
Average	**2.6914**	2.7454	3.2757	2.8777		

Table 2. $RMSE$ Error over the first 10 variables. (+ = positive), (− = negative).

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.3970**	2.5448	4.1506	2.8153	28.04 (+)	20.73 (+)
CASI-THERM	**3.1912**	3.2866	3.3585	3.2501	1.17 (-)	0.93 (-)
Bank32NH	0.0930	0.0913	**0.0910**	0.0913	4.14 (-)	6.21 (+)
Boston Housing	**4.2035**	4.3264	4.7022	4.8756	6.73 (+)	5.98 (+)
Average	**2.4712**	2.5623	3.0756	2.7581		

Table 3. $RMSE$ Error over the first 15 variables (13 for *Boston Housing*). (+ = positive), (− = negative).

Database	MI_{Dist}	PS−FS	FSR	EN	Friedman Test	Quade Test
CASI-AHS-CHLOR	**2.1964**	2.2922	3.7828	2.5497	49.13 (+)	30.43 (+)
CASI-THERM	**3.2055**	3.2771	3.3027	3.2308	2.17 (-)	1.68 (-)
Bank32NH	0.0920	0.0908	**0.0905**	0.0907	1.21 (-)	7.71 (+)
Boston Housing	**4.3171**	4.5162	4.7995	4.9491	7.84 (+)	8.08 (+)
Average	**2.4517**	2.5436	2.9950	2.7051		

10 and 15 variables (13 variables for *Boston Housing*) for all the methods and datasets selected. Friedman and Quade Tests [3], [4] were applied on the results with a confidence level of $p = 0.005$. The Fisher distribution critical value for the four methods and over the first 5, 10 and 15 variables (13 for *Boston Housing*) was set up, obtaining $F(3, 12) = 7.20, F(3, 27) = 5.36, F(3, 42) = 4.94$. The differences in $RMSE$ ranked for the four methods are not significant for the first 5 features, but they are for 10 to 15 features. Thus, the difference between the methods increases with the number of selected features, although in the case of *CASI-THERM* database the statistical tests are not significant.

The proposed method, MI_{Dist}, obtains better performance with respect to the rest of methods for all the cases (5, 10 and 15 variables) for the *CASI-AHS-CHLOR* and *CASI-THERM* datasets and for two out of the three (10 and 13 variables) for the *Boston Housing* dataset. *PS−FS* method is the second best one in most cases followed by the *EN* method, while *FSR* behaves worst. When averaging each method over the four databases, MI_{Dist} obtains lower values in the three tables.

The selection of the first one and the first two variables is better for MI_{Dist} compared to the rest of the methods. In this case, the clustering strategy plays an important role in the formation of different groupings of features, obtaining better results than a greedy

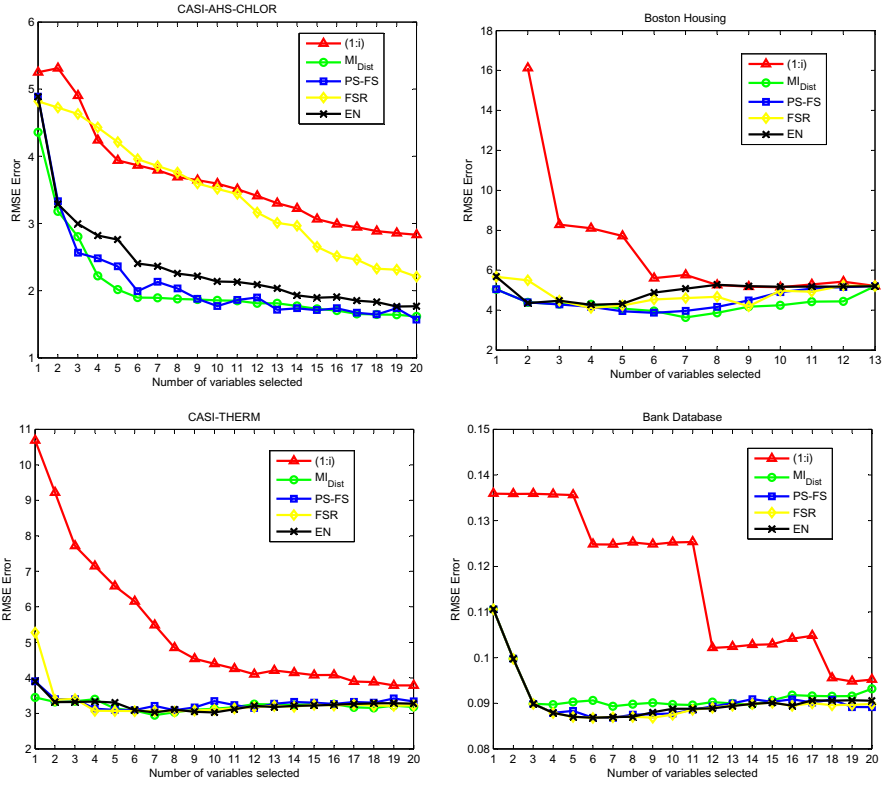


Fig. 1. $RMSE$ Error using SVR for the *CASI-AHS-CHLOR*, *Boston Housing*, *CASI-THERM* and *Bank*32*NH* datasets, respectively. The first point in the $(1:i)$ line for *Boston Housing* is not shown because it is of the order of ~ 2000.

selection algorithm as is the case of *FSR*. In the cases of *PS−FS* and *EN* methods the possible advantage of our method consists in a proper adjustment of the parameters from Nadaraya−Watson function estimator and its use through a distance metric in feature space that takes into account the internal relationships between features.

5 Conclusions

This paper presents a *filter*-type method to do feature selection for regression, using a *Dissimilarity matrix* based on conditional mutual information measures. This matrix is an extension for continuous variables of a Dissimilarity matrix used by Sotoca *et al* in [18] for classification. The method is compared against three other methods on four datasets of different nature, using the $RMSE$ Error given by the SVR technique. Authors are currently working on the performance analysis with other types of regressors and datasets and analyzing the effect of noise in the data on the performance of the selection strategy.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons Inc., Chichester (1991)
2. Drucker, H., Burges, C., Kaufman, L., Kaufman, L., Smola, A., Vapnik, V.: Support Vector Regression Machines. In: NIPS 1996, pp. 155–161 (1996)
3. Friedmann, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32(200), 675–701 (1937)
4. García, S., Fernández, A., Luengo, J., Herrera, F.: Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180, 2044–2064 (2010)
5. Gill, P.E., Murray, W., Wright, M.H.: Practical Optimization. Academic Press, London (1981)
6. Harrison, D., Rubinfeld, D.L.: Hedonic prices and the demand for clean air. J. Environ. Econ. Manag. 5, 81–102 (1978)
7. Holmes, M.P., Gray, A., Isbell, C.L.: Fast kernel conditional density estimation: A dual-tree Monte Carlo approach. Comput. Stat. Data Analysis 54(7), 1707–1718 (2010)
8. Hyndman, R.J., Bashtannyk, D.M., Grunwald, G.K.: Estimating and visualizing conditional densities. Journal of Computation and graphical Statistics 5(4), 315–336 (1996)
9. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. PAMI 22(1), 4–37 (2000)
10. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: IEEE ICNN, pp. 1942–1948 (1995)
11. Kwak, N., Choi, C.-H.: Input feature selection by mutual information based on parzen window. IEEE Trans. PAMI 24(12), 1667–1671 (2002)
12. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis 1, 131–156 (1997)
13. Monteiro, S.T., Kosugi, Y.: Particle Swarms for Feature Extraction of Hyperspectral Data. IEICE Trans. Inf. and Syst. E90D(7), 1038–1046 (2007)
14. Moreno, J.F.: Sen2flex data acquisition report, Universidad de Valencia, Tech. Rep. (2005)
15. Ney, H.: On the relationship between classification error bounds and training criteria in statistical pattern recognition. In: Perales, F.J., Campilho, A.C., Pérez, N., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 636–645. Springer, Heidelberg (2003)
16. Pudil, P., Ferri, F.J., Novovicova, J., Kittler, J.: Floating search methods for feature selection with nonmonotonic criterion functions. Pattern Recognition 2, 279–283 (1994)
17. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832–837 (1956)
18. Sotoca, J.M., Pla, F.: Supervised feature selection by clustering using conditional mutual information-based distances. Pattern Recognition 43(6), 2068–2081 (2010)
19. Verleysen, M., Rossi, F., François, D.: Advances in Feature Selection with Mutual Information. In: Biehl, M., Hammer, B., Verleysen, M., Villmann, T. (eds.) Similarity-Based Clustering. LNCS, vol. 5400, pp. 52–69. Springer, Heidelberg (2009)
20. Ward, J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58(301), 236–244 (1963)
21. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 67(part 2), 301–320 (2005)

Dual Layer Voting Method for Efficient Multi-label Classification

Gjorgji Madjarov[1,2], Dejan Gjorgjevikj[1], and Sašo Džeroski[2]

[1] FEEIT, Ss. Cyril and Methodius University, Skopje, Macedonia
[2] DKT, Jožef Stefan Institute, Ljubljana, Slovenia
{madzarovg,dejan}@feit.ukim.edu.mk, Saso.Dzeroski@ijs.si

Abstract. A common approach for solving multi-label classification problems using problem-transformation methods and dichotomizing classifiers is the pairwise decomposition strategy. One of the problems with this approach is the need for querying a quadratic number of binary classifiers for making a prediction that can be quite time consuming, especially in classification problems with large number of labels. To tackle this problem we propose a Dual Layer Voting Method (DLVM) for efficient pair-wise multiclass voting to the multi-label setting, which is related to the calibrated label ranking method. Five different real-world datasets (enron, tmc2007, genbase, mediamill and corel5k) were used to evaluate the performance of the DLVM. The performance of this voting method was compared with the majority voting strategy used by the calibrated label ranking method and the quick weighted voting algorithm (QWeighted) for pair-wise multi-label classification. The results from the experiments suggest that the DLVM significantly outperforms the concurrent algorithms in term of testing speed while keeping comparable or offering better prediction performance.

Keywords: Multi-label classification, calibration label, calibrated label ranking, voting strategy.

1 Introduction

The problem of traditional single-label classification is concerned with learning from examples each of which is associated with a single label λ_i from a finite set of disjoint labels $L = \{\lambda_1, \lambda_2, ..., \lambda_Q\}, Q > 1$. If $Q = 2$, then the learning problem is called a binary classification problem, while if $Q > 2$, we are talking about a multi-class classification problem. On the other hand, multi-label classification is concerned with learning from a set of examples each of which is associated with a set of labels $Y \subseteq L$ i.e each example can be a member of more than one class.

A common approach to address the multi-label classification problem is utilizing class binarization methods, i.e. decomposition of the problem into several binary subproblems that can then be solved using a binary base learner. The simplest strategy in the multi-label setting is the one-against-all strategy also

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 232–239, 2011.

referred to as the binary relevance method. It addresses the multi-label classification problem by learning one classifier (model) M_k $(1 \leq k \leq Q)$ for each class, using all the examples labeled with that class as positive examples and all other (remaining) examples as negative examples. At query time, each binary classifier predicts whether its class is relevant for the query example or not, resulting in a set of relevant labels.

Another approach for solving the multi-label classification problem using binary classifiers is pair-wise classification or round robin classification [1][2]. Its basic idea is to use $Q * (Q-1)/2$ classifiers covering all pairs of labels. Each classifier is trained using the samples of the first label as positive examples and the samples of the second label as negative examples. To combine these classifiers, the pair-wise classification method naturally adopts the majority voting algorithm. Given a test instance, each classifier delivers a prediction for one of the two labels. This prediction is decoded into a vote for one of the labels. After the evaluation of all $Q * (Q-1)/2$ classifiers the labels are ordered according to their sum of votes. To predict only the relevant classes for each instance a label ranking algorithm is used.

Brinker et al. [3] propose a conceptually new technique for extending the common pair-wise learning approach to the multi-label scenario named calibrated label ranking (CLR). The key idea of calibrated label ranking is to introduce an artificial (calibration) label λ_0, which represents the split-point between relevant and irrelevant labels. The calibration label λ_0 is assumed to be preferred over all irrelevant labels, but all relevant labels are preferred over it. At prediction time (when majority voting strategy is usually used), one will get a ranking over $Q+1$ labels (the Q original labels plus the calibration label).

Besides the majority voting that is usually used strategy in the prediction phase of the calibrated label ranking algorithm, Park et al. [4] propose another more effective voting algorithm named Quick Weighted Voting algorithm (QWeighted). QWeighted computes the class with the highest accumulated voting mass avoiding the evaluation of all possible pair-wise classifiers. An adaptation of QWeighted to multi-label classification (QWeightedML) [5] is to repeat the process while all relevant labels are not determined i.e. until the returned class is the artificial label λ_0, which means that all remaining classes will be considered to be irrelevant.

In this paper we propose an efficient Dual Layer Voting Method (DLVM) that modifies the majority voting algorithm for calibrated label ranking technique [6]. We have evaluated the performance of this algorithm on a selection of multi-label datasets that vary in terms of problem domain, number of labels and label cardinality. The results demonstrate that our modification outperforms the majority voting algorithm for calibrated label ranking algorithm [6] and the QWeightedML [5] algorithm in terms of testing speed, while keeping comparable prediction results.

For the readers' convenience, in Section 2 we will introduce the Dual Layer Voting Method. Section 3 presents the computational complexity of DLVM

comparing to CLR. The experimental results that compare the performance of the proposed DLVM with concurrent methods are presented in Section 4. Section 5 gives a conclusion.

2 Dual Layer Voting Method (DLVM)

Conventional pair-wise approach learns a model M_{ij} for all combinations of labels λ_i and λ_j with $1 \leq i < j \leq Q$. This way $Q*(Q-1)/2$ different pair-wise models are learned. The main disadvantage of this approach is that in the prediction phase a quadratic number of base classifiers (models) have to be consulted for each test example.

As a result of introducing the artificial calibration label λ_0 in the calibrated label ranking algorithm [6], the number of the base classifiers is increased by Q i.e. additional set of Q binary preference models M_{0k} $(1 \leq k \leq Q)$ is learned. The models M_{0k} that are learned by a pair-wise approach to calibrated ranking, and the models M_k that are learned by conventional binary relevance are equivalent.

The binary relevance models M_k $(1 \leq k \leq Q)$ almost always have bigger time complexity than pair-wise models M_{ij} $(1 \leq i < j \leq Q)$ because they are learned with all the examples from the training set, while the pair-wise models M_{ij} are learned only with the examples labeled with labels λ_i and λ_j. In standard voting algorithm for the calibrated label ranking method each test example needs to consult all the models (classifiers) M_k $(1 \leq k \leq Q)$ and M_{ij} $(1 \leq i < j \leq Q)$ in order to rank the labels by their order of preference. As a result of increased number of classifiers the CLR method leads to more accurate prediction but also leads to slower testing time and bigger computational complexity especially when the number of the labels in the problem is big.

In this paper we propose an efficient Dual Layer Voting Method (DLVM) for multi-label classification that is related to the CLR algorithm [6]. It reduces the number of base classifiers that are needed to be consulted in order to make a final prediction for a given test example. This method leads to an improvement in recognition speed, while keeping comparable or offering better prediction results.

The architecture of the DLVM is organized in two layers (Figure 1). In the first layer Q classifiers are located, while in the second layer there are $Q*(Q-1)/2$ classifiers. The classifiers in the first layer are binary relevance models M_{0k}, while the pair-wise models M_{ij} are located in the second layer of the architecture. Each model M_{0k} from the first layer is connected to $Q-1$ models M_{ij} from the second layer, where $k = i$ or $k = j$ $(1 \leq i \leq Q-1, i+1 \leq j \leq Q)$.

In the prediction phase, each model M_{0k} tries to determine the relevant labels for the corresponding test example. Each model M_{0k} gives the probability (the output value of the model M_{0k} is converted to probability) that the test example is associated with the label λ_k. If that probability is appropriately small (under some predetermined threshold), we can conclude that the artificial calibration label λ_0 is preferred over the label λ_k i.e. the label λ_k belongs to the set of irrelevant labels. In such case, one can conclude that for the corresponding test example, the pair-wise models of the second layer M_{ij} where $i = k$ or $j = k$, need not be consulted, because the binary relevance model M_{0k} from the first

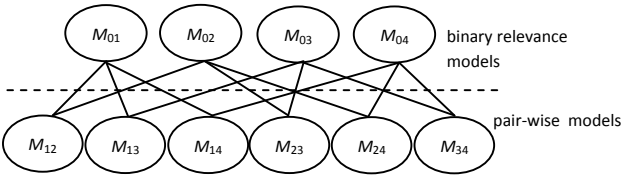


Fig. 1. Architecture of the DLVM

layer has suggested that the label λ_k belongs to the set of irrelevant labels. For each test example for which it is known that the label λ_k belongs to the set of irrelevant labels, the number of pair-wise models that should be consulted decreases for $Q-1$.

In order to decide which labels belong to the set of the irrelevant labels i.e. which pair-wise models M_{ij} from the second layer do not have to be consulted a threshold t $(0 \leq t \leq 1)$ is introduced. In our experiments the value of the threshold t was determined by cross-validation.

As previously described, every test example first consults all binary relevance models M_{0k} of the first layer. The output value of each corresponding model M_{0k} $(1 \leq k \leq Q)$ is converted to probability and compared to the threshold t.

- If the prediction probability is above the threshold, the test example is forwarded to all the models M_{ij} of the second layer of the architecture of the DLVM that are associated to the model M_{0k}.
- If the prediction probability is under the threshold, the test example is not forwarded to any model of the second layer.

Observed from the side of the pair-wise models and considering the prediction probabilities of the binary relevance models of the first layer, three distinct cases in the voting process of each pair-wise model of the second layer can appear:

1. The prediction probabilities of both binary relevance models M_{0i} and M_{0j} that are connected to the pair-wise model M_{ij} are above the threshold t.
2. The prediction probability of only one of the binary relevance models (M_{0i} or M_{0j}) is above the threshold t.
3. The prediction probabilities of the binary relevance models M_{0i} and M_{0j} are both under the threshold t.

In the first case the model M_{ij} is consulted and its prediction is decoded into a vote for one of the labels λ_i or λ_j. In the second case, the model M_{ij} is not consulted and its vote goes directly to the label whose binary relevance model prediction probability is above the threshold t. In the third case the model M_{ij} is not consulted and it does not vote at all. Following this approach, each consulted pair-wise model M_{ij} of the second layer tries to determine which of the labels (λ_i or λ_j) is preferred over the other.

By increasing the value of the threshold, the number of pair-wise models that should be consulted decreases. For $t = 1$ no example is forwarded to the second layer. On the other hand, for $t = 0$, for each test example all pair-wise models of the second layer are consulted.

3 Computational Complexity

The computational complexity of the calibrated label ranking method can be defined as a sum of the computational complexity of the binary relevance models and the pair-wise models:

$$O_{CLR} = O_{BR} + O_P \tag{1}$$

where O_{BR} and O_P are the computational complexities of the binary relevance models and the pair-wise models consequently.

On the other hand, computational complexity of DLVM can be defined as a sum of computational complexity of the models located in the first layer of the architecture (O_{FL}) and computational complexity of the models located in the second layer of the architecture (O_{SL}):

$$O_{DLVM} = O_{FL} + O_{SL} \tag{2}$$

The computational complexity of the first layer of the DLVM and the computational complexity of the binary relevance models of the CLR method are equal ($O_{BR} = O_{FL}$). The main difference of computational complexity between CLR and DLVM is in the computational complexity of the pair-wise models of the CLR and the second layer of DLVM. As noted in the previous section, if the threshold $t = 1$ no models of the second layer will be consulted so O_{SL} will be 0 and $O_{DLVM} = O_{FL} = O_{BR}$. If the threshold $t = 0$ in the DLV method, all models of the second layer will be consulted and the number of consulted pair-wise models becomes $Q * (Q - 1)/2$ ($O_{SL} = O_P$). For the threshold values $0 < t < 1$, $O_{SL} = r * O_P$ where r is a reduction parameter specific for each multi-label dataset ($0 < r < 1$). For a real world problem the reduction parameter r can be determined as:

$$r = \frac{amf * (amf - 1)}{Q * (Q - 1)} \tag{3}$$

where amf is the average number of binary relevance models located in the first layer of DLVM that give a probability that is above the threshold t in the prediction process. For an ideal case (prediction accuracy of 100% by the binary relevance models) amf is getting equal to the label cardinality lc ($amf = lc$) which is introduced by Tsoumakas et al. [14]. The label cardinality is the average number of labels per example in a multi-label dataset.

4 Experimental Results

The performance of the proposed method was measured with five different multi-label evaluation metrics (Hamming loss, one error, coverage, ranking loss and average precision) proposed in [7] on the problems of recognition of text, video,

Table 1. Datasets description

	Domain	$\#training\ s.$	$\#test\ s.$	$\#features$	$\#labels$	lc
enron	text	1123	579	1001	53	3.38
tmc2007	text	21519	7077	49060	22	2.16
genbase	biology	463	199	1186	27	1.25
mediamill	video	30993	12914	120	101	4.376
corel5k	images	4500	500	499	374	3.52

Table 2. The prediction performances of each method for every dataset

	Evaluation metrics	**enron**	**tmc2007**	**genbase**	**mediamill**	**corel5k**
CLR	Hamming Loss	0.0476	0.0177	0.0011	0.2082	0.0095
	One-error	0.2297	0.0411	0.0	0.4426	0.6740
	Coverage	11.519	1.4213	0.5778	321.883	99.57
	Ranking Loss	0.0756	0.0065	0.0077	0.1527	0.1106
	Avg. Precision	0.7018	0.9630	0.9923	0.4310	0.2918
QWeightedML	Hamming Loss	0.0481	0.0263	0.0011	0.2082	0.0095
	One-error	0.2262	0.0821	0.0	0.4426	0.6640
	Coverage	20.333	2.0333	0.2864	364.8	218.12
	Ranking Loss	0.1516	0.0265	0.0012	0.1645	0.2826
	Avg. Precision	0.6543	0.9233	0.9950	0.4290	0.2190
DLVM	Hamming Loss	0.0501	0.0228	0.0011	0.2082	0.0094
	One-error	0.2193	0.0631	0.0	0.4429	0.6640
	Coverage	14.431	1.7452	0.2864	364.8	172.81
	Ranking Loss	0.0969	0.0212	0.0012	0.2045	0.1417
	Avg. Precision	0.6970	0.9342	0.9950	0.4298	0.2870
	threshold t	0.03	0.1	0.35	0.015	0.01
	amf	18.46	4.93	1.261	27.79	74.24
	reduction r	0.117	0.0419	0.00047	0.0737	0.039

Table 3. Testing times of each method for every dataset measured in seconds

	enron	**tmc2007**	**genbase**	**mediamill**	**corel5k**
CLR	605.1	6106.6	0.187	4555.7	84.4
QWeightedML	174.31	2534.3	0.172	2282.6	16.5
DLVM	146.57	1135.8	0.141	456.8	9.85

images and protein function. DLVM is compared with the calibrated label ranking method with majority voting strategy for pair-wise multi-label classification (CLR) [6] and the QWeightedML algorithm [5].

In our experiments, five different multi-label classification problems were addressed by each classifying method. The recognition performance and the testing time were recorded for every method. The problems considered in the experiments include text [11](enron) [12](tmc2007), protein function [13] (genbase), video [15] (mediamill) and images [16] (corel5k) classification. The complete

description of the datasets (domain, number of training and test instances, number of features, number of labels and label cardinality) is shown in Table 1.

The training and testing of the DLVM was performed using a custom developed application that uses the MULAN library [8] for the machine learning framework Weka [9]. The LIBSVM library [10] utilizing the SVMs with radial basis kernel were used for solving the partial binary classification problems for the enron, tmc2007 and mediamill datasets. The kernel parameter *gamma* and the penalty C for the datasets were determined by 5-fold cross validation using only the samples of the training sets. For the remaining datasets (genbase and corel5k), the C4.5 decision tree was used for solving the binary classification problems. For these two datasets, C4.5 was chosen instead of SVM in order to avoid the computational complexity in the training phase which appeared when we tried to solve these problems using SVMs. The outputs of the SVM classifiers are converted to probabilities according to [17], while the outputs of the C4.5 classifiers are already a probabilities [18]. In all classification problems the classifiers were trained using all available training examples and were evaluated by recognizing all test examples from the corresponding dataset.

Tables 2 and 3 give the predictive performance and the testing times measured in seconds of each method applied on each of the datasets. The first column of the Table 2 describes the methods. The second column describes the evaluation metrics and the remaining columns show the performance of each method for every dataset. The last three rows show the values of the threshold t for each dataset separately, for which the presented results of DLVM are obtained, the values of the amf and the reduction parameter r. The value of the threshold t for each dataset was determined by 5-fold cross validation using only the samples of the training set in order to achieve maximum benefit in terms of prediction results on testing speed. Table 2 shows that DLVM offers better predictive performance than QWeightedML method in most of the evaluation metrics, while showing comparable performance to CLR. On the other hand, Table 3 clearly shows that among the three tested approaches DLVM offers best performance in terms of testing speed. The results show that for the five treated classification problems DLVM is 1.5 to 10 times faster than the calibrated label ranking algorithm with majority voting and 20% to 400% faster than the QWeightedML method. The reduction of the testing time of the DLVM over the CLR becomes even more notable as the number of labels in the treated classification problem increases. The experiments showed that for the enron, mediamill and corel5k datasets with quite big number of labels the testing time of DLVM is four, ten and eight times shorter compared to CLR, respecively.

5 Conclusion

A Dual Layer Voting Method (DLVM) for efficient pair-wise multiclass voting to the multi-label setting was presented. The performance of the proposed method was compared with the calibrated label ranking method with majority voting strategy for pair-wise multi-label classification and the QWeightedML

algorithm on five different real-world datasets (enron, tmc2007, genbase, mediamill and corel5k). The results show that the DLVM significantly outperforms the calibrated label ranking method with majority voting and the QWeightedML algorithm in term of testing speed while keeping comparable or offering better prediction performance. DLVM was 1.5 to 10 times faster than calibrated label ranking algorithm with majority voting and 20% to 400% faster than the QWeightedML method.

References

1. Fürnkranz, J.: Round robin classification. Journal of Machine Learning Research 2(5), 721–747 (2002)
2. Wu, T.F., Lin, C.J., Weng, C.R.: Probability estimates for multiclass classification by pairwise coupling. Journal of Machine Learning Research 5(8), 975–1005 (2004)
3. Brinker, K., Fürnkranz, J., Hullermeier, E.: A unified model for multilabel classification and ranking. In: 17th European Conference on Artificial Intelligence, Riva Del Garda, Italy, pp. 489–493 (2006)
4. Park, S.H., Fürnkranz, J.: Efficient pairwise classification. In: 18th European Conference on Machine Learning, Warsaw, Poland, pp. 658–665 (2007)
5. Loza Mencía, E., Park, S.H., Furnkranz, J.: Efficient voting prediction for pairwise multi-label classification. Neurocomputing 73, 1164–1176 (2010)
6. Fürnkranz, J., Hullermeier, E., Loza Mencia, E., Brinker, K.: Multi-label classification via calibrated label ranking. Machine Learning 73(2), 133–153 (2008)
7. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization. Machine Learning 39(2), 135–168 (2000)
8. `http://mulan.sourceforge.net/`
9. `http://www.cs.waikato.ac.nz/ml/weka/`
10. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
11. `http://bailando.sims.berkeley.edu/enron_email.html`
12. Srivastava, A., Zane-Ulman, B.: Discovering recurring anomalies in text reports regarding complex space systems. In: Proceedings of the IEEE Aerospace Conference, pp. 55–63 (2005)
13. Diplaris, P.M.S., Tsoumakas, G., Vlahavas, I.: Protein classification with multiple algorithms. In: Proceedings of 10th Panhellenic Conference on Informatics, Volos, Greece, pp. 448–456 (2005)
14. Tsoumakas, G., Katakis, I.: Multi label classification: An overview. International Journal of Data Warehousing and Mining 3 (2007)
15. Snoek, C.G.M., Worring, M., Van Gemert, J.C., Geusebroek, J.-M., Smeulders, A.W.M.: The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In: Proceedings of ACM Multimedia, Santa Barbara, USA, pp. 421–430 (2006)
16. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
17. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. MIT Press, Cambridge (1999)
18. Quinlan, J.R.: C4.5:Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

Passive-Aggressive for On-Line Learning in Statistical Machine Translation

Pascual Martínez-Gómez, Germán Sanchis-Trilles, and Francisco Casacuberta

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{pmartinez,gsanchis,fcn}@dsic.upv.es

Abstract. New variations on the application of the passive-aggressive algorithm to statistical machine translation are developed and compared to previously existing approaches. In online adaptation, the system needs to adapt to real-world changing scenarios, where training and tuning only take place when the system is set-up for the first time. Post-edit information, as described by a given quality measure, is used as valuable feedback within the passive-aggressive framework, adapting the statistical models on-line. First, by modifying the translation model parameters, and alternatively, by adapting the scaling factors present in state-of-the-art SMT systems. Experimental results show improvements in translation quality by allowing the system to learn on a sentence-by-sentence basis.

Keywords: on-line learning, passive-aggressive, statistical machine translation.

1 Introduction

Online passive-aggressive (PA) algorithms [4] are a family of margin-based online learning algorithms that are specially suitable for adaptation tasks where a convenient change in the value of the parameters of our models is desired after every sample is presented to the system. The general idea is to learn a weight vector representing a hyperplane such that differences in quality also correspond to differences in the margin of the instances to the hyperplane. The update is performed in a characteristic way by trying to achieve at least a unit margin on the most recent example while remaining as close as possible to the current hyperplane.

Different ways to apply the PA framework to statistical machine translation (SMT) are analised. SMT systems use mathematical models to describe the translation task and to estimate the probability of translating a source sentence $\mathbf{x}$ into a target sentence $\mathbf{y}$. Recently, a direct modelling of the posterior probability $\Pr(\mathbf{x} \mid \mathbf{y})$ has been widely adopted. To this purpose, different authors [11,8] propose the use of the so-called log-linear models, where the decision rule is given by the expression

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) = \operatorname*{argmax}_{\mathbf{y}} \boldsymbol{\lambda}\mathbf{h}(\mathbf{x}, \mathbf{y}) = \operatorname*{argmax}_{\mathbf{y}} s(\mathbf{x}, \mathbf{y}) \tag{1}$$

where $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, M is the number of models (or features) and λ_m are the weights of the log-linear combination. $s(\mathbf{x}, \mathbf{y})$ is a score representing how good $\mathbf{x}$ translates

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 240–247, 2011.

into $\mathbf{y}$. Common feature functions $h_m(\mathbf{x}, \mathbf{y})$ include different translation models (TM), but also distortion models or even the target language model. $\mathbf{h}(\cdot|\cdot)$ and $\boldsymbol{\lambda}$ are estimated by means of training and development sets, respectively.

In order to capture context information, *phrase-based* models [16] were introduced, widely outperforming single word models [3]. The main idea is to segment source sentence $\mathbf{x}$ into *phrases* (i.e. word sequences), and then to translate each source phrase $\tilde{x}_k \in \mathbf{x}$ into a target phrase $\tilde{y}_k$. Those models were employed throughout this work.

Adjusting both $\mathbf{h}$ or $\boldsymbol{\lambda}$ leads to an important problem in SMT: whenever the text to be translated belongs to a different domain than the training corpora, translation quality diminishes significantly [3]. For this reason, the problem of *adaptation* is very common in SMT, where the objective is to improve the performance of systems trained and tuned on out-of-domain data by using very limited amounts of in-domain data.

Adapting a system to changing tasks is specially interesting in the Computer Assisted Translation (CAT) [2] and Interactive Machine Translation (IMT) paradigms [1], where the collaboration of a human translator is essential to ensure high quality results. Here, the SMT system proposes a hypothesis to a human translator, who may amend the hypothesis to obtain an acceptable translation, and after that expects the system to learn from its own errors, so that it is not necessary to correct the same error again. The challenge is then to make the best use of every correction provided by adapting the system *online*, i.e. without performing a complete retraining which is too costly.

In this work, the performance of PA with two adaptation strategies is analysed, namely feature vector and scaling factor adaptation, with the purpose of using feedback information to improve subsequent translations in a sentence-by-sentence basis.

Similar work is briefly detailed in the following Section. PA algorithms are reviewed in Section 3. Their application to SMT is described in Section 4. Experiments conducted are analysed in Section 5, and conclusions and future work are listed in Section 6.

2 Related Work

In [10], an online learning application is presented for IMT, incrementally updating model parameters by means of an incremental version of the Expectation-Maximisation algorithm and allowing for the inclusion of new phrase pairs. We propose the use of a dynamic reranking algorithm which is applied to a $nbest$ list, regardless of its origin. In addition, in [10], only $\mathbf{h}$ is adapted, whereas here we also analyse the adaptation of $\boldsymbol{\lambda}$.

In [13] the authors propose the use of the PA framework [4] for updating the feature functions $\mathbf{h}$. The obtained improvements were very limited, since adapting $\mathbf{h}$ is a very sparse problem. Hence, in the present paper, the adaptation of the $\boldsymbol{\lambda}$ will be compared to the adaptation of $\mathbf{h}$, which is shown in [14] to be a good adaptation strategy. In [14], the authors propose the use of a Bayesian learning technique in order to adapt the scaling factors based on an adaptation set. In contrast, our purpose is to perform online adaptation, i.e. to adapt system parameters after each new sample has been provided.

Another difference between [13] and the present work is that they propose to model the user feedback by means of BLEU score [12], which is quite commonly used in SMT. Such score measures precision of n-grams with a penalty for sentences that are too short. However, BLEU is not well defined on the sentence level, since it implements a geometric average which is zero whenever no common 4-gram exists between reference

and hypothesis. In the present work, we propose the use of TER [15] instead. TER is similar to the word error rate criterion of speech recognition, but allowing shifts of word sequences. TER is well defined on the sentence level, and, furthermore, in [15] it is shown to correlate better with human judgement.

3 The Passive-Aggressive Algorithm

PA [4] is a family of margin-based, on-line learning algorithms that update model parameters after each observation has been seen. In this case, PA is applied to a regression problem, where target value $\hat{\mu}(\mathbf{y})_t \in \mathbb{R}$ has to be predicted by the system for input observation $\mathbf{x}_t \in \mathbb{R}^n$ at time t by using a linear regression function $\hat{\mu}(\mathbf{y})_t = \mathbf{w}_t \cdot \mathbf{x}_t$.

After every prediction, the true target value $\mu(\mathbf{y})_t \in \mathbb{R}$ is received and the system suffers an instantaneous loss according to a sensitivity parameter ϵ:

$$l_\epsilon(\mathbf{w}; (\mathbf{x}, \mu(\mathbf{y}))) = \begin{cases} 0 & \text{if } |\mathbf{w} \cdot \mathbf{x} - \mu(\mathbf{y})| \leq \epsilon \\ |\mathbf{w} \cdot \mathbf{x} - \mu(\mathbf{y})| - \epsilon & \text{otherwise} \end{cases} \quad (2)$$

If the system's error falls below ϵ, the loss suffered by the system is zero and the algorithm remains *passive*, that is, $\mathbf{w}_{t+1} = \mathbf{w}_t$. Otherwise, the loss grows linearly with the error $|\hat{\mu}(\mathbf{y}) - \mu(\mathbf{y})|$ and the algorithm *aggressively* forces an update of the parameters.

The idea behind the PA algorithm is to modify the parameter values of the regression function so that it achieves a zero loss function on the current observation $\mathbf{x}_t$, while remaining as close as possible to the previous weight vector $\mathbf{w}_t$. That is, formulated as an optimisation problem subject to a constraint [4]:

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w} - \mathbf{w}_t||^2 + C\xi^2 \qquad \text{s.t.} \quad l_\epsilon(\mathbf{w}; (\mathbf{x}, \mu(\mathbf{y}))) = 0 \quad (3)$$

where ξ^2 is, according to the so-called PA Type-II, a squared slack variable scaled by the aggressivity factor C. As in classification tasks, it is common to add a slack variable into the optimisation problem to get more flexibility during the learning process.

It is only left to add the constraint together with a Lagrangian variable and set the partial derivatives to zero to obtain the closed form of the update term. In Section 4, the update term for every adaptation strategy ($\boldsymbol{\nu}_t$ and $\hat{\boldsymbol{\lambda}}_t$) is detailed.

4 Passive-Aggressive in SMT

4.1 Feature Vector Adaptation

As described in [13], PA can be used for adapting the translation scores within state-of-the-art TMs. First, we need to define $h_{TM}(\mathbf{x}, \mathbf{y})$ as the combination of n TMs implicit in current translation systems, which are typically specified for all phrase pairs $(\tilde{x}_k, \tilde{y}_k)$:

$$h_{TM}(\mathbf{x}, \mathbf{y}) = \sum_n \lambda_n \sum_k h_n(\tilde{x}_k, \tilde{y}_k) \quad (4)$$

where h_{TM} can be considered as a single feature function h in Eq. 1. Then, we can study the effect of adapting the TMs in an online manner by adapting h_{TM}. Although there might be some reasons for adapting all the score functions $\mathbf{h}$, in the present paper

we focus on analysing the effect of adapting only the TMs. By considering $\forall m \notin TM$: $h_m^t(\cdot,\cdot) = h_m(\cdot,\cdot)$, and defining an auxiliary function $\mathbf{u}_t(\mathbf{x},\mathbf{y})$ such that

$$h_{TM}^t(\mathbf{x},\mathbf{y}) = \sum_n \lambda_n \sum_k u_t(\tilde{x}_k,\tilde{y}_k) h_n(\tilde{x}_k,\tilde{y}_k) = \mathbf{u}_t(\mathbf{x},\mathbf{y})\mathbf{h}_{TM}(\mathbf{x},\mathbf{y}),$$

the decision rule in Eq. 1 is approximated, by only adapting $h_{TM}(\cdot|\cdot)$, as

$$\hat{\mathbf{y}}_t = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m \neq TM} \lambda_m h_m(\mathbf{x},\mathbf{y}) + h_{TM}^t(\mathbf{x},\mathbf{y}). \tag{5}$$

Let $\mathbf{y}$ be the hypothesis proposed by the system, and $\mathbf{y}^*$ the best hypothesis the system is able to produce in terms of translation quality (i.e. the most similar sentence with respect to reference translation proposed by the user $\mathbf{y}^\tau$). Ideally, we would like to adapt the model parameters (be it $\boldsymbol{\lambda}$ or $\mathbf{h}$) so that $\mathbf{y}^*$ is rewarded.

We define the difference (or loss) in translation quality between the proposed hypothesis $\mathbf{y}$ and the best hypothesis $\mathbf{y}^*$ in terms of a given quality measure $\mu(\cdot)$:

$$l(\mathbf{y}) = |\mu(\mathbf{y}^\tau,\mathbf{y}) - \mu(\mathbf{y}^\tau,\mathbf{y}^*)|, \tag{6}$$

where the absolute value has been introduced in order to preserve generality, since in SMT some of the quality measures used, such as TER [15], represent an error rate (i.e. the lower the better), whereas others such as BLEU [12] measure precision (i.e. the higher the better). The difference in probability between $\mathbf{y}$ and $\mathbf{y}^*$ is proportional to

$$\phi(\mathbf{y}) = s(\mathbf{x},\mathbf{y}^*) - s(\mathbf{x},\mathbf{y}). \tag{7}$$

Ideally, we would like that increases or decreases in $l(\cdot)$ correspond to increases or decreases in $\phi(\cdot)$, respectively: if a candidate hypothesis $\mathbf{y}$ has a translation quality $\mu(\mathbf{y})$ which is very similar to the translation quality provided by $\mu(\mathbf{y}^*)$, we would like that such fact is reflected in the translation score s, i.e. $s(\mathbf{x},\mathbf{y})$ is very similar to $s(\mathbf{x},\mathbf{y}^*)$. The purpose of our online procedure should be to promote such correspondence after processing sample t. The update step for $\mathbf{u}_t(\mathbf{x},\mathbf{y})$ can be defined as $\mathbf{u}_{t+1}(\mathbf{x},\mathbf{y}) = \mathbf{u}_t(\mathbf{x},\mathbf{y}) + \boldsymbol{\nu}_t$, where $\mathbf{u}_t(\mathbf{x},\mathbf{y})$ is the update function learnt after observing the previous $(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ pairs, and $\boldsymbol{\nu}_t$ is the solution to the optimisation problem

$$\min_{\mathbf{u},\xi>0} \left(\frac{1}{2}||\mathbf{u} - \mathbf{u}_t||^2 + C\xi^2 \right) \tag{8}$$

subject to constraint $\mathbf{u}_t(\mathbf{x},\mathbf{y})\Phi_t(\mathbf{y}) \geq \sqrt{l(\mathbf{y})} - \xi$, with $\Phi_t(\mathbf{y}) = [\phi(\tilde{y}_1),\ldots,\phi(\tilde{y}_K)]' \approx \mathbf{h}_{TM}(\mathbf{x},\mathbf{y}^*) - \mathbf{h}_{TM}(\mathbf{x},\mathbf{y})$, since all the rest of score functions except $\mathbf{h}_{TM}$ remain constant and the only feature functions we intend to adapt are $\mathbf{h}_{TM}$. Then, the solution to Equation 8 according to PA Type-II has the form [13]:

$$\boldsymbol{\nu}_t = \Phi_t(\mathbf{y}) \frac{\sqrt{l_t(\mathbf{y})} - \mathbf{u}_t(\mathbf{x},\mathbf{y})\Phi_t(\mathbf{y})}{||\Phi_t(\mathbf{y})||^2 + \frac{1}{C}} \tag{9}$$

In [13], the update is triggered only when the proposed hypothesis violates the constraint $\mathbf{u}_t(\mathbf{x},\mathbf{y})\Phi_t \geq \sqrt{l_t(\mathbf{y})}$.

4.2 Scaling Factor Adaptation

A coarse-grained technique for tackling with the online learning problem in SMT implies adapting the log-linear weights $\boldsymbol{\lambda}$. After the system has received the sentence $\mathbf{y}_t^\tau$ as

correct reference for an input sentence $\mathbf{x}_t$, the idea is to compute the best weight vector $\hat{\boldsymbol{\lambda}}_t$ corresponding to the sentence pair observed at time t. Once $\hat{\boldsymbol{\lambda}}_t$ has been computed, $\boldsymbol{\lambda}_t$ can be updated towards a new weight vector $\boldsymbol{\lambda}_{t+1}$, for a certain learning rate α, as:

$$\boldsymbol{\lambda}_{t+1} = (1-\alpha)\boldsymbol{\lambda}_t + \alpha\hat{\boldsymbol{\lambda}}_t \tag{10}$$

As done with $\boldsymbol{\nu}_t$ in Section 4.1, the update term for computing $\boldsymbol{\lambda}_{t+1}$ is given by

$$\hat{\boldsymbol{\lambda}}_t = \Phi_t(\mathbf{y})\frac{\sqrt{l_t(\mathbf{y})} - \boldsymbol{\lambda}_t\Phi_t(\mathbf{y})}{||\Phi_t(\mathbf{y})||^2 + \frac{1}{C}}, \tag{11}$$

where $\Phi_t(\mathbf{y}) = [\phi_1(\mathbf{y}), \ldots, \phi_M(\mathbf{y})]' = \mathbf{h}(\mathbf{x}, \mathbf{y}^*) - \mathbf{h}(\mathbf{x}, \mathbf{y})$, including all feature functions. An update is triggered only when constraint $\boldsymbol{\lambda}_t\Phi_t(\mathbf{y}) \geq \sqrt{l_t(\mathbf{y})}$ is violated.

4.3 Heuristic Variations

Several update conditions different to the ones described above have been explored in this paper. The most obvious is to think that an update has to be performed every time that the quality of a predicted hypothesis $\mathbf{y}$ is lower than the best possible hypothesis $\mathbf{y}_t^*$ in terms of a given quality measure μ. That is, when $\exists \mathbf{y}^* : |\mu(\mathbf{y}_t, \mathbf{y}^*) - \mu(\mathbf{y}_t, \mathbf{y})| > 0$.

In feature vector adaptation, the key idea is to reward those phrases that appear in $\mathbf{y}^*$ but did not appear in $\mathbf{y}$, and, symmetrically, to penalise phrases that appeared in $\mathbf{y}$ but not in $\mathbf{y}^*$. When adapting $\boldsymbol{\lambda}$, the idea is to adjust the discriminative power of models by means of shifting the value of their scaling factors towards the desired value.

5 Experiments

5.1 Experimental Setup

Given that a true CAT scenario is very expensive for experimentation purposes, since it requires a human translator to correct every hypothesis, we will be simulating such scenario by using the reference present in the test set. However, such reference will be fed one at a time, given that this would be the case in an online CAT process.

Translation quality will be assessed by means of BLEU and TER scores. It must be noted that BLEU measures precision, i.e. the higher the better, whereas TER is an error rate, i.e. the lower the better. As mentioned in Section 2, BLEU may often be zero for all hypotheses, which means that $\mathbf{y}^*$ is not always well defined and it may not be possible to compute it. Such samples will not be considered within the online procedure.

As baseline system, we trained a SMT system on the Europarl [6] training data, in the partition established in the Workshop on SMT of the NAACL 2009[1]. Since our purpose is to analyse the performance of the PA algorithm in an online adaptation scenario, we also considered the use of the News Commentary (NC) test set of the 2009 ACL shared task on SMT. Statistics are provided in Table 1. The open-source MT toolkit Moses [7] was used in its default setup, and the 14 weights of the log-linear combination were estimated using MERT [9] on the Europarl development set. Additionally, an interpolated 5-gram language model and Kneser-Ney smoothing [5] was estimated.

[1] `http://www.statmt.org/wmt09/`

Table 1. Characteristics of the Europarl corpus and NC09 test set. OoV stands for "Out of Vocabulary" words, K for thousands of elements and M for millions of elements.

		Es	En	Fr	En	De	En
Training	Sentences	1.3M		1.2M		1.3M	
	Run. words	27.5M	26.6M	28.2M	25.6M	24.9M	26.2M
	Vocabulary	125.8K	82.6K	101.3K	81.0K	264.9K	82.4K
Development	Sentences	2000		2000		2000	
	Run. words	60.6K	58.7K	67.3K	48.7K	55.1K	58.7K
	OoV. words	164	99	99	104	348	103
NC 09 test	Sentences	2525		2051		2051	
	Run. words	68.1K	65.6K	72.7K	65.6K	62.7K	65.6K
	OoV. words	1358	1229	1449	1247	2410	1247

Experiments were performed on the English–Spanish, English–German and English–French language pairs, in both directions and for NC test sets of 2008 and 2009. However, in this paper only the results for English $\rightarrow$ French are presented, for space reasons. In addition, we only report results for the 2009 test set. Nevertheless, the results presented here were found to be coherent in all experiments conducted.

As for the different parameters adjustable in the algorithms described in Section 4.1 and 4.2, they were set according to preliminary investigation to $C \rightarrow \infty$ ($\frac{1}{C} = 0$ was used) in both approaches and $\alpha = 0.01$ in scaling factor adaptation. Instead of using the true best hypothesis, the best hypothesis within a given $nbest(\mathbf{x})$ list was selected.

5.2 Experimental Results

We analysed the performance of the different PA variations described in Section 4, both in terms of BLEU and in terms of TER, and both for adapting $\mathbf{h}$ and $\boldsymbol{\lambda}$. Results for varying order of $nbest$ can be seen in Fig. 1. Although the final scores are reported for the whole test set, all experiments described here were performed following an online CAT approach: each reference sentence was used for adapting the system parameters after such sentence has been translated and its translation quality has been assessed.

It can be seen that the heuristic PA variation yields a small improvement when optimising TER. However, such improvement is not mirrored when optimising BLEU, and hence we assume it is not significant. It can also be seen that adapting $\boldsymbol{\lambda}$ leads to consistently better performance than adapting $\mathbf{h}$. Although adapting $\mathbf{h}$ provides much more flexibility, we understand that adapting $\mathbf{h}$ is a very sparse problem.

The techniques analysed perform much better in terms of TER than in terms of BLEU. Again, it is worth remembering that BLEU is not well defined at the sentence level, and hence the fact that PA has more trouble using it was expected.

In Fig. 2, the evolution of TER throughout the whole test set is plotted for the adaptation of $\mathbf{h}$ and $\boldsymbol{\lambda}$ when setting the size of the $nbest$ list to 1000. In this figure, average TER scores up to the t-th sentence is considered. The reason for plotting average TER is that plotting individual sentence TER scores would result in a very chaotic, unreadable plot, as it can still be seen in the first 100 sentences. Again, in this Figure it also emerges that adapting $\boldsymbol{\lambda}$ leads to much larger improvements than adapting $\mathbf{h}$.

Although it appears that the learning curves peak at about 1500 sentences, this finding is not coherent throughout all experiments carried out, since such peak ranges from

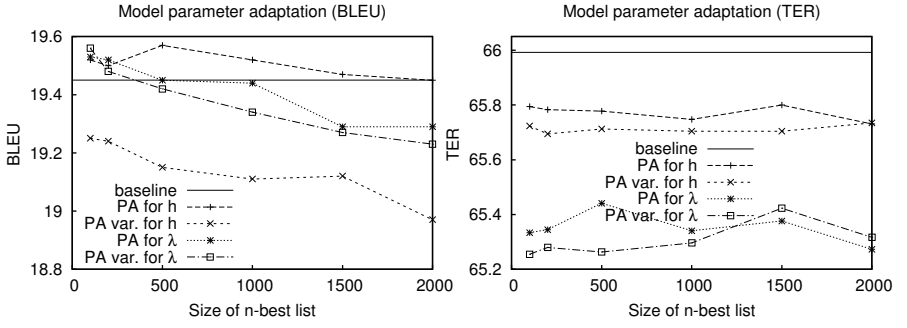


Fig. 1. Final BLEU and TER scores for the NC 2009 test set, English → French when adapting feature functions $\mathbf{h}$ and when adapting scaling factors $\boldsymbol{\lambda}$. `PA` stands for PA as described in Section 4 and `PA var.` for the heuristic variation described in Section 4.3.

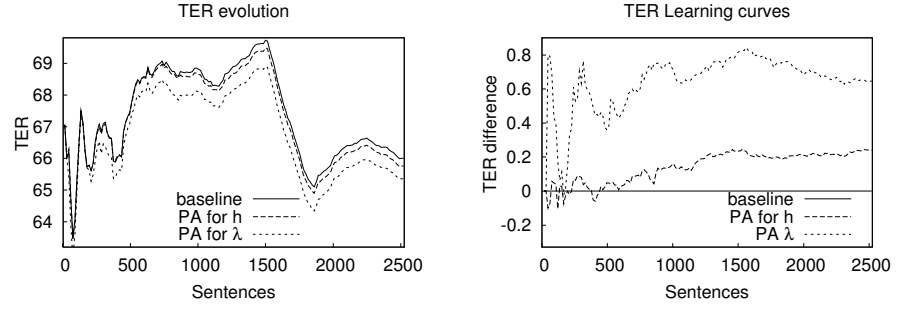


Fig. 2. TER evolution and learning curves when adapting feature functions $\mathbf{h}$ and scaling factors $\boldsymbol{\lambda}$, considering all 2525 sentences within the NC 2009 test set. So that the plots are clearly distinguishable, only 1 every 15 points has been drawn.

300 to 2000 in other cases. This means that the particular shape of the learning curves depends strongly on the chosen test set, and that the information that can be extracted is only whether or not the algorithms implemented provide improvements.

One last consideration involves computation time. When adapting $\boldsymbol{\lambda}$, implemented procedures take about 100 seconds to rerank the complete test set, whereas in the case of adapting $\mathbf{h}$ the time is about 25 minutes. We consider this fact important since in a CAT scenario the user is waiting actively for the system to produce a hypothesis.

6 Conclusions and Future Work

The passive-aggressive algorithm has been analysed for its application in an online scenario, adapting system parameters after each observation. Feedback information has been included into an SMT system, increasing the perception of its own performance.

The passive-aggressive algorithm and a proposed heuristic variation have been applied to two tasks with different characteristics. Feature function adaptation is a sparse problem in the order of thousands of parameters that need to be adapted, whereas the scaling factor adaptation only has around 14 parameters to adapt. This might be one of the reasons for the passive-aggressive algorithm to perform better in the latter task.

Two quality scores have also been used during the experiments and the behaviour of the system allows us to extract one more conclusion. When optimising BLEU, the performance of the algorithm is consistently lower than when optimising TER. We believe that the reason for this is that BLEU is not well defined at the sentence level.

In future work, it would be interesting to observe the impact of smoothed quality scores on the performance of the algorithms.

Acknowledgements

This paper is based upon work supported by the EC (FEDER/FSE) and the Spanish MICINN under projects MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) and iTrans2 (TIN2009-14511). Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project, by the Generalitat Valenciana under grant Prometeo/2009/014 and scholarship GV/2010/067 and by the UPV under grant 20091027.

References

1. Barrachina, S., et al.: Statistical approaches to computer-assisted translation. Computational Linguistics 35(1), 3–28 (2009)
2. Callison-Burch, C., Bannard, C., Schroeder, J.: Improving statistical translation through editing. In: Proc. of 9th EAMT Workshop Broadening Horizons of Machine Translation and its Applications, Malta (April 2004)
3. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Proc. of the Workshop on SMT, pp. 136–158. ACL (June 2007)
4. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research 7, 551–585 (2006)
5. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing II, pp. 181–184 (May 1995)
6. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit X, pp. 79–86 (2005)
7. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proc. of the ACL Demo and Poster Sessions, Prague, Czech Republic, pp. 177–180 (2007)
8. Och, F., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proc. of the ACL 2002, pp. 295–302 (2002)
9. Och, F.: Minimum error rate training for statistical machine translation. In: Dignum, F.P.M. (ed.) ACL 2003. LNCS (LNAI), vol. 2922, pp. 160–167. Springer, Heidelberg (2004)
10. Ortiz-Martínez, D., García-Varea, I., Casacuberta, F.: Online learning for interactive statistical machine translation. In: Proceedings of NAACL HLT, Los Angeles (June 2010)
11. Papineni, K., Roukos, S., Ward, T.: Maximum likelihood and discriminative training of direct translation models. In: Proc. of ICASSP 1998, pp. 189–192 (1998)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: A method for automatic evaluation of machine translation. In: Proc. of ACL 2002, pp. 311–318 (2002)
13. Reverberi, G., Szedmak, S., Cesa-Bianchi, N., et al.: Deliverable of package 4: Online learning algorithms for computer-assisted translation (2008)
14. Sanchis-Trilles, G., Casacuberta, F.: Log-linear weight optimisation via bayesian adaptation in statistical machine translation. In: Proc. of COLING 2010, Beijing, China, pp. 1077–1085 (August 2010)
15. Snover, M., et al.: A study of translation edit rate with targeted human annotation. In: Proc. of AMTA 2006, Cambridge, Massachusetts, USA, pp. 223–231 (August 2006)
16. Zens, R., Och, F., Ney, H.: Phrase-based statistical machine translation. In: Jarke, M., Koehler, J., Lakemeyer, G. (eds.) KI 2002. LNCS (LNAI), vol. 2479, pp. 18–32. Springer, Heidelberg (2002)

Feature Set Search Space for FuzzyBoost Learning*

Plinio Moreno, Pedro Ribeiro, and José Santos-Victor

Instituto de Sistemas e Robótica & Instituto Superior Técnico
Portugal
{plinio,pedro,jasv}@isr.ist.utl.pt

Abstract. This paper presents a novel approach to the weak classifier selection based on the GentleBoost framework, based on sharing a set of features at each round. We explore the use of linear dimensionality reduction methods to guide the search for features that share some properties, such as correlations and discriminative properties. We add this feature set as a new parameter of the decision stump, which turns the single branch selection of the classic stump into a fuzzy decision that weights the contribution of both branches. The weights of each branch act as a confidence measure based on the feature set characteristics, which increases the accuracy and robustness to data perturbations. We propose an algorithm that consider the similarities between the weights provided by three linear mapping algorithms: PCA, LDA and MMLMNN [14]. We propose to analyze the row vectors of the linear mapping, grouping vector components with very similar values. Then, the created groups are the inputs of the FuzzyBoost algorithm. This search procedure generalizes the previous temporal FuzzyBoost [10] to any type of features. We present results in features with spatial support (images) and spatio-temporal support (videos), showing the generalization properties of the FuzzyBoost algorithm in other scenarios.

1 Introduction

Boosting algorithms combine efficiency and robustness in a very simple and successful strategy for classification problems. The advantages of this strategy have led several works to improve the performance of boosting on different problems by proposing modifications to the key elements of the original AdaBoost algorithm [5]: (i) the procedure to compute the data weights, (ii) the selection of the base classifier and (iii) the loss function it optimizes.

The focus of this work is the careful selection of the weak (base) classifier. Since the weak classifier could be any function that performs better than chance, the choice of the weak classifiers is usually motivated by the particular context of the problem. When the objective is to find meaningful sets of data samples,

* This work was supported by FCT (ISR/IST plurianual funding through the PIDDAC Program) and partially funded by EU Project First-MM (FP7-ICT-248258) and EU Project HANDLE (FP7-ICT-231640).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 248–255, 2011.

several dimensions of the original samples are gathered to build augmented base classifiers. This approach has been followed by the SpatialBoost [2], the TemporalBoost [11] and the temporally consistent learners used in [10]. A general drawback of these works is the specificity of the application for which they are constructed. The aim of this work is to generalize the fuzzy decision function of [10] to any type of data, while maintaining its main advantages. The generalized fuzzy decision function selects jointly the usual parameters of a decision stump and the set of features to use, a procedure that turns the single branch selection of the decision stump into a linear combination of the branches. Such a combination of the stump branches is commonly referred to as a fuzzy decision on [7,6,12]. Moreover, [8] shows empirically that the fuzzy tree decreases the variance and consequently improves the classification output.

The generalization of the fuzzy decision function brings a difficult problem to solve: the selection of the feature set. Exhaustive search is prohibitive, so we propose an algorithm that extracts the feature sets from (linear) dimensionality reduction techniques. These techniques map the original feature space to a more meaningful subspace, by the minimization of a cost function. Thus, the linear mapping contains relevant information about the similarity between feature dimensions on the original space. We analyze the rows of the linear mapping, selecting the components with very similar values and disregarding components with very low values. We consider three dimensionality reduction algorithms: Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA) and Multiple Metric Learning for large Margin Nearest Neighbor (MMLMNN) Classification [14]. We apply the feature set search for fuzzy decision stumps in two data domains: spatial (face and car detection) and spatio-temporal (moving people and robots).

2 The FuzzyBoost Algorithm

Boosting algorithms provide a framework to sequentially fit additive models in order to build a final strong classifier, $H(x_i)$. The final model is learned by minimizing, at each round, the weighted squared error

$$J = \sum_{i=1}^{N} w_i (y_i - h_m(x_i))^2, \tag{1}$$

where $w_i = e^{-y_i h_m(x_i)}$ are the weights and N the number of training samples. At each round, the optimal weak classifier is then added to the strong classifier and the data weights adapted, increasing the weight of the misclassified samples and decreasing correctly classified ones [13].

In the case of GentleBoost it is common to use simple functions such as decision stumps. They have the form $h_m(x_i) = a\delta\left[x_i^f > \theta\right] + b\delta\left[x_i^f \leq \theta\right]$, where f is the feature index and δ is the indicator function (i.e. $\delta[condition]$ is one if *condition* is *true* and zero otherwise). Decision stumps can be viewed as decision trees with only one node, where the indicator function sharply chooses branch a

or b depending on threshold θ and feature value x_i^f. In order to find the stump at each round, one must find the set of parameters $\{a, b, f, \theta\}$ that minimizes J w.r.t. h_m. A closed form for the optimal a and b are obtained and the value of pair $\{f, \theta\}$ is found using an exhaustive search [13].

2.1 Fuzzy Weak Learners Optimization

We propose to include the feature set as an additional parameter of the decision stump, as follows:

$$h_m^*(x_i) = \frac{1}{||F||}\left(a\, F^T\delta\left[x_i > \theta\right] + b\, F^T\delta\left[x_i \leq \theta\right]\right), \tag{2}$$

where $x_i \in \mathbb{R}^d$ and the vector $F \in \mathbb{Z}_2^m, m \in \{1, \ldots, d\}$, so the non-zero components of F define a feature set. The vector F chooses a group of original sample dimensions that follow the indicator function constraints of Eq. (2) in order to compute the decision stump $h_m^*(x_i)$. Note that by choosing $m = 1$, Eq. (2) becomes the classic decision stump and the selection of different m values induce different decision stump functions. In this work we choose $F \in \mathbb{Z}_2^d$, thus we do not assume any a priori information about the samples x. Eq. (2) can be rearranged in order to put a and b in evidence,

$$h_m^*(x_i) = a\,\frac{F^T\delta\left[x_i > \theta\right]}{||F||} + b\,\frac{F^T\delta\left[x_i \leq \theta\right]}{||F||}. \tag{3}$$

From Eq. 3 it is easier to see that the the selector F is replacing the indicator function (i.e. a true or false decision) by an average of decisions. The new functions are:

$$\Delta_+(x_i, \theta, F) = \frac{F^T\delta\left[x_i > \theta\right]}{||F||}, \quad \Delta_-(x_i, \theta, F) = \frac{F^T\delta\left[x_i \leq \theta\right]}{||F||}, \tag{4}$$

and they compute the percentage of features selected by F that are above and below the threshold θ. The functions Δ_+ and $\Delta_- = 1 - \Delta_+$ of Eq. 4 sample the interval [0 1] according to the number of features selected by F (i.e. according to $||F||$). For example, if $||F|| = 2$ this yields to $\Delta \in \{0,\ 1/2,\ 1\}$, if $||F|| = 3$ to $\Delta \in \{0,\ 1/3,\ 2/3,\ 1\}$ and so on. The new weak learners, the fuzzy decision stumps, are expressed as $h_m^*(x_i) = a\Delta_+ + b\Delta_-$.

We illustrate in Fig. 1 the difference between the classic decision stumps and our proposed fuzzy stumps. The response of the decision stump is either a or b according to the feature point x_i^f, while the fuzzy stump response is a linear function of Δ_+ that weights the contribution of the decisions a and b, thus the name fuzzy stump.

Replacing the fuzzy stumps of Eq. 3 in the cost function (Eq. 1), the optimal decision parameters a and b are obtained by minimization,

$$a = \frac{\bar{\nu}_+\bar{\omega}_- - \bar{\nu}_-\bar{\omega}_\pm}{\bar{\omega}_+\bar{\omega}_- - (\bar{\omega}_\pm)^2}, \qquad b = \frac{\bar{\nu}_-\bar{\omega}_+ - \bar{\nu}_+\bar{\omega}_\pm}{\bar{\omega}_+\bar{\omega}_- - (\bar{\omega}_\pm)^2}, \tag{5}$$

(a) (b)

Fig. 1. Response of the weak learners: (a) decision stumps and (b) fuzzy stumps

with $\bar{\nu}_+ = \sum_i^N w_i y_i \Delta_+^T$, $\bar{\nu}_- = \sum_i^N w_i y_i \Delta_-^T$,
$\bar{\omega}_+ = \sum_i^N w_i \Delta_+^T$, $\bar{\omega}_- = \sum_i^N w_i \Delta_-^T$, $\bar{\omega}_\pm = \sum_i^N w_i \Delta_-^T \Delta_+^T$.

There is no closed form to compute the optimal θ and F, thus exhaustive search is usually performed. Although finding the optimal θ is a tractable problem, the search for the best F is NP-hard thus generally impossible to perform. This problem can be viewed as a feature selection problem with the objective of choosing, at each boosting round, the set of features that minimizes the error. In order to guide the search and reduce the number of possible combinations we apply dimensionality reduction algorithms in the original feature space, using the projection matrix in order to find feature set candidates.

2.2 The Search Space for the Feature Set

Linear dimensionality reduction techniques aim to find a subspace where regression and classification tasks perform better than in the original feature space. These feature extraction methods aim to find more meaningful and/or discriminative characteristics of the data samples by minimizing task-defined cost functions. Finally, the original features are substituted by the transformed ones in order to perform classification.

Although the dimensionality reduction methods differ in the way that they use labeled or unlabeled data to compute the linear transformation of the input, the projection vectors of all methods code the way that the original feature space should be combined to create a new dimension. Since the linear mapping contains relevant information about the correlations between dimensions of the original feature space, we propose to analyze each projection vector of the mapping by selecting vector components with similar values. Our rationale follows the weight similarity approach: if the weight of a dimension in the projection vector is similar to other dimension(s), this implies some correlation level between those dimensions. We apply this idea to three projection algorithms: PCA, LDA and MMLMNN.

The idea behind PCA is to compute a linear transformation $x^* = Lx$ that projects the training inputs into a variance-maximizing subspace. The linear transformation L is the projection matrix that maximizes the variance of the projected inputs, and the rows of L are the leading eigenvectors of the input data covariance matrix.

LDA uses the labels information to maximize the amount of between-class variance relative to the amount of within-class variance. The linear transformation $x^* = Lx$ outputted by LDA is also a projection matrix, and the rows of L are the leading eigenvectors of the matrix computed as the ratio of the between-class variance matrix and the within-class variance matrix. Unlike PCA, LDA operates in a supervised setting, restricting the number of linear projections extracted to the number of classes present in the problem.

The recently proposed MMLMNN method [14] attempts to learn a linear transformation $x^* = Lx$ of the input space, such that each training input should share the same labels as its k nearest neighbors (named target neighbors) and the training inputs with different label (named impostors) should be widely separated. This two terms are combined into a single loss function that has the competing effect of attracting target neighbors on one hand, and repel impostors on the other (see [14] for details).

Computing F from L. Given a linear mapping L computed by PCA, LDA or MMLMNN, we scale the values of the projection matrix as follows: $\mathcal{L}_{ij} = \frac{|L_{ij}|}{\max(L)}$. The scaling ensures that $0 < \mathcal{L}_{ij} \leq 1$, which allow us to define lower thresholds (s_0 in Alg. 1) and number of intervals (n_s in Alg. 1) that have the same meaning for all the linear mappings. The algorithm for generating the feature sets is as follows:

input : s_0 lower threshold, n_s number of intervals, $\mathcal{L}$ projection matrix
output: $F_j \quad j = 1 \ldots n_s$
1 **for** *each projection (row) vector* $\mathcal{L}_i$ **do**
2 compute $\Delta_s = (\max(\mathcal{L}_i) - s_0)/n_s$;
3 **for** $j = 1 \ldots n_s$ **do**
4 compute $s_j = s_0 + (j-1)\Delta_s$;
5 $F_j = \delta[s_j \leq \mathcal{L}_i < s_j + j\Delta_s]$;
6 **end**
7 **end**

Algorithm 1. Generation of feature sets F of Eq. (2) from a scaled linear mapping $\mathcal{L}$

The lower threshold $s_0 \in [0, 1[$ removes components of $\mathcal{L}_i$ having low projection weights, which are the less meaningful dimensions. The number of intervals $n_s \in \mathbb{N}$ defines the criterion to group dimensions with similar weights (line 5 of Alg. 1), so a high number of intervals will group a few dimensions and a low number of intervals will generate a larger feature set F_j. In order to see the effect of several choices of s_0 and n_s, we apply the Algorithm 1 using several pairs (s_0, n_s) for each linear mapping L.

3 Experimental Results

We evaluate the recognition rate difference between the decision stumps and the fuzzy stumps on two binary problems: (i) face vs. background and (ii) people vs.

(a) CBCL MIT face dataset (b) People vs. robot

Fig. 2. Positive and negative examples of the datasets used in this paper

robot discrimination. We use the CBCL MIT face database [1] and the people vs. robot database introduced in [4]. Figure 2 shows some examples of each class for both datasets.

Parameter selection of the feature search. We define a set of pairs (s_0, n_s) for each linear mapping L in order to see the effect of the parameter selection in the performance of the generated feature sets in the FuzzyBoost algorithm. We set three low thresholds $s_0 \in \{0.1, 0.2, 0.3\}$, and for each s_0 we set three number of intervals, as follows: (i) $(0.1, 9)$, $(0.1, 18)$ and $(0.1, 27)$ for the first s_0, (ii) $(0.2, 8)$, $(0.2, 16)$ and $(0.2, 24)$ for the second s_0 and (iii) $(0.3, 7)$, $(0.3, 14)$ and $(0.3, 21)$ for the third s_0. The rationale behind this choice is to have Δ_s intervals with the same length across the different s_0 values, which allows to evaluate the pairs (s_0, n_s) fairly. For each pair (s_0, n_s), we apply the Alg. 1 on the three projection methods in order to generate the feature sets F. Then, F is applied on the weak learner selection of Eq. (2) in a fixed number of rounds $M = 1000$. The quantitative evaluation is the maximum recognition rate attained on the testing set.

Faces database. The feature vector in this problem is constructed with the raw images (19x19 pixels). This is a high dimensional space, where the linear mappings are not able to find feature sets that provide a large improve when compared to GentleBoost. Nevertheless, Figure 3 shows that MMLMNN performs better than GentleBoost for most of the tests and a lot better than its competitors PCA and LDA.

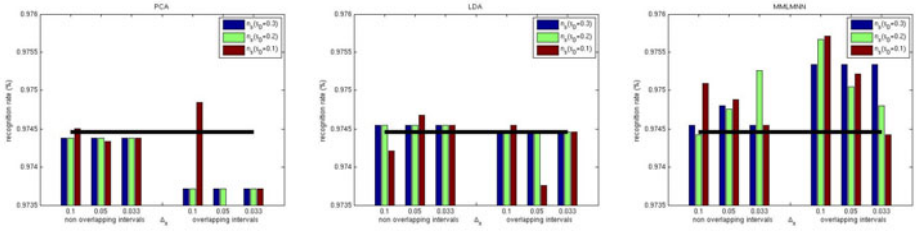


Fig. 3. Recognition rate of the FuzzyBoost algorithm for feature sets generated from: PCA, LDA and MMLMNN

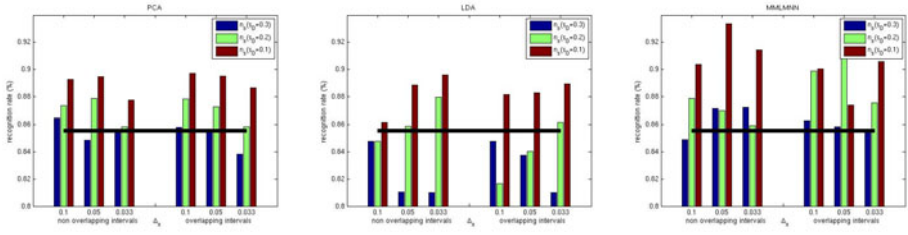


Fig. 4. Recognition rate of the FuzzyBoost algorithm for feature sets generated from: PCA, LDA and MMLMNN, FOA weighted histogram over 5 frames

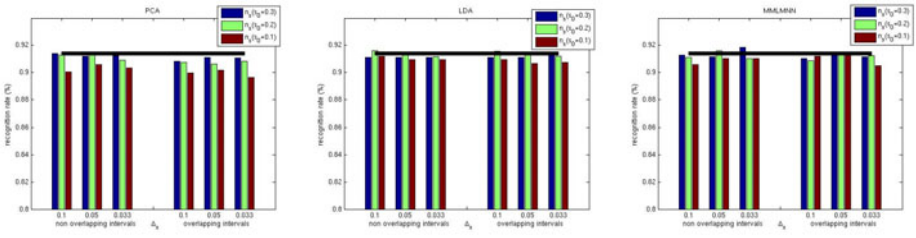


Fig. 5. Recognition rate of the FuzzyBoost algorithm for feature sets generated from: PCA, LDA and MMLMNN, MBH feature over 5 frames

Robot versus people database. We apply two types of features: The weighted histogram of the Focus Of Attention (FOA) feature [9] and the Motion Boundary Histogram (MBH) [3] using a polar sampling grid cell. The spatio-temporal volume is constructed by stacking the feature vector of the current frame with the vectors of the previous four frames. The FOA feature is a $64d$ vector per frame, and the MBH is a $128d$ feature vector per frame. Thus, the spatio-temporal feature based on MBH lies in a very high dimensional space. Figs. 4 and 5 show the results of the FOA and MBH features respectively. We notice the same trend of the previous tests, where MMLMNN performs better than the other dimensional reduction algorithms. Remark also the large gap improvement when using the FOA feature against the GentleBoost in Fig. 4. On the other hand, little improvement is achieved with the MBH feature. We believe that is more difficult to find the right spatio-temporal groupings in this feature space and it looks like the simple stacking of the feature vectors is able to attain good classification results. On the other hand, the FuzzyBoost is able to improve the performance of the FOA feature even higher than the GentleBoost with the stacked MBH features of five frames.

4 Conclusions

We introduce the generation of appropriate feature sets for the FuzzyBoost algorithm. The algorithm for feature set generation analyzes the row vectors of any linear dimensionality reduction algorithm in order to find feature dimensions

with similar vector components. We generalize the formulation of the fuzzy decision stump, which now can be applied to any learning problem. We present two types of domains where the fuzzy decision stump brings robustness and generalization capabilites, namely: face recognition and people vs. robot detection by their motion patterns.

References

1. CBCL face database #1, MIT center for biological and computation learning, `http://cbcl.mit.edu/projects/cbcl/software-datasets/FaceData2.html`
2. Avidan, S.: SpatialBoost: Adding spatial reasoning to adaBoost. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 386–396. Springer, Heidelberg (2006)
3. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
4. Figueira, D., Moreno, P., Bernardino, A., Gaspar, J., Santos-Victor, J.: Optical flow based detection in mixed human robot environments. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnação, M.L., Silva, C.T., Coming, D. (eds.) ISVC 2009. LNCS, vol. 5875, pp. 223–232. Springer, Heidelberg (2009)
5. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
6. Jang, J.S.R.: Structure determination in fuzzy modeling: a fuzzy cart approach. In: Proceedings IEEE Conference on Fuzzy Systems, vol. 1, pp. 480–485 (June 1994)
7. Janikow, C.Z.: Fuzzy decision trees: issues and methods. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 28(1), 1–14 (1998)
8. Olaru, C., Wehenkel, L.: A complete fuzzy decision tree technique. Fuzzy Sets and Systems 138(2), 221–254 (2003)
9. Pla, F., Ribeiro, P.C., Santos-Victor, J., Bernardino, A.: Extracting motion features for visual human activity representation. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3522, pp. 537–544. Springer, Heidelberg (2005)
10. Ribeiro, P.C., Moreno, P., Santos-Victor, J.: Boosting with temporal consistent learners: An application to human activity recognition. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Paragios, N., Tanveer, S.-M., Ju, T., Liu, Z., Coquillart, S., Cruz-Neira, C., Müller, T., Malzbender, T. (eds.) ISVC 2007, Part I. LNCS, vol. 4841, pp. 464–475. Springer, Heidelberg (2007)
11. Smith, P., da Vitoria Lobo, N., Shah, M.: Temporalboost for event recognition. In: International Conference on Computer Vision, vol. 1, pp. 733–740 (October 2005)
12. Suarez, A., Lutsko, J.F.: Globally optimal fuzzy decision trees for classification and regression. IEEE Transactions on PAMI 21(12), 1297–1311 (1999)
13. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE PAMI 29(5), 854–869 (2007)
14. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. 10, 207–244 (2009)

Interactive Structured Output Prediction: Application to Chromosome Classification*

Jose Oncina[1] and Enrique Vidal[2]

[1] Dept. Lenguajes y Sistemas Informáticos
Universidad de Alicante
oncina@dlsi.ua.es
[2] Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
evidal@iti.upv.es

Abstract. Interactive Pattern Recognition concepts and techniques are applied to problems with structured output; i.e., problems in which the result is not just a simple class label, but a suitable structure of labels. For illustration purposes (a simplification of) the problem of Human Karyotyping is considered. Results show that a) taking into account label dependencies in a karyogram significantly reduces the classical (non-interactive) chromosome label prediction error rate and b) they are further improved when interactive processing is adopted.

Keywords: interactive pattern recognition, machine learning, structured output prediction, chromosome classification, karyotype recognition.

1 Introduction

Classification is one of the most traditional Pattern Recognition (PR) frameworks [2]. For a given input x, the set of possible output hypotheses is a finite (and typically small) set of class-labels, or just integers $\{1, \ldots, C\}$, where C is the number of classes. In this case, the *search* needed to solve the recognition problem amounts to a straightforward exhaustive exploration of the corresponding C posterior probability values, $\Pr(h \mid x)$; that is,

$$\hat{h} = \arg\max_{1 \le h \le C} \Pr(h \mid x) \tag{1}$$

While classification is in fact a useful framework within which many applications can be naturally placed, there are many other practical problems of increasing interest which need a less restrictive framework where hypotheses are not just labels, but some kind of *structured* information. This is the case, for

* Work supported by the MIPRCV Spanish MICINN "Consolider Ingenio 2010" program (CSD2007-00018). Work supported by the Spanish CICyT through project TIN2009-14205-C04-01. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 256–264, 2011.

example, of Automatic Speech or Handwritten Text Recognition (ASR, HTR), Machine Translation (MT), etc. In these cases, the inputs, x, are structured as *sequences* of feature vectors (ASR, HTR) or words (MT) and the outputs, h, are *sequences* of words or other adequate linguistic units. Many applications admit this kind of input and output sequential structuring, but there are also other practical problems, many of them in the field of Image Processing and Computer Vision, which require more complex structures such as input and output *arrays* or *graphs* of vectors and labels, respectively.

Let $\mathcal{H}$ be the *structured hypotheses space*. Now (1) is written as:

$$\hat{h} = \arg\max_{h \in \mathcal{H}} \Pr(h \mid x) \tag{2}$$

Depending on the exact nature of $\mathcal{H}$, this optimization can become quite complex, but several adequate algorithmic solutions or approximations, such as *Viterbi search* [12,3], $A^\star$ [1,7], etc., have been developed over the last few decades.

In this paper we are interested in applying Interactive PR (IPR) [11] approaches to problems with structured output because it is in this kind of problems where the IPR framework is likely to be most fruitful.

To illustrate concepts, problems and approaches in this framework, we will consider here a simplification of a classical PR problem: the recognition of human karyotypes. While individual chromosome recognition [10,5] is a typical PR example of *classification*, the recognition of a whole karyotype [8,6] properly corresponds to the case of *structured input/output*, as will be discussed below.

A *karyotype* is the number and appearance of chromosomes in the nucleus of a eukaryote cell. Normal human karyotypes contain 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Normal karyotypes for females contain two X chromosomes, males have both an X and a Y chromosomes. Any variation from the standard karyotype may lead to developmental abnormalities. The chromosomes are depicted (by rearranging a microphotograph) in a standard format known as a *karyogram* or idiogram: in pairs, ordered by size and position of centromere for chromosomes of the same size[1]. Each chromosome is assigned a label from $\{\text{“1”}, \ldots, \text{“22”}, \text{“}X\text{”}, \text{“}Y\text{”}\}$, according with its position in the karyogram [8].

In this work we consider the problem of karyotype recognition and we explore IPR approaches to increase the productivity with respect to a traditional, *non interactive* or *“offline”* PR approach.

In order to focus on the most relevant aspects of the problem we will not consider the real, full karyotype recognition problem, but a simpler setting in which only single chromosome images, rather than pairs, are considered and sex chromosomes, “X”, “Y”, are ignored. Then a *karyotype* is represented by a sequence

[1] For the sake of simplicity, we ignore here the initial image segmentation task and assume that each of the 46 chromosomes in a normal unsorted karyotype is already represented as an individual image. Moreover, we do not take into account recent advances in karyotype analysis, such as fluorescent dye based spectral karyotyping [9], which allow obtaining coloured chromosome images and may significantly simplify the real human karyotyping problem.

or vector of chromosome images $\mathbf{x} = (x_i)_{i=1..22}$. Our task is to obtain the corresponding *karyogram*; i.e., a corresponding sequence or vector $\mathbf{h} = (h_i)_{i=1..22}$, where each h_i is the label or class of the chromosome image x_i, $i \in \{1, \ldots, 22\}$. For example, $h_4 = 7$ means that the chromosome image x_4 belongs to class 7 or has the label "7" in the karyogram.

From now on, we assume that a reliable PR system is available for classifying individual chromosome images. For each image x_i, $i \in \{1, \ldots, 22\}$, the system provides us with adequate approximations, $P(j \mid x_i)$, to the posterior probabilities $\Pr(j \mid x_i)$, $j \in \{\text{"1"}, \ldots, \text{"22"}\}$.

2 Non-interactive and Interactive Frameworks

We explore three different frameworks: One non interactive or *"offline"* and two interactive called *"active"* and *"passive"*. The names *active* and *passive* refer to who takes the supervision "initiative". In the active case, the system "actively" proposes items to supervise, while in the passive case, it just "passively" waits for the user to decide which items need supervision and/or correction.

Offline: The system proposes a vector of labels $\mathbf{h}$. This vector is supervised by a user who corrects all the errors. User's effort is measured in two ways: a) the number of karyograms with at least one misclassified chromosome. In this case we are assuming that the same effort is needed to correct a single chromosome label as to correct several; b) the number of misclassified chromosomes. We assume that the effort is proportional to the number of label corrections needed.

Passive: The system proposes a karyogram hypothesis $\mathbf{h}$. Then the user examines its labels $h_1, h_2, \ldots$ sequentially until the first error is found. After correcting this error, the system proposes a new karyogram consistent with all the previously checked and/or corrected elements. Note that this protocol can be equivalently formulated as follows: The system, sequentially for $i = 1, \ldots, 22$, proposes the candidate label h_i for the chromosome image x_i. At each step, the user corrects the possible label error. In this framework the obvious measure of effort is counting the number of corrections the user has to make. However, we will also report the number of karyograms that need at least one correction.

Active: The system sequentially proposes a pair (i, j) as an hypothesis that the chromosome x_i is of the class $h_i = j$ in the karyogram. Like in the previous case, the effort is measured as the number of times the user should correct the possible system hypothesis error.

3 Offline Framework

In this case, classical, non-interactive processing is assumed. Different scenarios are considered, depending on which errors we want to minimize.

3.1 Offline Individual Chromosomes

This is perhaps the simplest setting in which individual chormosome images have to be classified without taking into account that they may belong to a

karyotype. This is the setting we find in the majority of PR papers dealing with chromosome recognition (e.g., [10,5]).

In traditional PR [2], decision theory is used to minimize the cost of wrong hypotheses. A 0/1 cost or *loss* function corresponds to minimizing the number of wrong hypotheses. Under this minimal error loss, the best hypothesis is shown to be one which maximises the hypothesis posterior probability.

In this case, the individual chromosome error is minimised by maximizing the posterior probability for each chromosome image; that is, for all $\mathbf{x}$ and for each $i \in \{1, \ldots, 22\}$:

$$\hat{h}_i = \underset{j \in \{1,\ldots,22\}}{\arg\max}\ P(j \mid x_i) \tag{3}$$

3.2 Offline Karyotype Global

Here we aim to minimize complete-karyogram errors. According to decision theory, for each $\mathbf{x}$ we have to search for the most probable karyogram, $\hat{\mathbf{h}}$:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}}{\arg\max} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{4}$$

Assuming independence beyond the impossibility of assigning two different labels to the same chromosome image, we can write:

$$\Pr(\mathbf{h} \mid \mathbf{x}) = \begin{cases} C \prod_{i=1..22} \Pr(h_i \mid x_i) & \text{if } \mathbf{h} \in \mathcal{H}' \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where C is a normalization constant and $\mathcal{H}' = \{\mathbf{h} \in \mathcal{H} : h_i \neq h_j\ \forall i \neq j\}$ is the set of valid hypothesis (those without repeated labels). This way (4) becomes:

$$\hat{\mathbf{h}} = \underset{\mathbf{h} \in \mathcal{H}'}{\arg\max} \prod_{i=1..22} P(h_i \mid x_i) \tag{6}$$

To approximately solve this difficult maximization problem, a greedy strategy is adopted. First we compute $(\hat{i}, \hat{j}) = \arg\max_{i,j} P(j \mid x_i)$ and we assign $h_{\hat{i}} = \hat{j}$. Then, we eliminate the chromosome $x_{\hat{i}}$ and label $\hat{j}$ from the arg max searching set and repeat the process until all elements of $\mathbf{h}$ have been assigned.

3.3 Offline Karyotype Unconstrained

This setting is similar to the previous one in that each batch of 22 chromosome images, $\mathbf{x}$, is considered to be a complete karyotype. But here we aim to minimize the number of chromosome (rather than complete-karyogram) errors.

Let $\mathbf{h}$ be a proposed hypothesis and $\mathbf{h}^\star$ the "correct" hypothesis. The *loss* function in this case is not 0/1, but the total number of misclassified chromosomes in $\mathbf{h}$. This loss is given by the Hamming distance:

$$d(\mathbf{h}, \mathbf{h}^\star) = \sum_{1=1..22} [h_i \neq {h^\star}_i] \tag{7}$$

where $[\mathcal{P}]$ denotes the Iverson bracket, which is 1 if $\mathcal{P}$ is *true* and 0 otherwise. Then, the conditional risk [2] (i.e., the expected number of errors when a hypothesis $\mathbf{h}$ is proposed for a given $\mathbf{x}$) is:

$$R(\mathbf{h} \mid \mathbf{x}) = \sum_{\mathbf{h}' \in \mathcal{H}} d(\mathbf{h}, \mathbf{h}') \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{8}$$

and the hypothesis that minimises this risk is:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}\in\mathcal{H}} \sum_{\mathbf{h}'\in\mathcal{H}} d(\mathbf{h},\mathbf{h}') \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{9}$$

$$= \arg\min_{\mathbf{h}\in\mathcal{H}} \sum_{i=1..22} \sum_{\mathbf{h}'\in\mathcal{H}} [h_i \neq h'_i] \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{10}$$

$$= \arg\max_{\mathbf{h}\in\mathcal{H}} \sum_{i=1..22} \sum_{\mathbf{h}'\in\mathcal{H}} [h_i = h'_i] \Pr(\mathbf{h}' \mid \mathbf{x}) \tag{11}$$

And now, since the i-summation terms are independent, the maximisation can be split into 22 maximization problems, one for each h_i.

$$\hat{h}_i = \arg\max_{j\in\{1,\ldots,22\}} \sum_{\mathbf{h}\in\mathcal{H},\, h_i=j} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{12}$$

$$= \arg\max_{j\in\{1,\ldots,22\}} \sum_{\mathbf{h}\in\mathcal{H}',\, h_i=j} \prod_{k=1..22} P(h_k \mid x_k) \tag{13}$$

$$= \arg\max_{j\in\{1,\ldots,22\}} P(j \mid x_i) \sum_{\mathbf{h}\in\mathcal{H}',\, h_i=j} \prod_{k=1,\, k\neq i..22} P(h_k \mid x_k) \tag{14}$$

Finally, it is interesting to see that, if we assume in (14) that the individual chromosome probabilities are reasonably well approximated, the summation would not vary enough to dominate the big variations of $P(i \mid x_i)$ and, therefore,

$$\hat{h}_i \approx \arg\max_{j\in\{1,\ldots,22\}} P(j \mid x_i) \tag{15}$$

which is identical to the classical solution to the *Offline Individual Chromosome* setting. Note that, as in that setting, here we are not restricting $\hat{\mathbf{h}}$ to be a valid hypothesis. That is, in the optimal $\hat{\mathbf{h}}$ we may have $\hat{h}_i = \hat{h}_j, i \neq j$.

We can enforce finding only valid hypothesis through a simple heuristic: at each step, select the chromosome label that maximises $P(j \mid x_i)$, provided j was not used in a previous step. It may be argued that introducing this restriction will lead to more accurate predictions. However, with the approximation (15), this heuristic exactly leads to the greedy solution to the *Offline Karyotype Global* problem discussed at the end of section 3.2.

4 Interactive Passive Framework

In this framework two approaches have been considered: *Karyotype* and *Karyotype Unconstrained*. In both cases, it is assumed that the karyogram elements are explored in a *left–to–right* sequential order. In what follows, suppose we are at the i^{th} interaction step and let $\mathbf{h}'$ denote the hypothesis provided by the system in the previous step, $i-1$. Given the left–to–right exploration, all the elements h'^{i-1}_1 of $\mathbf{h}'$ are known to be correct.

4.1 Interactive Passive Left–to–Right Karyotype Global

In this strategy we look for a hypothesis, compatible with the known correct information in ${h'}_1^{i-1}$, which minimises the expected number of whole karyogram errors. That is:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}\in\mathcal{H}} \Pr(\mathbf{h} \mid \mathbf{x}, {h'}_1^{i-1}) \tag{16}$$

$$= \arg\max_{\mathbf{h}\in\mathcal{H},\, h_1^{i-1}={h'}_1^{i-1}} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{17}$$

$$= \arg\max_{\mathbf{h}\in\mathcal{H}',\, h_1^{i-1}={h'}_1^{i-1}} \prod_{k=i..22} P(h_k \mid x_k) \tag{18}$$

As in *Offline Karyotype Global*, a greedy approach is used for this maximisation. First all the labels known from the previous step are assigned; i.e., $h_1^{i-1} = {h'}_1^{i-1}$. Next we obtain $(\hat{k}, \hat{j}) = \arg\max_{i\le k\le 22, j\notin {h'}_1^{i-1}} P(j \mid x_k)$ and assign $\hat{h}_{\hat{k}} = \hat{j}$. Then, the chromosome $x_{\hat{k}}$ and the label $\hat{j}$ are removed from the searching set and the process is repeated until all elements of $\hat{\mathbf{h}}$ have been assigned.

4.2 Interactive Passive Left–to–Right Karyotype Unconstrained

In this case, at each step i^{th} we just look for the most probable label for the i^{th} chromosome image, assuming all the labels assigned in previous steps are correct. Clearly, in this way we do not explicitly care about possible label repetitions for the labels to be assigned in further steps and this is why this strategy is called *"unconstrained"*. However, since the single label to be assigned at each step is restricted to be different from those assigned in previous steps, the final result obtained at the end of the process is guaranteed to be valid karyogram.

Formally, we look for the most probable label h_i for the chromosome image x_i, given that all the labels ${h'}_1^{i-1}$ of $\mathbf{h}'$ are correct. That is:

$$\hat{h}_i = \arg\max_{j\in\{1,\ldots,22\}} \sum_{\mathbf{h}\in\mathcal{H},\, h_i=j} \Pr(\mathbf{h} \mid \mathbf{x}, {h'}_1^{i-1}) \tag{19}$$

$$= \arg\max_{j\in\{1,\ldots,22\}} \sum_{\mathbf{h}\in\mathcal{H},\, h_i=j,\, h_1^{i-1}={h'}_1^{i-1}} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{20}$$

$$= \arg\max_{j\in\{1,\ldots,22\}} P(j \mid x_i) \sum_{\mathbf{h}\in\mathcal{H}',\, h_i=j,\, h_1^{i-1}={h'}_1^{i-1}} \prod_{k\in\{i+1,\ldots,22\}} P(h_k \mid x_k) \tag{21}$$

As in the *Offline Karyotype Unconstrained* case, if we assume that the summation is going to change less than $P(j \mid x_i)$. Then,

$$\hat{h}_i \approx \arg\max_{j\in\{1,\ldots,22\},\, j\notin {h'}_1^{i-1}} P(j \mid x_i) \tag{22}$$

4.3 Interactive Active Framework

In this framework, at the step i^{th}, the system chooses which chromosome and class label has to be supervised. In the previous karyogram, $\mathbf{h}'$, we write ${h'}_k = 0$

if an only if we don't know whether the k^{th} label in $\mathbf{h}'$ is correct. Let $c(\mathbf{h}') = \{j : j = h_k \neq 0,\ 1 \leq k \leq 22\}$ be the set of correct labels in $\mathbf{h}'$. An optimal chromosome-label pair to be supervised is:

$$(\hat{k}, \hat{j}) = \underset{k:h_k=0,\, j \notin c(\mathbf{h}')}{\arg\max} \sum_{\mathbf{h} \in \mathcal{H},\, h_k = j} \Pr(\mathbf{h} \mid \mathbf{x}) \tag{23}$$

$$= \underset{k:h_k=0,\, j \notin c(\mathbf{h}')}{\arg\max} P(j \mid x_k) \sum_{\mathbf{h} \in \mathcal{H}',\, h_k = j} \prod_{l=1..22,\, l \neq k} P(h_l \mid x_l) \tag{24}$$

As in previous cases, if the variation is dominated by $\Pr(j \mid x_k)$:

$$(\hat{k}, \hat{j}) \approx \underset{(k,j),\, h'_l \neq 0,\, j \notin c(\mathbf{h}')}{\arg\max} P(j \mid x_k) \tag{25}$$

5 Experiments

The experiments presented in this work have been carried out using the so-called *"Copenaghen Chromosomes Data Set"*. The raw data, preprocessing and representation are described in detail in [4,10]. Chromosome images are represented as strings of symbols, each of which represents the variation of the grey-level along the chromosome median axis. No centromere position information [10] was used. 200 karyotypes and 4,400 chromosome samples are available in this data set.

These samples were split into two blocks of 100 karyotipes (2,200 chromosome samples), every experiment entailed two runs following a two-blocks Cross-Validation scheme. The reported error-rates are the average of these two runs.

The probabilities needed to apply the methods described in the previous sections where obtained with the so-called ECGI approach [10]. Models used in this approach can be seen as a kind of Hidden Markov Models where the topology is automatically derived from the training strings. For each chromosome class, a model was trained using the training strings of this class. Then, for each test string x, its corresponding 22 class-likelihoods $P(x \mid j)$ were computed by parsing x through the 22 trained models. The posterior probabilities were obtained by normalizing the likelihoods assuming uniform priors for the 22 possible classes.

Using these probabilities under experimintal conditions similar to those adopted here, the error rate reported in [10] for individual chromosome classification was close to 8%. This is the baseline for the present experiments.

The following methods have been tested: *Offline Individual Chromosomes* (OIC, eq. (3)), *Offline Karyotype Global* (OKG, eq. (6)), *Interactive Passive Left-to-Right Karyotype Unconstrained* (PKU, eq. (22)), *Interactive Passive Left-to-Right Karyotype Global* (PKG, eq. (18)) and *Interactive Active* (IAC, eq. (25)).

The *Offline Karyotype Unconstrained* (eq. (15)) and the *Interactive Passive Left-to-Right Karyotype Unconstrained* (eq. (22)) frameworks have not been tested because, as noted in Sections 3.3 and 4.2, approximations and greedy solutions make solutions for these frameworks identical to those of *Offline Individual Chromosomes* or *Offline Karyotype Global* respectively.

It is worth noting that all these methods, except IPU (eq. 22), are insensible to the order in which the chromosomes appear in $\mathbf{x}$. Nevertheless, each experiment

Table 1. Karyotype and chormosome error corrections needed (in %). The first raw (OIC) is the baseline.

Method	Equation	Karyotype	Chromosome
Offline Indivual Chromosome (OIC)	(3)	76	8.0
Offline Karyotype Global (OKG)	(6)	27	3.7
Pasive Karyotype Unconstrained (PKU)	(22)	55	4.6
Pasive Karyotype Global (PKG)	(18)	27	2.1
Active (IAC)	(25)	27	1.9

has been carried out twice, one with the original order of the chromosomes in the data set and another with the reverse of this order. Results are averaged for these two runs.

6 Results

Empirical results are shown in Table 1. As expected the *Global* methods lead to the best karyotype-level results (and also at the chromosome level). On the other hand, interactive processing clearly requires far fewer label correction: about 43% fewer for both PKU relative to OIC and PKG relative to OKG. Finally, the IAC approach achieves the overall best results.

7 Discussion and Conclusions

This work shows how to apply interactive Pattern Recognition concepts and techniques to problems with structured output. For illustration purposes these techniques are applied to (a simplification of) the problem of Human Karyotyping. Results show that a) taking into account label dependencies in a karyogram significantly reduces the classical (noninteractive) chromosome label prediction errors and b) performance is further improved when interactive processing is adopted. These results have been obtained using both search and probability computation approximations. Further improvements are expected by improving the accuracy of these computations.

References

1. Dechter, R., Pearl, J.: Generalized best-first search strategies and the optimality of A*. J. ACM 32, 505–536 (1985), http://doi.acm.org/10.1145/3828.3830
2. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, Chichester (1973)
3. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1998)
4. Kao, J.H., Chuang, J.H., Wang, T.: Chromosome classification based on the band profile similarity along approximate medial axis. Pattern Rec. 41, 77–89 (2008)
5. Martínez, C., García, H., Juan, A.: Chromosome classification using continuous hidden markov models. In: Perales, F.J., Campilho, A.C., Pérez, N., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 494–501. Springer, Heidelberg (2003)

6. Martínez, C., Juan, A., Casacuberta, F.: Iterative contextual recurrent classification of chromosomes. Neural Processing Letters 26(3), 159–175 (2007)
7. Pearl, J.: Heuristics - Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley, Reading (1985)
8. Ritter, G., Gallegos, M., Gaggermeier, K.: Automatic context-sensitive karyotyping of human chromosomes based on elliptically symmetric statistical distributions. Pattern Recognition 28(6), 823–831 (1995)
9. Schröck, E., du Manoir, S., Veldman, T., Schoell, B., Wienberg, J., Ferguson-Smith, M.A., Ning, Y., Ledbetter, D.H., Bar-Am, I., Soenksen, D., Garini, Y., Ried, T.: Multicolor spectral karyotyping of human chromosomes. Science 273(5274), 494–497 (1996)
10. Vidal, E., Castro, M.: Classification of Banded Chromosomes using Error-Correcting Grammatical Inference (ECGI) and Multilayer Perceptron (MLP). In: Sanfeliu, A., Villanueva, J., Vitriá, J. (eds.) In VII National Symposium on Pattern Recognition and Image Analysis, Barcelona, pp. 31–36 (1997)
11. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive pattern recognition. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) MLMI 2007. LNCS, vol. 4892, pp. 60–71. Springer, Heidelberg (2008)
12. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information Theory 13, 260–269 (1967)

On the Use of Diagonal and Class-Dependent Weighted Distances for the Probabilistic k-Nearest Neighbor*

Roberto Paredes[1] and Mark Girolami[2]

[1] Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia
Camino de vera S/N, 46022 Valencia, Spain
rparedes@dsic.upv.es
[2] University of Glasgow, UK
girolami@dcs.gla.ac.uk

Abstract. A probabilistic k-nn (PKnn) method was introduced in [13] under the Bayesian point of view. This work showed that posterior inference over the parameter k can be performed in a relatively straightforward manner using Markov Chain Monte Carlo (MCMC) methods. This method was extended by Everson and Fieldsen [14] to deal with metric learning. In this work we propose two different dissimilarities functions to be used inside this PKnn framework. These dissimilarities functions can be seen as a simplified version of the full-covariance distance functions just proposed. Furthermore we propose to use a class-dependent dissimilarity function as proposed in [8] aim at improving the k-nn classifier. In the present work we pursue a simultaneously learning of the dissimilarity function parameters together with the parameter k of the k-nn classifier. The experiments show that this simultaneous learning lead to an improvement of the classifier with respect to the standard k-nn and state-of-the-art technique as well.

1 Introduction

The k-Nearest-Neighbor (knn) rule is one of the most popular and straightforward pattern recognition techniques. This classification rule is clearly competitive when the available number of prototypes is (very) large, relative to the intrinsic dimensionality of the data involved. The knn has different parameters to be tuned beyond the k parameter. We can consider the distance function and the set of reference-prototypes as two different (*hyper*-)parameters to be tuned as well. Normally the k value uses to be estimated by means of cross-validation but more complicate strategies must be consider to tune the other parameters. For the *distance* function different approaches are proposed in oder to obtain a suitable distance function depending on each particular problem [1–8]. And for the set of prototypes to be used as *reference* set, several efforts have been carried out to optimize this set of prototypes, for instance *editing* and

* Work supported by the Spanish MEC/MICINN under the MIPRCV Consolider Ingenio 2010 program (CSD2007-00018).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 265–272, 2011.

condensing techniques [9–11], and other techniques where a reduced set of prototypes are modified in order to minimize some classification error estimation [12].

The work presented here is based on our previous studies in distance functions for nearest neighbor classification [8]. In [8] a new technique for weighted-distance learning is proposed based on the minimization of an index related to the leaving-one-out error estimation of the training set. The results of this technique over a pool datasets was quite competitive. Despite of this good behaviour this technique could be improved by means of two different contributions. First, the use of a larger neighborhood than the nearest neighbor by means of using k-nn with $k > 1$. Clearly, in most of the classification tasks, the use $k > 1$ could improve drastically the accuracy of the classifier. The problem is that the index to minimize proposed in [8] is an approximation to the error estimation of the *Nearest Neighbor* classifier and such index is not feasible to be defined for the *k-Nearest Neighbor*. Second, in some classification tasks the leaving-one-out error estimation could be an *optimistic* estimator because in some data distributions the samples could pair up. In these situations the parameters obtained could be overfitted. Bayesian inference can help to solve this two problems, to learn both, a weighted distance and a suitable k value, but within the Bayesian framework that ensure a good generalization capabilities of the parameters obtained.

A probabilistic k-nn (PKnn) method was introduced in [13] under the Bayesian point of view. By defining an approximate joint distribution the authors show that posterior inference over k can be performed in a relatively straightforward manner employing standard Markov Chain Monte Carlo (MCMC) methods. On the other hand, Everson and Fieldsen [14] extended this inferential framework to deal with metric learning. However, the experimental evaluation of this approach was very reduced and not completely convincing.

A more extensive empirical evaluation was performed by Manocha and Girolami in [15]. This work compares the conventional K-nn with the PKnn and shows that there is no significant statistical evidence to suggest that either method is a more accurate classifier than the other one. This empirical study was performed considering only the parameter k without considering the metric learning possibility.

In the present work we are going to deal with the metric learning together with the estimation of the parameter k. To this end, and due to our own expertise, we are going to use a simpler version of the metric learning proposed in [14] but at the same time we extend this metric learning adding a class-dependency to the distance function to be used. Moreover in order to overcome the computational effort of the Bayesian approach in the classification phase we propose a maximum a likelihood selection of the classifier parameters. An experimental evaluation is carry out showing the benefits of the here proposed approaches.

The present paper is organized as follows. Section 2 presents the PKnn approach and the modifications proposed. Section 3 draws the different considerations made in order to apply the Monte Carlo Markov Chain (MCMC) in our framework. Section 4 presents the experiments carried out and the results obtained. Finally section 5 draws the final conclusions and future work.

2 Approach

Let be $T = \{(t_1, \mathbf{x}_1), \cdots, (t_N, \mathbf{x}_N)\}$ the training data, where each $t_n \in \{1, \cdots, C\}$ denotes the class label associated with the D-dimensional feature vector $\mathbf{x}_n \in \mathbb{R}^D$. An approximate conditional joint likelihood is defined in [13] such that:

$$p(\mathbf{t}|\mathbf{X}, \beta, k, M) \approx \prod_{n=1} \frac{\exp\left\{\frac{\beta}{k} \sum\limits_{j\sim n|k}^{\mathbf{M}} \delta_{t_n t_j}\right\}}{\sum\limits_{c=1}^{C} \exp\left\{\frac{\beta}{k} \sum\limits_{j\sim n|k}^{\mathbf{M}} \delta_{c t_j}\right\}} \tag{1}$$

where $\mathbf{t}$ is the vector $[t_1, \cdots t_N]^\mathsf{T}$, $\mathbf{X}$ is the matrix $[\mathbf{x}_1, \cdots, \mathbf{x}_N]^\mathsf{T}$ and $\mathbf{M}$ denotes the metric used in the feature space:

$$d(\mathbf{x}, \mathbf{y}) = \{(\mathbf{x} - \mathbf{y})^T \mathbf{M} (\mathbf{x} - \mathbf{y})\}^{1/2} \tag{2}$$

The number of nearest neighbors is k and β defines a scaling variable.

The most interesting part of this likelihood is concentrated in the expression:

$$\sum_{j\sim n|k}^{\mathbf{M}} \delta_{t_n t_j} \tag{3}$$

This expression denotes the number of prototypes that belongs to the class t_n among the k nearest neighbors of $\mathbf{x}_n$ measured under the metric $\mathbf{M}$. This k nearest neighbors are considered from the $N - 1$ samples of the data set $\mathbf{X}$ when $\mathbf{x}_n$ is removed. The same expression appears in the denominator for each class c.

This expression is closely related to the nearest-neighbor error estimator used in [8]. In both cases the estimation is done over $N - 1$ samples of the data in a Leaving-One-Out way. But in the present work, the estimator is taking into account a wider neighborhood with $k > 1$.

In the work presented in [14] the metric parameter $\mathbf{M}$ is decomposed into two parts, $\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\mathbf{T}$, where $\mathbf{\Lambda}$ is a diagonal matrix with non-negative entries and $\mathbf{Q}$ is an orthogonal matrix.

In the present work, in order to make a direct comparison with [8], and using simpler distance functions, we propose two different simplified versions of the metric $\mathbf{M}$. For the first version we consider that $\mathbf{M} = \mathbf{\Lambda}$, so the distance function becomes diagonal:

$$d(\mathbf{x}, \mathbf{y}) = \{\sum_{i=1}^{D} \lambda_i (x_i - y_i)^2\}^{1/2} \tag{4}$$

For the second version we propose to use the class-dependent weights scheme $\mathbf{M} = \mathbf{\Lambda_c}$ where c is the class of the target sample in the distance function. This class-dependent dissimilarity function was proposed in [8]:

$$d(\mathbf{x}, \mathbf{y}) = \{\sum_{i=1}^{D} \lambda_{ci} (x_i - y_i)^2\}^{1/2} \tag{5}$$

where $c = class(y)$.

This last dissimilarity function is not a distance metric because it does not fulfill the symmetric property neither the triangle inequality. Anyway, this class-dependent scheme uses to improve the accuracy of the nearest-neighbor classifier in real applications. [8].

3 MCMC Approximation

Equipped with the approximate conditional joint likelihood (1) the full posterior inference will follow by obtaining the parameter posterior distribution $p(\beta, k, \mathbf{M}|\mathbf{t}, \mathbf{X}, \mathbf{M})$. Therefore, predictions of the target class label t_* of a new sample $\mathbf{x}_*$ are made by posterior averaging:

$$p(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}) = \sum_k \int p(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}, \beta, k, \mathbf{M}) p(\beta, k, \mathbf{M}|\mathbf{t}, \mathbf{X}) d\beta d\mathbf{M}$$

As the required posterior takes an intractable form a MCMC procedure is proposed in [13] and extended in [14] to enable metric inference so that the following Monte-Carlo estimate is employed

$$\hat{p}(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}) = \frac{1}{N_s} \sum_{s=1}^{N_s} p(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}, \beta^{(s)}, k^{(s)}, \mathbf{M}^{(s)})$$

where each $\beta^{(s)}, k^{(s)}, \mathbf{M}^{(s)}$ are samples obtained from the full parameter posterior $p(\beta, k, \mathbf{M}|\mathbf{t}, \mathbf{X})$ using a Metropolis style sampler.

For the Metropolis sampler we are going to assume the following priors: uniform distribution for parameter k, normal distribution for β and Dirichlet distribution for $\mathbf{\Lambda}$, further details about this parameter distributions appear in [14] and [15].

3.1 Maximum Likelihood Parameter Selection

One of the major drawbacks of this Bayesian inference approach is to deal with a large set (N_s) of parameter combinations ($k^{(s)}$,$\beta^{(s)}$, and $\mathbf{\Lambda^{(s)}}$ or $\mathbf{\Lambda_c^{(s)}}$) in the classification stage. Aim at reducing this drawback we propose to select an unique parameter combination, the one that maximizes likelihood estimation over the training data, equation (1). This parameter combination is going to be denoted as $\beta^{(m)}, k^{(m)}, \mathbf{\Lambda}^{(m)}$ or $\mathbf{\Lambda_c}^{(m)}$ in case of class-dependent dissimilarity function.

The predictions of the target class label t_* of a new datum $\mathbf{x}_*$ are going to be made by using an unique parameter combination:

$$\hat{p}(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}) = p(t_*|\mathbf{x}_*, \mathbf{t}, \mathbf{X}, \beta^{(m)}, k^{(m)}, \mathbf{M}^{(m)})$$

Note that this approximation lead to a classifier with the same computational effort than the standard K-nn. Moreover, fast nearest neighbor search methods can be easily adapted to work with diagonal distance functions. Thus, this maximum likelihood approximation can be use in a large-scale scenario without any computational degradation with respect to the standard K-nn.

Table 1. Benchmark data sets used in the experiments. N, C, and D are, respectively, the total number of vectors, the number of classes and the dimension of each data set. In two data sets, Australian and Heart, D is the dimension after expanding categorical features (the corresponding original dimensions were 14 and 13, respectively).

Task	N	C	D
Australian	690	2	42
Balance	625	3	4
Cancer	685	2	9
Diabetes	768	2	8
German	1,000	2	24
Glass	214	6	9
Heart	270	2	25
Ionosphere	351	34	2
Liver	345	2	6
Sonar	208	60	2
Vehicle	846	4	18
Vote	435	2	10
Vowel	440	11	16
Wine	178	3	13

4 Experiments

The capabilities of the proposed approach have been empirically assessed through several *standard benchmark corpus* from the well known UCI Repository of Machine Learning Databases and Domain Theories [16] and the STATLOG Project [17].

The experiments have been carried out using 14 different classification task from the UCI Machine Learning repository. The results of the proposed method are compared with the results of the conventional k-nn method with Euclidean distance and estimating the k value by leaving-one-out.

For all the data sets *B-Fold Cross-Validation* [18] (B-CV) has been applied to estimate error rates. Each corpus is divided into B blocks, B is fixed to 5, using $B - 1$ blocks as a training set and the remaining block as a test set. Therefore, each block is used exactly once as a test set. This process is repeated 10 times with different random partitions of the data into B folds. It is worth mentioning that each feature is normalized to have $\mu = 0$ and $\sigma = 1.0$ and these parameters are estimated using only the training set of each cross validation.

The experiments in this sub-section were carried out to compare the results obtained using the conventional k-nn and four different PKnn-based approaches. The first two are the $PKnn$ with diagonal distance and the same but using a unique combination of parameters, those that obtain the maximum likelihood through the MCMC procedure, $PKnn_m$. The second two are the PKnn with class-dependent diagonal distances, $PKnnc$ and its maximum likelihood version, $PKnnc_m$.

Results are shown in Table 2. Most of the proposed techniques achieved better results than the baseline k-nn classifier. It is important to note the good behaviour of the maximum versions of the $PKnn$ which have exactly the same computational cost on the classification phase than the conventional k-nn algorithm. On the other hand the class-dependent version improves the accuracy over the $PKnn$ on six datasets.

Table 2. The results using boldface are the best for each task

	$k-nn$	$PKnn$	$PKnn_m$	$PCknn$	$PCknn_m$
Australian	17.1	**13.4**	13.6	14.5	15.1
Balance	13.8	**11.9**	12.4	12.0	12.6
Cancer	8.3	**3.5**	3.8	3.7	3.9
Diabetes	27.5	**23.4**	23.7	23.5	24.1
German	26.8	26.8	26.8	**25.5**	25.7
Glass	33.8	**31.3**	31.9	33.0	33.9
heart	17.3	16.1	17.1	**15.5**	16.9
Ionosphere	14.0	13.1	13.7	**10.1**	10.2
Liver	39.4	37.8	38.2	**37.4**	39.2
Sonar	26.4	25.4	**24.9**	25.2	25.3
Vehicle	31.3	**28.7**	**28.7**	29.8	29.9
Vote	7.4	5.0	**4.9**	5.3	5.3
Vowel	1.7	**1.7**	1.9	2.1	2.4
Wine	**2.2**	3.2	3.7	3.0	3.7

Table 3. The results using boldface are significantly (95%) better

	CW	$PKnn$	$PKnn_m$
australian	17.4	**13.4**	**13.6**
balance	18.0	**11.9**	**12.4**
cancer	3.7	3.5	3.8
diabetes	30.2	**23.4**	**23.7**
german	28.0	26.8	26.8
glass	28.5	31.3	31.9
heart	22.3	**16.1**	**17.1**
liver	40.2	37.8	38.2
vehicle	29.4	28.7	28.7
vote	6.6	5.0	4.9
vowel	1.4	1.7	1.9
wine	1.4	3.2	3.7

In order to asses the capabilities of the proposed approach it is interesting to compare it with the CW method proposed in [8]. This method obtain a class-dependent distance optimized for the 1-nn classifier. This approach can be considered among the state of the art for nearest neighbor distance learning. Table 3 shows the comparison between the here proposed methods $PKnn$ and $PKnn_m$, and the CW. The results reported by CW in different datasets showed that it is a very accurate algorithm taking into account that is based on the 1-nn. But clearly the here proposed method has the capability to learn the local neighborhood size to be considered, the k value. This capability is very important because clearly some applications are solved more naturally by means of considering a local neighborhood grater than the 1-nearest neighbor in order to define properly the class posterior probabilities. It is important to note the benefits obtained by the here proposed method $PKnn_m$ taking into account that it has the same complexity than the CW.

5 Conclusions

The present paper have shown the benefits of the Bayesian approach applied to the well-known k-nearest neighbor classifier. The results show an improvement of such classifier when both, the k value and the distance used, are optimized simultaneously by means of a MCMC procedure. The computational cost of the classification stage can be alleviated using only an unique combination of parameters, the parameters that obtained the maximum of the likelihood. This important reduction of the computational cost still keeps on providing accuracy improvements over the standard K-nn algorithm.

Future work will be focused on speed up the learning stage over the MCMC that it is still a relevant drawback of the PKnn, more concretely if we want to use of the PKnn in large-scale problems.

References

1. Short, R., Fukunaga, K.: A new nearest neighbor distance measure. In: Proceedings 5th IEEE Int. Conf. Pattern Recognition, Miami Beach, FL, pp. 81–86 (1980)
2. Ricci, F., Avesani, P.: Data Compression and Local Metrics for Nearest Neighbor Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(4), 380–384 (1999)
3. Paredes, R., Vidal, E.: A class-dependent weighted dissimilarity measure for nearest neighbor classification problems. Pattern Recognition Letters 21, 1027–1036 (2000)
4. Domeniconi, C., Peng, J., Gunopulos, D.: Locally Adaptive Metric Nearest Neighbor Classification. IEEE Transaction on Pattern Analysis and Machine Intelligence 24(9), 1281–1285 (2002)
5. de Ridder, D., Kouropteva, O., Okun, O., PietikÃd'inen, M., Duin, R.P.W.: Supervised locally linear embedding. In: Kaynak, O., Alpaydın, E., Oja, E., Xu, L. (eds.) ICANN 2003 and ICONIP 2003. LNCS, vol. 2714, pp. 333–341. Springer, Heidelberg (2003)
6. Peng, J., Heisterkamp, D.R., Dai, H.: Adaptive Quasiconformal Kernel Nearest Neighbor Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5)
7. de Ridder, D., Loog, M., Reinders, M.J.T.: Local fisher embedding. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 2, pp. 295–298 (2004)
8. Paredes, R., Vidal, E.: Learning weighted metrics to minimize nearest-neighbor classification error. IEEE Transactions on Pattern Analisys and Machine Intelligence 28(7), 1100–1111 (2006)
9. Wilson, D.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trnas. Syst., Man, Cyber. SMC-2, 408–421 (1972)
10. Ferri, F., Albert, J., Vidal, E.: Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules. IEEE Trnas. Syst., Man, Cyber. Part B: Cybernetics 29(4), 667–672 (1999)
11. Paredes, R., Vidal, E.: Weighting prototypes. A new editing approach. In: Proceedings 15th. International Conference on Pattern Recognition, Barcelona, vol. 2, pp. 25–28 (2000)
12. Paredes, R., Vidal, E.: Learning prototypes and distances: a prototype reduction technique based on nearest neighbor error minimization. Pattern Recognition 39(2), 180–188 (2006)
13. Holmes, C.C., Adams, N.M.: A probabilistic nearest neighbour method for statistical pattern recognition. Journal of the Royal Statistical Society Series B 64(2), 295–306 (2002)

14. Everson, R., Fieldsend, J.: A variable metric probabilistic *k*-nearest-neighbours classifier. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 654–659. Springer, Heidelberg (2004)
15. Manocha, S., Girolami, M.A.: An empirical analysis of the probabilistic k-nearest neighbour classifier. Pattern Recogn. Lett. 28(13), 1818–1824 (2007)
16. Blake, C., Keogh, E., Merz, C.: UCI Repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`
17. D. Statistics, M. S. S. S. University., Statlog Corpora, `ftp.strath.ac.uk`
18. Raudys, S., Jain, A.: Small Sample Effects in Statistical Pattern Recognition: Recommendations for Practitioners. IEEE Trans. on Pattern Analysis and Machine Intelligence 13(3), 252–264 (1991)

Explicit Length Modelling for Statistical Machine Translation*

Joan Albert Silvestre-Cerdà, Jesús Andrés-Ferrer, and Jorge Civera

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{jsilvestre,jandres,jcivera}@dsic.upv.es

Abstract. Explicit length modelling has been previously explored in statistical pattern recognition with successful results. In this paper, two length models along with two parameter estimation methods for statistical machine translation (SMT) are presented. More precisely, we incorporate explicit length modelling in a state-of-the-art log-linear SMT system as an additional feature function in order to prove the contribution of length information. Finally, promising experimental results are reported on a reference SMT task.

Keywords: Length modelling, log-linear models, phrase-based models, statistical machine translation.

1 Introduction

Length modelling is a well-known problem in pattern recognition which is often disregarded. However, it has provided positive results in applications such as author recognition [21], handwritten text and speech recognition [25], and text classification [8], whenever it is taken into consideration.

Length modelling may be considered under two points of view. On the one hand, the so-called implicit modelling in which the information about the length of the sequence is indirectly captured by the model structure. This is the case of handwritten text and speech recognition [9] and language modelling [4], which often include additional states to convey length information.

On the other hand, we may perform an explicit modelisation by incorporating a probability distribution in the model to represent length variability in our data sample [20]. Explicit modelling can be found in language modelling [10,16], and bilingual sentence alignment and segmentation [3,7], among others.

This work focuses on explicit length modelling for statistical machine translation (SMT). The aim of SMT is to provide automatic translations between

* Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018) and iTrans2 (TIN2009-14511) projects. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067, and by the "Vicerrectorado de Investigación de la UPV" under grant 20091027.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 273–280, 2011.

languages, based on statistical models inferred from translation examples. Two approaches to length modelling are proposed for a state-of-the-art SMT system. The rest of the paper is structured as follows. Next section describes related work in SMT. Section 3 introduces the log-linear framework in the context of SMT and Section 4 explains the two proposed length models. Experimental results are reported in Section 5. Finally, conclusions are discussed in Section 6.

2 Related Work

Length modelling in SMT has received little attention since Brown's seminal paper [2] until recently. Nowadays state-of-the-art SMT systems are grounded on the paradigm of phrase-based translation [14], in which sentences are translated as segments of consecutive words. Thereby, most recent work related to length modelling has been performed at the phrase level with a notable exception [23]. Phrase length modelling was initially presented in [22] where the difference ratio between source and target phrase length is employed to phrase extraction and scoring with promising results. Zhao and Vogel [24] discussed the estimation of a phrase length model from a word fertility model [2], using this model as an additional score in their SMT system. In [6], a word-to-phrase model is proposed which includes a word-to-phrase length model. Finally, [1] describes the derivation and estimation of a phrase-to-phrase model including the modelisation of the source and target phrase lengths.

However, any of the previous works report results on how phrase length modelling contributes to the performance of a state-of-the-art phrase-based SMT system. Furthermore, phrase-length models proposed so far depend on their underlying model or phrase extraction algorithm, which differ from those employed in state-of-the-art systems. The current work is inspired on the phrase length model proposed in [1], but applied to a state-of-the-art phrase-based SMT system [15] in order to study the contribution of explicit length modelling in SMT.

3 Log-Linear Modelling

In SMT, we formulate the problem of translating a sentence as the search of the most probable target sentence $\hat{y}$ given the source sentence x

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, Pr(y \mid x)\,. \tag{1}$$

However, state-of-the-art SMT systems are based on log-linear models that combine a set of feature functions to directly model this posterior probability

$$Pr(y \mid x) = \tfrac{1}{Z(x)}\, \exp\Big(\textstyle\sum_i \lambda_i\, f_i(x,y)\Big)\,, \tag{2}$$

being λ_i, the weight for the i-th feature function $f_i(x,y)$ and $Z(x)$, a normalisation term so that the posterior probability sums up to 1. Feature weights are usually optimised according to minimum error rate training (MERT) on a development set [17].

Common feature functions in SMT systems are those related with the translation, $f_i(x, y) = \log p(x \mid y)$; and language models, $f_j(x, y) = \log p(y)$. The language model is represented by an n-gram language model [4] and is incorporated as a single feature function into the log-linear model. In contrast, the translation model is decomposed into several factors that are treated as different feature functions. These factors are introduced in the following section.

4 Explicit Length Modelling

In the phrase-based approach to SMT, the translation model considers that the source sentence x is generated by segments of consecutive words defined over the target sentence y. As in [1], in order to define these segments we introduce two hidden segmentation variables

$$p(x \mid y) = \sum_l \sum_m p(x, l, m \mid y)\,, \tag{3}$$

being l, the source segmentation variable and m, the target segmentation variable. Thus, we can factor Eq. (3) as follows

$$p(x, l, m \mid y) = p(m \mid y)\, p(l \mid m, y)\, p(x \mid l, m, y)\,, \tag{4}$$

where $p(m \mid y)$ and $p(l \mid m, y)$ are phrase length models, whilst $p(x \mid l, m, y)$ constitutes the phrase-based translation model.

We can independently factorise terms in Eq. (4) from left to right,

$$p(m \mid y) = \prod_t\ p(m_t \mid m_1^{t-1}, y)\,, \tag{5}$$

$$p(l \mid m, y) = \prod_t\ p(l_t \mid l_1^{t-1}, m, y)\,, \tag{6}$$

$$p(x \mid l, m, y) = \prod_t\ p(x(t) \mid x(1), \ldots, x(t-1), l, m, y)\,, \tag{7}$$

where t ranges over the possible segmentation positions of the target sentence, m_t and l_t are the length of the t-th source and target phrase, respectively, and $x(t)$ is the t-th source phrase.

The model in Eq. (5) becomes a feature function, known as phrase penalty, intended to control the number of phrases involved in the construction of a translation. Eq. (6) is used in the following sections to derive phrase-length models that become the phrase-length feature functions of our log-linear model.

Finally, Eq. (7) is simplified by only conditioning on the t-th target phrase to obtain the conventional phrase-based feature function

$$p(x(t) \mid x(1), \ldots, x(t-1), l, m, y) := p(x(t) \mid y(t))\,, \tag{8}$$

with parameter set, $\theta = \{p(u \mid v)\}$, for each source, u, and target, v, phrase.

4.1 Standard Length Model

The standard length model is derived from Eq. (6) by taking the assumption

$$p(l_t \mid l_1^{t-1}, m, y) := p(l_t \mid m_t)\,, \tag{9}$$

in which $p(l_t \mid m_t)$ is a source phrase-length model conditioned on the target phrase length with parameter set $\gamma = \{p(l \mid m)\}$ for each source, l, and target, m, lengths.

4.2 Specific Length Model

In the specific model, we take a more general assumption for Eq. (6) than that of Eq. (9) by just considering the dependency on $y(t)$,

$$p(l_t \mid l_1^{t-1}, m, y) := p(l_t \mid y(t))\,, \tag{10}$$

being $p(l_t \mid y(t))$, a source phrase-length model conditioned on the t-th target phrase. The parameter set of this model, $\psi = \{p(l \mid v)\}$, is considerably more sparse than that of the standard model. Hence, to alleviate overfitting problems, the specific parameters were smoothed with the standard parameters as follows

$$\tilde{p}(l \mid v) := (1-\varepsilon) \cdot p(l \mid v) + \varepsilon \cdot p(l \mid |v|)\,, \tag{11}$$

denoting by $|\cdot|$ the length of the corresponding phrase.

4.3 Estimation of Phrase-Length Models

The parameters of the models introduced in the previous section could be estimated by maximum likelihood criterion using the EM algorithm [5]. As shown in [1], the phrase-based translation model is estimated as

$$p(u \mid v) = \frac{N(u,v)}{\sum_{u'} N(u',v)}\,, \tag{12}$$

being $N(u,v)$, the expected counts for the bilingual phrase (u,v). The estimation of $p(l \mid m)$ is computed as

$$p(l \mid m) = \frac{N(l,m)}{\sum_{l'} N(l',m)}\,, \tag{13}$$

where

$$N(l,m) = \sum_{u,v} \delta(l,|u|)\,\delta(m,|v|)\,N(u,v)\,, \tag{14}$$

being δ, the Kronecker delta. The parameter $p(l\,|\,v)$ is estimated analogously.

Nevertheless, the parameter estimation of conventional log-linear phrase-based systems approximate the expected counts $N(u,v)$ with approximated counts $N^*(u,v)$, derived from a heuristic phrase-extraction algorithm [13] . Similarly, our first approach is also to approximate $N(l,m)$ in Eq. (13) as follows

$$N^*(l,m) = \sum_{u,v} \delta(l,|u|)\,\delta(m,|v|)\,N^*(u,v)\,. \tag{15}$$

This approach is referred as to phrase-extract estimation.

A second approach to the estimation of phrase-length parameters is based on the idea of a Viterbi approximation to Eq. (3). This approach only considers the source and target segmentation that maximises Eq. (3)

$$\hat{l}, \hat{m} = \underset{l,m}{\operatorname{argmax}} \{Pr(x, l, m \mid y)\}\,. \tag{16}$$

So, the hidden segmentation variables are uncovered and the counts in Eq. (12) are not expected but exact counts.

The search denoted by Eq. (16) is performed using a conventional log-linear phrase-based system which is based on a A^* search algorithm. It must be noted that the source and target sentences are available during the training phase, so this search becomes a guided search in which the target sentence is known.

5 Experimental Results

In this section, we study the benefits of explicit length modelling in phrase-based SMT. The experiments were carried out on the English-Spanish Europarl-v3 parallel corpora [11], which is a reference task in the SMT field. These corpora are provided in three separate sets for evaluation campaign purposes: training, development and test. Basic statistics are shown in Table 1.

Table 1. Basic statistics for Europarl-v3

	Training		Development		Test	
Language pairs	En	Es	En	Es	En	Es
Bilingual sentences	730740		2000		2000	
Vocabulary size	72.7K	113.9K	6.5K	8.2K	6.5K	8.3K
Running words	15.2M	15.7M	58.7K	60.6K	58.0K	60.3K
Perplexity (5-grams)	-	-	79.6	78.8	78.3	79.8

To evaluate the performance of length modelling, the models proposed in Sections 4.1 and 4.2 were introduced as feature functions into Moses [15], which is a state-of-the-art phrase-based SMT system.

Moses includes a set of feature functions that are related not only to translation and language models, but also to specific linguistic phenomena such as phrase reordering and word fertility [13]. For the sake of brevity we only focus on the description of translation features, since the rest of features are the same in all the experiments performed. Translation feature functions are constituted by the conventional target-to-source phrase-based model presented in Eq. (8) and a smoothing target-to-source word-based model. Although it might seem awkward, their source-to-target counterpart features are also included; along with the phrase penalty mentioned in Section 4 [13].

This set of features defines the baseline SMT system to which we compare our systems extended with phrase-length models. In our phrase-length augmented systems, we include two additional feature functions to account for the source-to-target and target-to-source phrase-length models.

In order to gauge the translation quality of the different systems, the well-known BLEU score [18] was used. BLEU score is an accuracy measure of the degree of n-gram overlapping between the system and the reference translation.

BLEU scores reported in Figures 1 and 2 on the test set are intended to assess not only the two phrase-length models proposed, but also the two approaches for parameter estimation stated in Section 4.3. A selection of the most relevant and representative experiments are shown in this section. Similar conclusions can be drawn from the omitted results.

Figure 1 shows the evolution of the BLEU score (y-axis) as a function of the maximum phrase length (x-axis) in order to study the behaviour of the two estimation approaches on the same phrase-length model. On the left-hand side, we analyse the standard model (Std) on the Spanish-to-English (Es-En) task, while on the right side, the specific (Spc) model is studied on the English-to-Spanish (En-Es) task.

In both cases, there are no statistically significant differences between the phrase-extract and the Viterbi estimation approaches, and also when comparing to the baseline system [12]. Although differences are not statistically significant, the specific model systematically supersedes the baseline system when maximum phrase length ranges from 4 to 7 on the English-to-Spanish pair, being the maximum difference 0.6 BLEU for the reference SMT system (maximum phrase length = 7).

Surprisingly, on this latter pair, the parameter estimation based on the phrase-extract approach systematically improves the results provided by the Viterbi estimation. This is probably due to the fact that the Viterbi estimation is not iterated, as it should be in a standard Viterbi EM-based estimation algorithm. We intend to tackle this issue as future work.

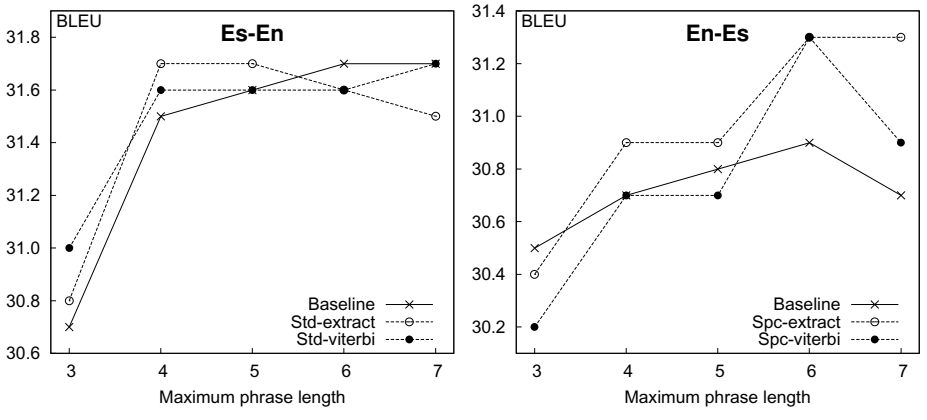


Fig. 1. Comparison of the two parameter estimation approaches for the standard (left) and specific (right) phrase-length models, and the baseline system. On the left-hand side, the evolution of the BLEU score (y-axis) as a function of the maximum phrase length (x-axis) on the Spanish-to-English (Es-En) task is shown. On the right-hand side, we consider the same experimental setting on the English-to-Spanish (En-Es) task.

In Figure 2, we directly compare the performance between the standard and specific length models in terms of BLEU scores. On left-hand side, we consider the Viterbi parameter estimation on the Spanish-to-English task, and on the right-hand side, the phrase-extract estimation on the English-to-Spanish task.

On the Spanish-to-English task, as happened in Figure 1, the three SMT systems obtain similar performance at different maximum phrase length. On the English-to-Spanish task, both phrase-length models achieve better performance than the baseline system, though not statistically significant, systematically superior from maximum phrase length equal to 4 up to 7. In this latter task, the specific length model obtains better figures than the more simple standard length model with maximum phrase length equal or greater than 5.

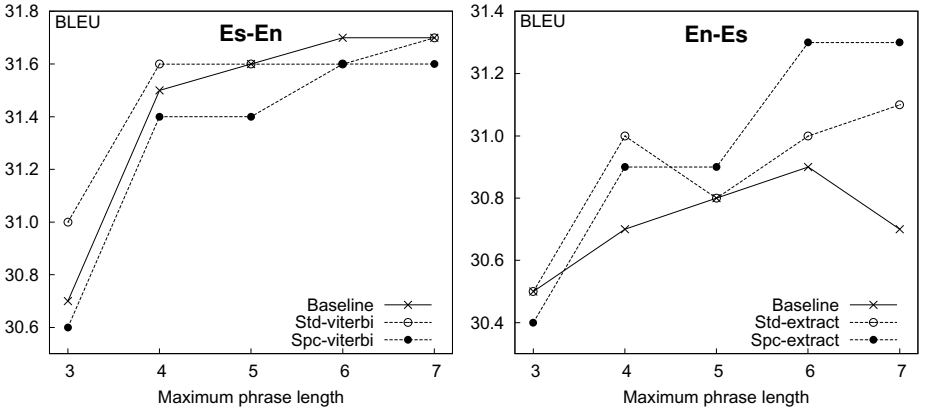


Fig. 2. BLEU figures as a function of the maximum phrase length for the Viterbi parameter estimation (left) and for the phrase-extract estimation (right), in order to compare the two proposed length models. Spanish-to-English and English-to-Spanish results are shown on the left and right-hand side, respectively.

6 Conclusions and Future Work

In this paper, we have presented two novel phrase-length models along with two alternative parameter estimation approaches. These phrase-length models have been integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing a significant boost of BLEU score in the reference SMT system on the English-to-Spanish task.

From the comparison of phrase-length models and parameter estimation approaches on the English-to-Spanish task, two conclusions can be drawn. First, the phrase-extract estimation is systematically better than the Viterbi approximation, and secondly, the specific model is superior to the standard model when dealing with long phrases. As future work, we plan to perform a full Viterbi iterative training algorithm that may improve the quality obtained by the proposed Viterbi-based estimation procedure.

Furthermore, we have observed an unstable behaviour of the evolution of the BLEU score for some maximum phrase lengths and more precisely, on the Spanish-to-English task. We believe this behaviour is explained by the weight optimisation process carried out by the MERT algorithm. To alleviate this problem we plan to use more robust weight optimisation techniques such as that proposed in [19]. Finally, we intend to extend the application of phrase-length models to the translation of more complex languages, such as Arabic or Chinese.

References

1. Andrés-Ferrer, J., Juan, A.: A phrase-based hidden semi-markov approach to machine translation. In: Proc. of EAMT, pp. 168–175 (2009)
2. Brown, P.F., et al.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics 19(2), 263–311 (1993)

3. Brown, P.F., et al.: Aligning sentences in parallel corpora. In: Proc. of ACL, pp. 169–176 (1991)
4. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proc. of ACL, pp. 310–318 (1996)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statistical Society. Series B 39(1), 1–38 (1977)
6. Deng, Y., Byrne, W.: HMM word and phrase alignment for statistical machine translation. IEEE Trans. Audio, Speech, and Lang. Proc. 16(3), 494–507 (2008)
7. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Proc. ACL, pp. 177–184 (1991)
8. Giménez, A., et al.: Modelizado de la longitud para la clasificación de textos. In: Actas del I Workshop de Rec. de Formas y Análisis de Imágenes, pp. 21–28 (2005)
9. Günter, S., Bunke, H.: HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and gaussian components. Pattern Recognition 37(10), 2069–2079 (2004)
10. Kneser, R.: Statistical language modeling using a variable context length. In: Proc. of ICSLP (1996)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit X, pp. 79–86 (2005)
12. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proc. of EMNLP, pp. 388–395 (2004)
13. Koehn, P.: Stadistical Machine Translation. Cambridge University Press, Cambridge (2010)
14. Koehn, P., et al.: Statistical phrase-based translation. In: HLT, pp. 48–54 (2003)
15. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Proc. of ACL (2007)
16. Matusov, E., et al.: Automatic Sentence Segmentation and Punctuation Prediction for Spoken Language Translation. In: Proc. of IWSL, pp. 158–165 (2006)
17. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proc. of ACL, pp. 160–167 (2003)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a Method for Automatic Evaluation of Machine Translation. Tech. rep., Watson Research Center (2001)
19. Sanchis-Trilles, G., Casacuberta, F.: Log-linear weight optimisation via Bayesian Adaptation in Statistical Machine Translation. In: COLING, pp. 1077–1085 (2010)
20. Sichel, H.S.: On a distribution representing sentence-length in written prose. J. Roy. Statistical Society. Series A 137(1), 25–34 (1974)
21. Uzuner, Ö., Katz, B.: A comparative study of language models for book and author recognition. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 969–980. Springer, Heidelberg (2005)
22. Venugopal, A., et al.: Effective phrase translation extraction from alignment models. In: Proc. of ACL, pp. 319–326 (2003)
23. Zens, R., Ney, H.: N-gram posterior probabilities for statistical machine translation. In: Proceedings of WSMT, pp. 72–77 (2006)
24. Zhao, B., Vogel, S.: A generalized alignment-free phrase extraction. In: Proc. of ACL Workshop on Building and Using Parallel Texts, pp. 141–144 (1995)
25. Zimmermann, M., Bunke, H.: Hidden markov model length optimization for handwriting recognition systems. In: Proc. of IWFHR, pp. 369–374 (2002)

Age Regression from Soft Aligned Face Images Using Low Computational Resources

Juan Bekios-Calfa[1], José M. Buenaposada[2], and Luis Baumela[3]

[1] Dept. de Ingeniería de Sistemas y Computación, Universidad Católica del Norte
Av. Angamos 0610, Antofagasta, Chile
juan.bekios@ucn.cl
[2] Dept. de Ciencias de la Computación, Universidad Rey Juan Carlos
Calle Tulipán s/n, 28933, Móstoles, Spain
josemiguel.buenaposada@urjc.es
[3] Dept. de Inteligencia Artificial, Universidad Politécnica de Madrid
Campus Montegancedo s/n, 28660 Boadilla del Monte, Spain
lbaumela@fi.upm.es

Abstract. The initial step in most facial age estimation systems consists of accurately aligning a model to the output of a face detector (e.g. an Active Appearance Model). This fitting process is very expensive in terms of computational resources and prone to get stuck in local minima. This makes it impractical for analysing faces in resource limited computing devices. In this paper we build a face age regressor that is able to work directly on faces cropped using a state-of-the-art face detector. Our procedure uses K nearest neighbours (K-NN) regression with a metric based on a properly tuned Fisher Linear Discriminant Analysis (LDA) projection matrix. On FG-NET we achieve a state-of-the-art Mean Absolute Error (MAE) of 5.72 years with manually aligned faces. Using face images cropped by a face detector we get a MAE of 6.87 years in the same database. Moreover, most of the algorithms presented in the literature have been evaluated on single database experiments and therefore, they report optimistically biased results. In our cross-database experiments we get a MAE of roughly 12 years, which would be the expected performance in a real world application.

1 Introduction

Age is a demographic variable that can be estimated using visual cues such as facial appearance, gait, clothing or hair style and non-visual cues like the voice. Automatic age estimation has interesting applications to enforce legal age restrictions in vending machines, automate marketing studies in shopping centres, measure tv audience or recognise faces automatically from videos. The aim of this paper is to use facial appearance as a visual cue to estimate the age of a person.

The facial age estimation problem is difficult since we are trying to estimate the real age from the face appearance, which depends on environmental

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 281–288, 2011.

conditions like health, eating habits, sun exposure record, etc [13]. Facial age estimation can be seen either as a classification problem (i.e. different age groups or ranges) or a regression problem.

The state-of-the-art on age estimation can be organised into hard aligned (AAMs or manually) results and soft aligned results. There are two key references in the hard aligned group: the Bio-inspired Features (BIF) [5] and the Regression from Patch Kernel (RPK) [12]. The BIF approach uses a bank of Gabor filters at different scales and orientations with a combination layer and a PCA reduction step over manually aligned faces of 60×60 pixels. Although the result is 4.77 years of MAE in leave-one-person-out cross-validation, the best reported so far, the computational requirements of the method are quite high. The RPK approach breaks the 32×32 pixels input image into equally sized patches (8×8 pixels each). Then each patch is described by Discrete Cosine Transform (DCT) and the position in the image plane is added to the descriptor. The probability distribution of the patch descriptors within an image is modelled by a mixture of Gaussians and the age is finally estimated by Kernel Regression [12]. This approach achieves a MAE of 4.95 years on FG-NET, with standard leave-one-subject-out cross-validation.

Concerning soft aligned results, [6] performs training and testing directly on the output of the face detector. They extract Histogram of Oriented Gradients (HoG), Locally Binary Patterns (LBP) and local intensity differences from local patches in a regular image grid. The regressor is based on a Random Forest trained with 250 randomly selected images from FG-NET. They achieve a MAE of 7.54 years. Their result is optimistically biased since the same subject may be in the training and testing sets. In [3], they use semi-supervised learning using web queries, multiple face detectors and robust multiple instance learning. They use DCT local image descriptors and a fully automated pipeline from database collection to age regression estimation. The main limitation of this approach for a resource limited device is its computational complexity.

An important issue to consider is whether it is worth using computationally intensive face alignment procedures rather than learning to estimate face age with unaligned images. Most face age estimation results use Active Appearance Models (AAMs) for face alignment [13]. Unfortunately, fitting an AAM to unseen faces is prone to get stuck in local minima [10]. Moreover, fitting an AAM can be a computationally prohibitive task when there are many faces in the image or when the computation is performed on a resource limited device, such as a smart phone or an IP camera. An alternative here is using soft aligned algorithms, which require no accurate alignment to the input face image [3,6].

In this paper we follow a soft alignment approach and train our regressor with cropped faces obtained from a face detector. We use K-NN regression for age estimation using a learned metric. Our metric is derived from Fisher Linear Discriminant Analysis. By computing the LDA projection matrix using age groups we impose that similar aged faces be close to each other and far apart from different aged ones. With this approach we can get roughly state-of-the-art age estimation. By dealing with the misalignment during training, the on-line

classification algorithm is quite simple and efficient. We train our algorithms in one database and test them in a different one (see section 3.2). With leave-one-person-out cross-validation in FG-NET we get a MAE of 5. With cross-database tests we achieve a MAE of 12 years, which is a more realistic value for a real application.

2 Age Regression from Face Images

We use a non-linear regressor based on K-NN for age estimation. Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ be $p^2 \times 1$ vectors where each $\mathbf{x}_i$ corresponds to the gray levels of a $p \times p$ pixels image scanned by columns and y_i is the age label corresponding to $\mathbf{x}_i$. The euclidean distance in the image space does not take into account the age. This means that with euclidean distance two face image vectors with different age labels could have lower distance than two face images with similar age. Therefore, we use a Mahalanobis-like distance with a learned metric matrix $\mathtt{M}$ to have similar aged face images close to each other and dissimilar aged face images far apart, $d_{\mathtt{M}}(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathtt{M}} = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathtt{M}(\mathbf{x}_i - \mathbf{x}_j)$. In the following subsections we explain how to learn the metric matrix $\mathtt{M}$ using Fisher Linear Discriminant Analysis and how we make K-NN age regression.

2.1 PCA+LDA Projection as the Age Metric Matrix

We use PCA+LDA (Linear Discriminant Analysis after a Principal Component Analysis projection) to compute a projection matrix $\mathtt{W}$. We compute $d_{\mathtt{M}}$ as

$$d_{\mathtt{M}}(\mathbf{x}_i, \mathbf{x}_j) = ||\mathtt{W}(\mathbf{x}_i - \mathbf{x}_j)||^2 = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathtt{W}^\top \mathtt{W}(\mathbf{x}_i - \mathbf{x}_j), \tag{1}$$

which means that the metric matrix is given by $\mathtt{M} = \mathtt{W}^\top \mathtt{W}$.

LDA is a supervised technique for dimensionality reduction that maximises the data separation of different classes. Since age is a continuous variable, to perform LDA first we have to discretise it into c age groups (see section 3 for the actual age groups we use). Given a multi-class problem with c classes and p sample points, $\{\mathbf{x}_i\}_{i=1}^{p}$ the basis of the transformed subspace, $\{\mathbf{w}_i\}_{i=1}^{d}$, is obtained by maximising [4] $J(w) = \sum_{i=1}^{d} \frac{\mathbf{w}_i^\top \mathtt{S}_B \mathbf{w}_i}{\mathbf{w}_i^\top \mathtt{S}_m \mathbf{w}_i}$, where $\mathtt{S}_B$ and $\mathtt{S}_m$ are respectively the between-class and full scatter matrices.

Depending on the amount of training data, the performance of the regressor or classifier built on LDA subspace decreases when retaining all eigenvectors associated with non-zero eigenvalues. Thus, a crucial step here is to choose which PCA eigenvectors to keep so that no discriminant information is lost. We select the dimension of the subspace resulting from the PCA step using a cross-validation scheme instead of the usual approach based on retaining the eigenvectors accounting for a given percentage of the variance (usually 95% or 99%) [7]. To this end we sort PCA eigenvectors in descending eigenvalue order. We then perform cross-validation and select the dimension with the best performance for the age regression. This feature selection process is essential to correctly train a PCA+LDA procedure [2].

2.2 K-NN Regression

We may interpret (1) as a projection of the face image onto the PCA+LDA subspace with the $\mathtt{W}$ matrix and then a classification in the transformed subspace using the euclidean metric. This is the approach we use in our K-NN regression implementation.

We project each training data vector, $\mathbf{x}_i$, onto the PCA+LDA subspace obtaining $\mathbf{z} = \mathtt{W}\mathbf{x}_i$. Once the optimal number of neighbours, K, is estimated by cross-validation, the regression output for a given input vector $\mathbf{z}$ in the PCA+LDA subspace is given by $\hat{y} = \sum_{i=1}^{K} \hat{w}_i y_i$, $\hat{w}_i = \frac{w_i}{\sum_{j=1}^{K} w_j}$, $w_i = \frac{1}{||\mathbf{z}-\mathbf{z}_i||}$, where y_i is the age label (real valued) of the i-th nearest neighbour, $\mathbf{z}_i$, to $\mathbf{z}$. When some or all of the distances are close to zero, or bellow a small threshold α (i.e. $||\mathbf{z} - \mathbf{z}_i|| <= \alpha = 10^{-6}$) we choose the label, y_i, of the nearest neighbour as the regression age label, $\hat{y} = y_i$.

3 Experiments

In this section we evaluate the performance of our age regressor and compare it with other age estimation approaches in the literature. We have used the Productive Aging Lab Face (PAL) database [9], the Images of Groups Dataset [1] and the FG-NET Aging database. To train our algorithm we estimate the number of nearest neighbours, K, and the PCA dimension optimising for the MAE in a cross-validation scheme. We crop and re-size images to a base size of 25×25 pixels using OpenCV's[1] 2.0.0 face detector, which is based on [11]. Then we equalise the histogram to gain some independence from illumination changes. Finally, we also apply an oval mask to prevent the background from influencing our results. Additionally, on FG-NET, we perform two kinds of manual alignment: 1) a similarity transform using the location of the eyes and 2) an affine transform using the location of the eyes and the centre of the mouth.

To train PCA+LDA we have discretised the age of FG-NET and PAL databases into 11 groups: 0-2, 3-7, 8-12, 13-19, 20-28, 29-37, 38-46, 47-55, 56-64, 65-73 and 74-82. On the other hand, the GROUPS database already comes with discrete age labels, which are organised in groups 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. In our experiments we use those face detections from the GROUPS database that have at least a size of 60×60 pixels (13,051 out of a total of 28,231).

Our measure for face age regression error is the Mean Absolute Error $MAE = \frac{1}{N} \sum_{i=1}^{M} |y_i - \hat{y}_i|$ where y_i is the actual label of a face image and $\hat{y}_i$ is the estimated age by a given algorithm. This is a non robust measure. To highlight outlier's influence in MAE a cumulative score curve shows the percentage of testing data below a given age estimation error (see Fig. 1). We use cumulative score curve to compare two age estimation algorithms, the higher the curve the better the algorithm.

[1] `http://opencv.willowgarage.com`

3.1 Intra-database Tests

The first set of experiments use one database for training and testing.

In the FG-Net database case we perform leave-one-person-out cross-validation. In this way we avoid the bias introduced in the evaluation when classifying images of the same person both in the training and testing sets. This means that we keep all the images of a subject for testing (around 12), training with the rest.

To quantify the influence of alignment on age regression we compare raw face detection with manual alignment in FG-NET (see table 1). The difference in MAE between global affine transformation (using eyes and mouth) and a global similarity transformation (using only the eyes) is lower than 0.3 years. When using soft aligned faces with raw face detection the MAE degrades by roughly 1.2 years.

We compare our results (see Table 1) on FG-NET with the two best published results [12,5] using leave-one-person-out cross-validation with manual eye alignment. In terms of global MAE, our eye aligned results are one year worse than [12] and [5] while our face detection result is roughly 2 years worse. The cumulative score curves in Fig. 1 right, confirms that the RPK [12] or BIF [5] approaches are marginally better than our manually aligned algorithm. On the other hand, our algorithm is much simpler and with lower computational requirements. The BIF method relies on processing the image with a large bank of filters, while RPK relies on an a mixture model adaptation of a face image description based on the distribution of the DCT on all image patches.

The work of Jahanbekam et al. [6] uses also face detection alignment on FG-NET. Their MAE is 7.54, which is optimistically biased since they do not use a leave-one-subject-out evaluation, and consequently, the same subject can be in the training and testing sets. Even in this case we outperform their approach, since for our MAE in this experiment is 6.9 (see Table 1).

Table 1. MAE on each age range in the FG-NET database with 25×25 pixels images

Experiment/Age Range	0-9	10-19	20-29	30-39	40-49	50-59	60+	Global
Affine Alignment	2.72	3.84	5.62	11.19	19.68	29.43	40.53	5.56
Similarity Alignment	2.85	3.76	5.6	11.58	19.65	27.67	42.11	5.7
Face Detection	4.68	4.39	6.57	13.62	19.84	29.68	38.12	6.9
RPK [12]	2.3	4.86	4.02	7.32	15.24	22.2	33.15	4.95
BIF [5]	2.99	3.39	4.3	8.24	14.98	20.49	31.62	4.77

3.2 Cross-Database Tests

Most age estimation algorithms only perform single database tests. To evaluate the performance of an age estimation algorithm we are interested in the algorithm's generalisation capabilities. In this section we train our algorithm using one database and test it on a different database. In the case of GROUPS and

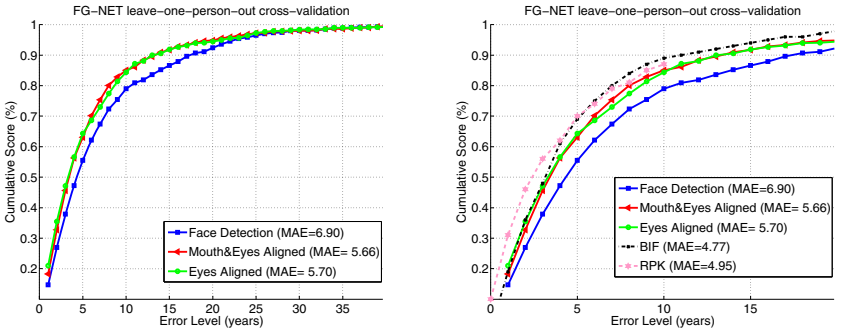


Fig. 1. Cumulative Score curves for FG-NET cross-validation experiments with 25×25 pixels sized images. Left: curves for different alignments using our method. Right: comparison with the two most competitive published methods.

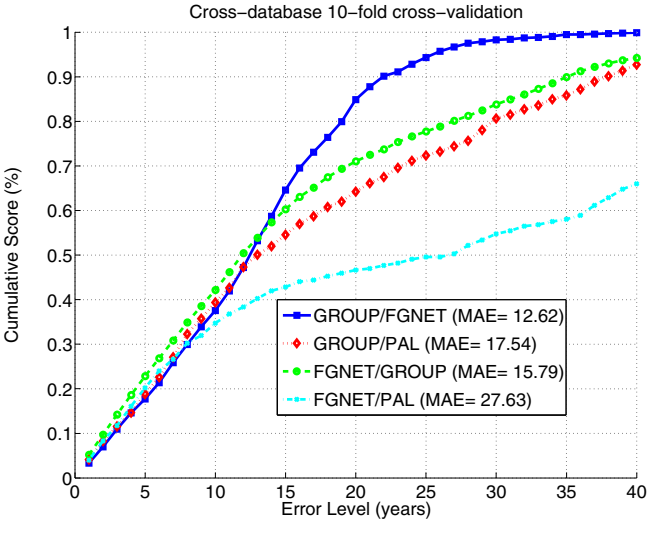


Fig. 2. Cumulative Score curves for cross-database experiments

PAL databases we train with 10-fold cross-validation. For training with FG-NET we perform leave-one-person-out cross-validation. In Fig. 2 we show the cumulative score curves and in Table 2 the MAE for our experiments.

We have made two groups of experiments. First train on a large database (GROUPS) and test on FG-NET and PAL. In this case we achieve a global MAE of about 15 years. In the second group of experiments we train with FG-NET, a small database, and test on GROUPS and PAL. The FG-NET/GROUPS experiment achieves also a MAE around 15 years. In the FG-NET/PAL case we achieve a much higher MAE because the age distribution in both databases is different. FG-NET has fewer people older than 40 whereas most of the subjects in PAL are above 40. This explains the differences on the results in Table 2.

Our results for GROUPS/FG-NET can be compared with others in the literature that use face detection and no further alignment [3]. In [3] a database from Internet with 219,892 samples is used for training. It is tested on FG-NET (see IAD/FG-NET in Table 2), being their MAE 9.49. Our result when training with a database with 13,051 samples is 12.62 for the GROUPS/FG-NET experiment in Table 2. We achieve a higher MAE because our database is one order of magnitude smaller and with a lower resolution age distribution. However, when looking at the per-age range MAEs, we get better results in 4 out of 7 age ranges (see columns IAD/FG-NET and GROUPS/FG-NET in Table 2).

Table 2. MAE on each age range in the cross-databases experiments

Experiment/Age Range	0-9	10-19	20-29	30-39	40-49	50-59	60+	Global
IAD/FG-NET[3]	10.98	8.15	6.05	7.92	13.42	22.75	29.96	9.49
GROUPS/FG-NET	15.55	12.98	6.88	5.65	12.20	19.66	22.64	12.62
GROUPS/PAL	—	10.42	7.59	6.69	9.30	17.27	28.90	17.54
FG-NET/GROUPS	9.56	5.77	9.41	—	—	29.55	53.52	15.79
FG-NET/PAL	—	5.56	5.84	14.27	23.62	32.85	49.10	27.63

4 Conclusions

In this paper we have presented a contribution to the age regression problem with results roughly within the state-of-the-art. Following the Occam Razor's principle we attack the problem from a simplicity driven perspective and with a low computational requirements solution in mind. We have realised that some solutions in the literature are computationally complex getting in return low gain age estimation performance.

With manual eye alignment we get competitive results using a very simple and fast algorithm. When using soft aligned images, by means of face detection, the MAE estimation is only one year worse than the manual alignment. It is thus unclear whether full automatic alignment, which is computationally intensive, is worthy. A similar result was reported in the gender recognition problem [8,2].

Moreover, we believe that the alignment problem can be solved by training, which would make the on-line computation much more efficient. By requiring no hard-alignment, our method is simple and fast both in training and in on-line classification. Given the low computational requirements, this method may be implemented in smart-phones or IP cameras.

The benchmark database for age estimation, FG-NET, has a very low number of images in some of the age ranges. This makes it difficult to train any learning algorithm and makes it difficult to get definitive conclusions by using only this database. Therefore, cross-database experiments are a must in order to push the state-of-the-art in facial age estimation.

Acknowledgement

The authors gratefully acknowledge funding from the Spanish *Ministerio de Ciencia e Innovación* under contracts `TIN2010-19654` and the *Consolider Ingenio* program contract `CSD2007-00018`.

References

1. Andrew, C., Gallagher, T.C.: Understanding images of groups of people. In: Proc. of CVPR, pp. 256–263 (2009)
2. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting linear discriminant techniques in gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (in press)
3. Bingbing, N., Zheng, S., Shuicheng, Y.: Web image mining towards universal age estimation. In: Proc. of ACM International Conference on Multimedia (October 2009)
4. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, London (1990)
5. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: Proc. of CVPR, pp. 112–119 (2009)
6. Jahanbekam, A., Bauckhage, C., Thurau, C.: Age recognition in the wild. In: Proc. of ICPR, pp. 392–395. IEEE, Los Alamitos (2010)
7. Johnson, R., Wichern, D.: Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs (1998)
8. Mäkinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(3), 541–547 (2008)
9. Minear, M., Park, D.C.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments and Computers 36, 630–633 (2004)
10. Ralph Gross, I.M., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing 23(11), 1080–1093 (2005)
11. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
12. Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., Huang, T.S.: Regression from patch-kernel. In: Proc. of CVPR (2008)
13. Yun Fu, G.G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 1955–1976 (2010)

Human Activity Recognition from Accelerometer Data Using a Wearable Device

Pierluigi Casale, Oriol Pujol, and Petia Radeva

Computer Vision Center, Bellaterra, Barcelona, Spain
Dept. of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain
pierluigi@cvc.uab.es

Abstract. Activity Recognition is an emerging field of research, born from the larger fields of ubiquitous computing, context-aware computing and multimedia. Recently, recognizing everyday life activities becomes one of the challenges for pervasive computing. In our work, we developed a novel wearable system easy to use and comfortable to bring. Our wearable system is based on a new set of 20 computationally efficient features and the Random Forest classifier. We obtain very encouraging results with classification accuracy of human activities recognition of up to 94%.

Keywords: Physical Activity Recognition, Wearable Computing, Pervasive Computing.

1 Introduction

Activity Recognition is an emerging field of research, born from the larger fields of ubiquitous computing, context-aware computing and multimedia. Recognizing everyday life activities is becoming a challenging application in pervasive computing, with a lot of interesting developments in the health care domain, the human behavior modeling domain and the human-machine interaction domain [3]. Even if first works about activity recognition used high dimensional and densely sampled audio and video streams [9], in many recent works ([2],[1]), activity recognition is based on classifying sensory data using one or many accelerometers. Accelerometers have been widely accepted due to their compact size, their low-power requirement, low cost, non-intrusiveness and capacity to provide data directly related to the motion of people.

In recent years, several papers have been published where accelerometer data analysis has been applied and investigated for physical activity recognition [5]. Nevertheless, few of them override the difficulty to perform experiments out-of-the-lab. The condition to perform experiments out-of-the-lab creates the need to build easy to use and easy to wear systems in order to free the testers from the expensive task of labeling the activities they perform.

In our work, we propose a new set of features extracted from wearable data that are competitive from computational point of view and able to ensure high classification results comparable with the state of the art wearable systems. The

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 289–296, 2011.

features proposed can be computed in real-time and provide physical meaning to the quantities involved in classification. The new set of features has been validated by mean of a reliable analysis comparing the new features with the majority of all the features commonly used in physical activity recognition using accelerometer data. Based on these features, we show that Random Forest classifier is an optimal classifier that reaches classification performances between 90% and 94%.

Moreover, we present a custom wearable system for human action recognition, developed in our lab, that is based on the analysis of accelerometer data. The wearable system is easy to use–users need only to start-stop the device, and comfortable to bring, having a reduced form which does not prevent any type of movement. Acceleration data can be acquired in many different, non-controlled environments allowing to overpass the laboratory limitation setting. Five basic every-day life activities like walking, climbing stairs, staying standing, talking with people and working at computer are considered in order to show its performance and robustness.

The paper is structured as follows. After discussing related work in Section 2, we describe in Section 3 how we create the dataset using in Section 3 we provide the technical details about the best features extraction for classifing human activities. In Section 4, we present the results of the classification of the activities. Finally, Section 5 concludes the paper.

2 Related Works

In [5], Mannini and Sabatini give a complete review about the state of the art of activity classification using data from one or more accelerometers. In their review, the best classification approaches are based on wavelet features using threshold classifiers. In their work, they separate high-frequency (AC) components, related to the dynamic motion the subject is performing from low-frequency (DC) components of the acceleration signal related to the influence of gravity and able to identify static postures. They extracted features from the DC components. The authors classify 7 basic activities and transitions between activities from data acquired in the lab, from 5 biaxial accelerometer placed in different part of the body, using a 17th-dimensional feature vector and a HMM-based sequential classifier, achieving 98.4% of accuracy.

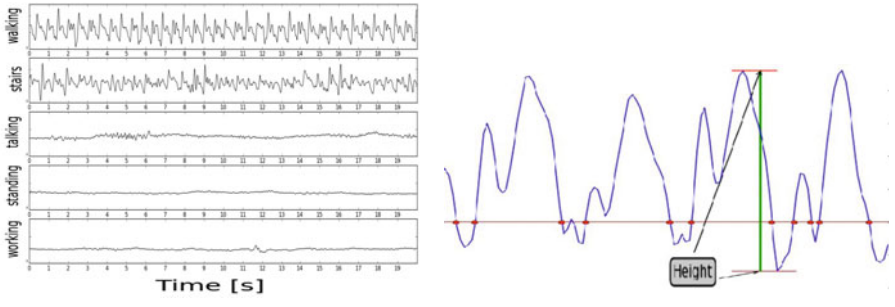
Lester, Choudhury and Borriello in [4] summarize their experience in developing an automatic physical activities recognition system. In their work, they answer some important questions about where sensors have to be placed in a person, if variation across users helps to improve the accuracy in activity classification and which are the best modalities for recognizing activities. They reach the conclusion that it does not matter where the users place the sensors, variation across users do help improving accuracy classification and the best modalities for physical activities recognition are accelerometers and microphones. Again, human activities are acquired in a controlled environment.

Our previous work in this research line [10], uses a prototype of wearable device completed by camera. Data of five everyday life activities have been

collected from people acting in two circumscribed environments. A GentleBoost classifier has been used for classifying the five activities with 83% of accuracy for each activity. Using the combination of a physical activity classifier and a face detector, face-to-face social activities have been detected with high confidence. In contrast, in this work we question how far we can get in human activities recognition using only wearable data.

3 The Problem of Human Activity Recognition

Recognizing human activities depends directly on the features extracted for motion analysis. Accelerometers provide three separated accelerometer data time series, one time series for acceleration on each axis A_x,A_y,A_z. An example of accelerometer data for five different activities is shown in Figure 1(a). Activities refer to regular walking, climbing stairs, talking with a person, staying standing and working at computer. In the figure, one can appreciate a pattern arising from a walking activity. In climbing stairs, an activity similar to walking, the same pattern seems not to be present, even if some common components between the two activities can be noted. The rest of activities differ significantly from the previous ones specially in the waveform and in the acceleration intensities involved, although forming another group of similar dynamic patterns. Small differences in the variation of the acceleration can help to discriminate the three activities. Complementary to the three axes data, an additional time series, A_m, have been obtained computing the magnitude of the acceleration:$A_m = \sqrt{A_x^2 + A_y^2 + A_z^2}$.



(a) (b)

Fig. 1. (a) Accelerometer Data for Five Different Activities..(b) Minmax sample in Accelerometer Data

3.1 Features Selection for Motion Data

Each time series A_i, with $i = \{x, y, z, m\}$ has been filtered with a digital filter in order to separate low frequencies components and high frequencies components as suggested in [5]. The cut-off frequency has been set to $1Hz$, arbitrarily. In this way, we obtain for each time series, three more time series A_{ij} with $j = \{b, dc, ac\}$, where b,dc,ac represent respectively the time series without filtering, the time

series resulting from a low pass filtering and the time series resulting from a high pass filtering. Finally, we extract features from each one of the time series.

A successful technique for extracting features from sequential motion data has been demonstrated to be windowing with overlapping. We extract features from data using windows of 52 samples, corresponding to 1 second of accelerometer data, with 50% of overlapping between windows. From each window, we propose to extract the following features: root mean squared value of integration of acceleration in a window, and mean value of Minmax sums. In next section, we will show that these two features play important role being two of the most discriminant ones because they provide informations about the physical nature of the activity being performed. The integration of acceleration corresponds to the Velocity. For each window, the integral of the signal and the RMS value of the series are computed. The integral has been approximated using running sums with step equals to 10 samples. The physical meaning that this feature provides is evident. The Minmax sums are computed as the sum of all the differences of the ordered pairs of the peaks of the time series. Note that minmax sums can be considered as a naive version of standard deviation. In Figure 1(b), an example of minmax sample is shown.

Still, in order to complete the set of features we add features that have proved to be useful for human activity recognition [5] like: mean value, standard deviation, skewness, kurtosis, correlation between each pairwise of accelerometer axis (not including magnitude), energy of coefficients of seven level wavelet decomposition. In this way, we obtain a 319-dimensional feature vector.

3.2 Classification and Derivation of Importance Measurement

Random forest [6] is an ensemble classifier that, besides classifying data, can be used for measuring attribute importance. Random Forest builds many classification trees, where each tree votes for a class and the forest choose the classification having the most votes over all the trees. Each tree is built as follows:

- If the number of cases in the training set is N, N cases are sampled at random with replacement. This sample is the training set.
- If there are M input variables, a number $m \ll M$ of variables is selected at random and the best split on these m variables is used to split the node. The value of m is held constant during the construction of the forest.
- Trees are not pruned.

When the training set for the current tree is drawn with replacement, about one-third of the cases is left out of the sample. This Out-Of-Bag (OOB) data is used to get an unbiased estimate of the classification error as trees are added to the forest. Random Forest has the advantage to assign explicitly an information measurement to each feature. Measuring the importance of attributes is based on the idea that randomly changing an important attribute between the m selected variables for building a tree affects the classification, while changing an unimportant attribute does not affect it in a significant way. Importance of all attributes for a single tree are computed as: correctly classified OOB examples

Table 1. List of Features selected by Random Forest

Feature	Importance	Feature	Importance
Mean Value A_{zdc}	4.64	Mean Value A_{ydc}	3.86
MinMax A_{zdc}	4.61	Rms Velocity A_{ydc}	3.67
RMS Velocity A_{zdc}	4.23	Mean Value A_{zb}	3.59
RMS Velocity A_{mdc}	4.2	Mean Value A_{xdc}	3.57
RMS Velocity A_{xac}	4.14	MinMax A_{xdc}	3.52
Mean Value A_{mdc}	4.07	MinMax A_{zb}	3.51
MinMax A_{ydc}	3.92	Mean Value A_{yb}	3.33
Standard Deviation A_{xb}	3.9	Rms Velocity A_{xdc}	3.22
MinMax A_{mdc}	3.89	Rms Velocity A_{zb}	3.2
Standard Deviation A_{xdc}	3.87	MinMax A_{yb}	2.96

minus correctly classified *OOB* examples when an attribute is randomly shuffled. The importance measure is obtained dividing the accumulated attribute by the number of used trees and multiplying the result by 100.

Using Random Forest, an importance measure of the features has been obtained. In Table 1, the best 20 features obtained out of 319 are reported with their respective importance value.

4 Validation and Discussions

First we discuss the architecture of our wearable system and then discuss the obtained results.

System architexture: Our wearable system, shown in Figure 2(a), is based on a Beagleboard, a low-price board built around the TI OMAP system on chip. We use Linux as operating system on the board. A low-cost USB webcam and a Bluetooth accelerometer are connected with the board. The system is powered using a portable lithium battery able to power up to four hours the system. Users can wear the system as in Figure 2(b), where the directions of the acceleration axis are printed upon the picture. More specifically, Z-axis represents the axis concordant to the direction of movement and the plane defined by the X and Y axis lies on the body of the person. The system works with three modalities, video, audio and accelerometer data. It takes photos, grabs audio continuously applying a filter for voice removal and it receives via bluetooth data from the accelerometer. All the sensors can be localized in the same part of the body. In our setting, sensors are located on the breast.

Data acquisition: Data have been collected from fourteen testers, three women and eleven men with age between 27 and 35. For labeling activities, people were asked to annotate the sequential order of the activities they performed and restart the system. Every time the system starts, data are named with a serial number. Once the user presses the starting button, she/he can start to perform the activity. The system boots in less then 2 minutes and the acquisition automatically starts while the user is already performing the activity. In this way, there are no "border effects" due to starting. The user can stop the acquisition in

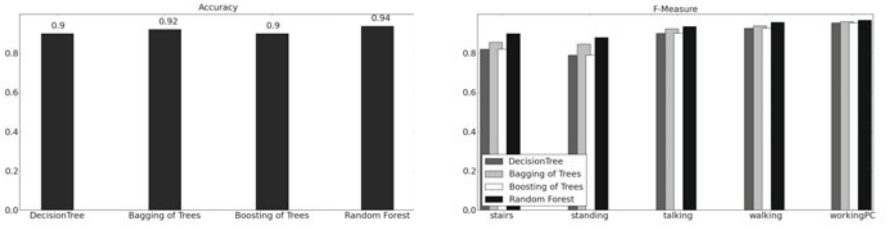
(a) (b)

Fig. 2. (a) The components of the wearable system, (b) The wearable system worn by an experimenter

every moment pressing again the start button. The data set collected is composed by 33 minutes of walking up/down stairs, 82 minutes of walking, 115 minutes of talking, 44 minutes of staying standing and 86 minutes of working at computer.
Human activity classification: Random Forest selects really meaningful features for classifying activities. The most important features selected are related to the Z axis that is, the direction of the movements. The majority of the features are relative to the DC components of movements and only the RMS velocity feature relative to the X axis from the AC components has been selected. The information relative to the variation of movements on the X axis can help to discriminate between activities like staying standing, talking and working at PC. On the other side, features relative to the variation of movements on Y axis, can help to discriminate between activities like walking and walking up/down stairs. Mean value, minmax features and RMS velocity are selected for all the DC components of all the time series. Random Forest selects the best features but it is not able to discriminate between features bringing the same information. For example, all the features selected that have been extracted from the time series without filtering are also selected from the DC time series and, in all the cases, the features selected from the DC time series have an importance value bigger than the corresponding value from the series without filtering. Features derived from higher level statistics (skewness and kurtosis) and features relative to the correlation between axis are features with the lowest importance.

In order to verify if the features selected are really informative, we use different classification methods for classifying the five activities. We compare the classification results obtained using Decision Trees, Bagging of 10 Decision Trees, AdaBoost using Decision Trees as base classifiers and a Random Forest of 10 Decision Trees. All the results are validated by 5-fold cross validation. The data set D_m has been created using the 20 features selected by the Random Forest classifier. In Figure 3(a) we show the classification accuracy of the classifiers trained on D_m. In Figure 3(b) we show the F-Measure of each activity for every classifier.

As can be seen from the graphics, the best classification accuracy is obtained using Random Forest. The F-Measure obtained for each class shows how each



(a) (b)

Fig. 3. (a) Classification Accuracy for Different Classifiers.(b) F-Measure for each Activity on the Motion Dataset.

activity can be classified with high precision and recall. In particular, activity with the best performances are walking and working at computer. Bagging and Random Forest are the classifier that give the best performances for each class. The confusion matrix obtained with the Random Forest classifier is reported in Table 2. Note how similar activity like walking and climbing stairs have some confusions between them. The biggest confusion in obtained between talking and standing, activity that can be easily confused from the perspective of motion. From Table 2 it can be concluded that all the classifiers have accuracy above

Table 2. Confusion Matrix of Random Forest trained on D_m

	stairs	walking	talking	standing	workingPC
stairs	**0.898**	0.029	0.004	0.002	0.001
walking	0.075	**0.959**	0.006	0.002	0.001
talking	0.015	0.007	**0.929**	0.093	0.012
standing	0.006	0.001	0.039	**0.888**	0.006
working	0.004	0.001	0.02	0.014	**0.977**

the 90% using only the motion modality. The Random Forest classifier trained on D_m shows confusions between similar activities like walking and walking up/down stairs, and between talking and standing. The F-Measure does not present significative differences between the classes that means that the five activities can be recognized with high confidence.

5 Conclusions

In this work, a study on the best features able to classify physical activities has been done. A new set of features has been taken into account and compared to the most commonly used features used for activity recognition in literature. The Random Forest classifier has been used to evaluate the informative measure of this new set of features. Results obtained show that the new set of features represent a very informative group of features for activity recognition. Using the features selected by Random Forest, different classifiers have been used for evaluating classification performances in activity recognition. Very high classification

performances have been reached, obtained up to 94% of accuracy using Random Forest. State of the art classification performances ([5],[4]) ensures classification performances higher than 94% when two-stages classification pipeline are used.

The validation of the new set of features has been performed using data collected using a custom wearable system, easy to use and comfortable to bring. The custom wearable device allows to perform experiments in uncontrolled environment overpassing the laboratory setting limitation. Testers perform activities in the environment they selected without the effort of labeling activities.

Based on these results obtained using only the motion sensor, future works plan to add the other sensors to increase the classification performances. We expect that adding further information from the camera and the microphone can help considerably in discriminating between activities like "standing", "talking" and "workingPC" or "walking" and "walking up/down stairs" activities where the biggest confusions are present. Moreover, we plan to extend the set of human activities in order to address the problem of short-term and long-term human behavior based on the accelerometer and video data.

Acknowledgments. This work is partially supported by a research grant from projects TIN2009-14404-C02, La Marato de TV3 082131 and CONSOLIDER (CSD2007-00018).

References

1. Ravi, N., Nikhil, D., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: IAAI, pp. 1541–1546 (2005)
2. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data, pp. 1–17. Springer, Heidelberg (2004)
3. Choudhury, T., Lamarca, A., Legr, L., Rahimi, A., Rea, A., Borriello, G., Hemingway, B., Koscher, K., Lester, J., Wyatt, D., Haehnel, D.: The Mobile Sensing Platform: An Embedded Activity Recognition System. IEEE Pervasive Computing 7, 32–41 (2008)
4. Lester, J., Choudhury, T., Borriello, G.: A practical approach to recognizing physical activities. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 1–16. Springer, Heidelberg (2006)
5. Mannini, A., Sabatini, A.M.: Machine Learning Methods for Classifying Human Physical Activities from on-body sensors. Sensors 10, 1154–1175 (2010)
6. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
7. Krause, A., Siewiorek, D., Smailagic, A., Farrigdon, J.: Unsupervised, dynamic identification of Physiological and Activity Context in Wearable Computing. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870. Springer, Heidelberg (2003)
8. Huynh, T., Fritz, M., Schiele, B.: Discovery of Activity Patterns using Topic Models. In: UbiComp 2008, pp. 10–19 (2008)
9. Clarkson, B., Pentland, A.: Unsupervised Clustering of ambulatory audio and video. In: ICASSP 1999, pp. 3037–3040 (1999)
10. Casale, P., Pujol, O., Radeva, P.: Face-to-Face Social Activity Detection Using Data Collected with a Wearable Device. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 56–63. Springer, Heidelberg (2009)

Viola-Jones Based Detectors: How Much Affects the Training Set?

Modesto Castrillón-Santana, Daniel Hernández-Sosa,
and Javier Lorenzo-Navarro

SIANI
Edif. Central del Parque Científico Tecnológico
Universidad de Las Palmas de Gran Canaria
35017 - Spain

Abstract. This paper presents a study on the facial feature detection performance achieved using the Viola-Jones framework. A set of classifiers using two different focuses to gather the training samples is created and tested on four different datasets covering a wide range of possibilities. The results achieved should serve researchers to choose the classifier that better fits their demands.

Keywords: Viola-Jones detectors, facial feature detection, training sets.

1 Introduction

The Viola-Jones face detector [16] has been extensively used thanks to the implementation available [10] in the OpenCV (Open Computer Vision) library [7]. However, Viola and Jones designed a general object detection framework that can be used for other objects. Its OpenCV implementation allows researchers to train their own classifier(s). Previously, during the sample gathering stage a large set of images is built with samples containing the object to detect (positive samples) and others not containing the target (negative samples).

Positive and negative samples gathering, data annotation, data preparation and training are uncomfortable and slow tasks that have been summarized in different brief tutorials, e.g. [14]. In this sense more recent implementations [2] have tried to keep the performance while reducing the training and test processing.

Within the facial analysis scenario, facial feature detection is a topic of interest as is may serve to reduce false positive detections when using a face detector, or to better align a detected face. Thanks to OpenCV, different face related classifiers are available to a large community of researchers [7,11]. Their performance have already been compared with different test sets, but no details related to the samples used during their training stage are available.

In this paper we train different facial feature classifiers making use of training sets of different nature, and test them with a large heterogeneous collection of face datasets, in terms of pose, illumination and resolution. We aim at providing researchers hints about how to build a detector for their particular application characteristics.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 297–304, 2011.

Section 2 summarizes the Viola-Jones object detection framework. The different datasets are briefly described in Section 3 and the results and conclusions in sections 4 and 5 respectively.

2 Viola-Jones General Object Detection Framework

Automatic face detectors have received researchers attention in last years, evolving notoriously [5,17]. In this sense recent approaches [13,16] have reduced dramatically the processing latency at high levels of accuracy, without requiring restricted heuristics based on cues such as skin color or motion. These approaches make use of a sliding window that is shifted at different scales across the whole image. Each time the area is checked with a classifier to verify whether the target pattern is present.

Following the sliding window approach, face detectors based on the framework described in [16] have achieved remarkable results while becoming well known thanks to the implementation [10] integrated in OpenCV [7]. This framework is based on the idea of a boosted cascade of weak classifiers, i.e. each one has a high detection ratio, with a reduced true reject ratio. Each classifier uses a set of Haar-like features, acting as a filter chain. Only those image regions that manage to pass through all the stages of the detector are considered as containing the target. For each stage in the cascade, a separate subclassifier is trained to detect almost all target objects while rejecting a certain fraction of those non-object patterns that have been incorrectly accepted by previous stage classifiers.

Theoretically for a cascade of K independent classifiers, the resulting detection rate, D, and the false positive rate, F, of the cascade are given by the combination of each single stage classifier rates:

$$D = \prod_{i=1}^{K} d_i \qquad\qquad F = \prod_{i=1}^{K} f_i \tag{1}$$

Each stage classifier is selected considering a combination of features which are computed on the integral image. These features are reminiscent of Haar wavelets and early features of the human visual pathway such as center-surround and directional responses. The implementation [10] integrated in the OpenCV [7] extends the original feature set [16].

With this approach, given a 20 stage detector designed for refusing at each stage 50% of the non-object patterns (target false positive rate) while falsely eliminating only 0.1% of the object patterns (target detection rate), its expected overall detection rate is $0.999^{20} \approx 0.98$ with a false positive rate of $0.5^{20} \approx 0.9 * 10^{-6}$. This schema allows a high image processing rate, due to the fact that background regions of the image are quickly discarded, while spending more time on promising object-like regions. Thus, the detector designer chooses the desired number of stages, the target false positive rate and the target detection rate per stage, achieving a trade-off between accuracy and speed for the resulting classifier.

Given an input image, the resulting classifier will report the presence and location of the object of interest.

The availability of different tutorials, e.g. [14], help OpenCV users to collect, annotate and structure the data before building the different classifiers that are later tested with an independent set of images.

3 Datasets

Being interested in testing a facial feature detector performance, we previously selected some face datasets to test. Different datasets have been used in the past to analyze face detection performance. However, we wanted to cover a wide range of situations to better characterize the classifiers under study. For that purpose four datasets of facial images have been selected:

- The CMU database [13] contains a collection of heterogeneous images divided into four different subsets *test*, *new-test*, *low-res* and *rotated* combining the test sets of Sung and Poggio [15] and Rowley, Baluja and Kanade [12]. The dataset and the annotation data corresponding to 721 faces can be obtained at [3].
- More recently initiatives such as FIW [6] have introduced new challenging situations to test the performance of the face related detectors with much larger datasets. The availability of annotation data [8] increases the number of annotated faces in real situations. In this dataset the authors provide face location information in terms of ellipses.
- The Yale Face database [1] contains a homogeneous collection of face images in different illumination conditions.
- Facity[1] is an online photo project presenting high quality frontal face images with natural illumination, no facial expression and open eyes.

Table 1 summarizes the number of images and faces available in each dataset. CMU and FDDB datasets can contain more than one face per image. The average image size (it is fixed for Yale and Facity sets), the average eye distance (in pixels) of each face annotated and the dataset standard deviation is also provided to indicate the dataset variability.

Excepting the CMU dataset, no other dataset is provided with information related to the facial features, therefore we have roughly annotated the center point of the main facial features: eyes, nose and mouth.

The criterion adopted to consider a facial feature, f_i, detection as correct, is that the euclidean distance between the annotated location, $pos_{f_i,annotated}$, and the detected location, $pos_{f_i,detected}$, must be lower than one fourth the actual eye distance. This criterion was used to estimate the eye detection success originally in [9].

[1] `www.facity.com`

Table 1. Datasets statistics. The average image dimension, average eye distance and standard deviation are expressed in pixels.

Dataset	Number of images	Average image dimension	Number of annotated faces	Average eye distance	Standard deviation
FDDB	2845	377×399	5171	99	177
CMU	180	421×422	721	64	203
Yale	165	320×243	165	55	3.4
Facity	3114	600×600	3114	206	14

4 Experiments

4.1 Classifiers

For each facial feature (eyes, nose and mouth) we have made use of two different training datasets:

- **Set A:** A collection of 6000 heterogeneous images taken randomly from the web. Using this dataset four different classifiers were trained: left eye, right eye, nose and mouth. These classifiers are already included in the current OpenCV release and have been analyzed in [4].
- **Set B:** A subset of 2300 faces of the Facity collection. Using this dataset different classifiers were trained: left eye, right eye, iris, nose, mouth, left mouth corner and right mouth corner.

For both training sets the flipped image was also used for training purposes, therefore we had around 12000 positive samples for the first family and around 4600 for the second. For both configurations around 15000 images were used as negative samples.

4.2 Results

The receiver operating characteristic (ROC) curve of each classifier is computed applying first the original release of each classifier, and two variants reducing its number of stages. Theoretically, this action must increase both correct, D, and false, F, detection rates.

The processing cost and detection precision are reflected in Table 2 for each classifier. The processing cost indicates the total time needed, in seconds, to process the whole dataset in a Core2 PC. The precision is related to the actual eye distance of the face, only for those detections considered true detections. It is observed that the classifiers computed making use of the training set B are much faster. This is justified by the simplicity of the resulting classifiers on each stage, the simpler the training images, the faster the resulting classifier. These classifiers are also similar or slightly more precise than those obtained using the training set A. They are particularly much more precise for images of similar nature than those used to train the classifiers, i.e. Facity.

Table 2. Classifier processing cost in seconds per dataset and positive detection precision, relative to the actual eye distance

Classifier	FDDB		CMU		Yale		Facity	
	Time	Precision	Time	Precision	Time	Precision	Time	Precision
Classifiers trained with set A								
Right eye	550	0.073	46	0.054	14	0.03	1546	0.02
Left eye	563	0.072	47	0.068	15	0.04	1625	0.04
Nose	927	0.069	72	0.11	19	0.04	1990	0.05
Mouth	677	0.17	58	0.11	16	0.06	1672	0.13
Classifiers trained with set B								
Iris	83	0.04	8	-	2	0.04	229	0.009
Left Eye	63	0.01	7	-	1.3	-	185	0.01
Right eye	61	0.09	7	-	1.4	-	174	0.006
Nose	90	0.1	9	0.16	2	0.07	233	0.08
Left mouth	88	-	9	0.05	2	0.04	256	0.02
Right mouth	100	-	10	0.04	2	0.04	289	0.02
Mouth	86	-	9	0.08	2	0.09	231	0.02

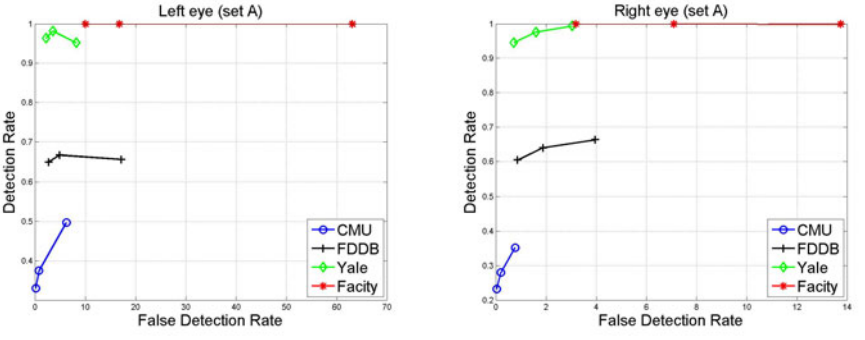


Fig. 1. Performance achieved using the left and right eye detector computed with the training set A

Those classifiers trained using set B present two important characteristics, they are faster, almost ten times for some facial features, and similar or more precise. Unfortunately, they are not so reliable to the whole dataset collection as seen in Figures 1-5. Their respective areas under the ROC curve are smaller than those presented bu the family of classifies computed with set A. To analyze each feature, Figure 1 and 2 compares the detection rate of the two classifiers specialized in the eye detection. The detectors based on the set A perform similarly for both eyes. However they are worst, as expected, for those datasets with unrestricted pose, while being really reliable with the frontal face datasets: Facity and Yale. On the other side, the left eye detector based on the set B offers a poor performance even for the Facity dataset. The iris detector presents better performance, and a reduced false detection rate, but far from that achieved using set A.

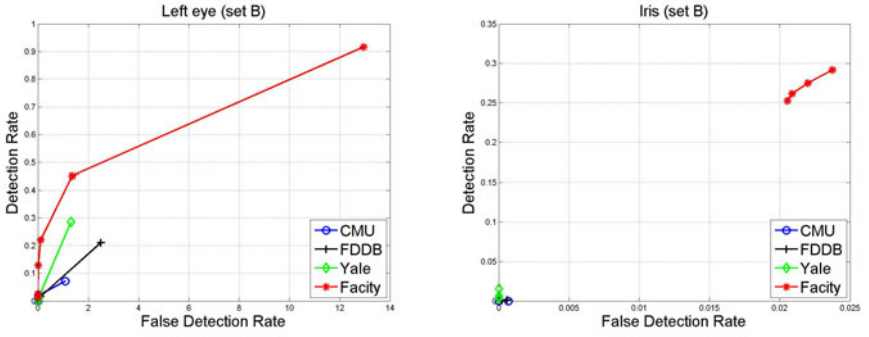


Fig. 2. Left) Left eye detection performance using the training set B. Right) Performance achieved using the iris detector computed with the training set B.

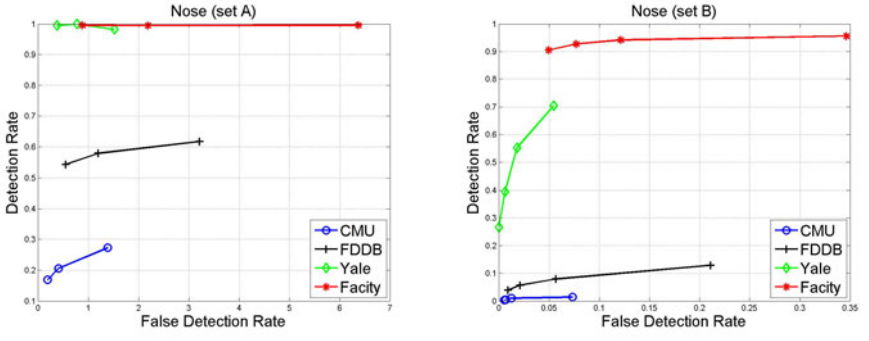


Fig. 3. Left) Nose detection performance using the training set A. Right) Nose detection performance using the training set B.

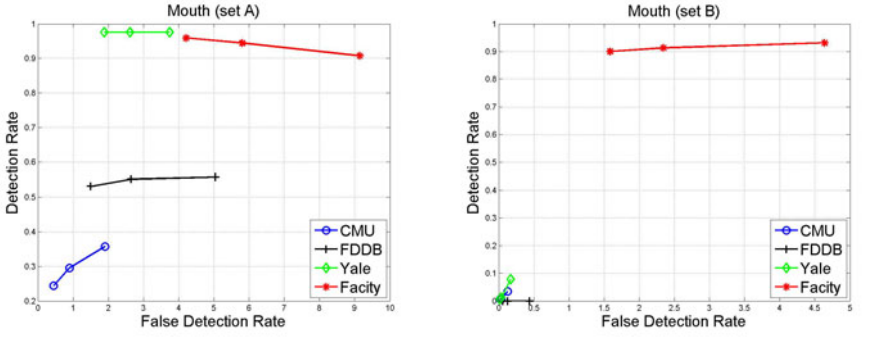


Fig. 4. Left) Mouth detection performance using the training set A. Right) Mouth detection performance using the training set B.

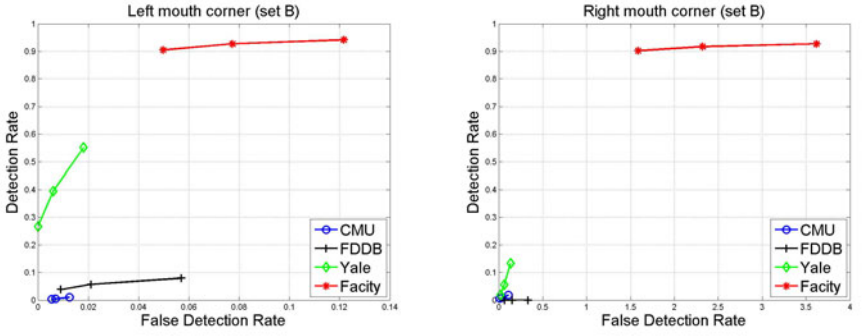


Fig. 5. Mouth corners detection performance using the training set B

The nose detection rates are presented in Figure 3. The behavior for the set A is similar to that observed for the eye pattern. The classifiers obtained with set B are now behaving better for the Yale and Facity but they never reach the reliability exhibited by those trained with set A. However, the reader must remember that this detector is much faster.

A similar performance is observed for the mouth detection, see Figure 4. We have also included the performance of the mouth corner classifiers, see Figure 5. The latter is only sensitive for the Facity dataset.

5 Conclusions

We have trained facial features detectors using two different kind of samples to build the training set. The training set A contains heterogeneous images under uncontrolled conditions, in contrast with the homogeneous training set B.

The results achieved with the training set A are more reliable than those achieved with the training set B. Those classifiers trained with set B are faster (almost ten times), with similar or better precision and present lower false detection rates, but their ROC curves suggest a clearly worse performance. They exhibit close performance only for datasets containing images of similar nature to those used for training. We can conclude that the training set does not encloses enough appearance information to build a robust facial feature detector

For future work we plan to combine the detectors and even the training sets. The effort must be done in terms of speeding up the process while keeping similar performance to those achieved with the training set A.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Innovation funds (TIN2008-06068).

References

1. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. on PAMI 19(7), 711–720 (1997)
2. Brubaker, S.C., Wu, J., Sun, J., Mullin, M.D., Rehg, J.M.: On the design of cascades of boosted ensembles for face detection. International Journal of Computer Vision 77, 65–86 (2008)
3. Carnegie Mellon University: CMU/VACS image database: Frontal face images (1999), http://vasc.ri.cmu.edu/idb/html/face/frontal_images/index.html (last accesed May 11, 2007)
4. Castrillón, M., Déniz, O., Hernández, D., Lorenzo, J.: A comparison of face and facial feature detectors based on the violajones general object detection framework. Machine Vision and Applications (2010) (in press)
5. Hjelmas, E., Low, B.K.: Face detection: A survey. Computer Vision and Image Understanding 83(3), 236–274 (2001), http://dx.doi.org/10.1006/cviu.2001.0921
6. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
7. Intel: Intel Open Source Computer Vision Library, v2.1 (April 2010), http://sourceforge.net/projects/opencvlibrary/ (last visited June 2010)
8. Jain, V., Learned-Miller., E.: Fddb: A benchmark for face detection in unconstrained settings. Tech. rep., University of Massachusetts, Amherst (2010)
9. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
10. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: IEEE ICIP 2002, vol. 1, pp. 900–903 (September 2002)
11. Reimondo, A.: Haar cascades repository (2007), http://alereimondo.no-ip.org/OpenCV/34 (last visited April 2010)
12. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(1), 23–38 (1998)
13. Schneiderman, H., Kanade, T.: A statistical method for 3d object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1746–1759 (2000)
14. Seo, N.: Tutorial: OpenCV haartraining (rapid object detection with a cascade of boosted classifiers based on haar-like features), http://note.sonots.com/SciSoftware/haartraining.html (last visited June 2010)
15. Sung, K.K., Poggio, T.: Example-based learning for view-based human face detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(1), 39–51 (1998)
16. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 151–173 (2004)
17. Yang, M.H., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002), http://dx.doi.org/10.1109/34.982883

Fast Classification in Incrementally Growing Spaces

Oscar Déniz-Suárez[1], Modesto Castrillón[2], Javier Lorenzo[2], Gloria Bueno[1], and Mario Hernández[2]

[1] E.T.S.I.Industriales, Universidad de Castilla-La Mancha
Avda. Camilo Jose Cela s/n, 13071 Ciudad Real, Spain
[2] Universidad de Las Palmas de Gran Canaria. Dpto. Informatica y Sistemas.
Edificio de Informatica, Campus de Tafira, 35017 Las Palmas, Spain

Abstract. The classification speed of state-of-the-art classifiers such as SVM is an important aspect to be considered for emerging applications and domains such as data mining and human-computer interaction. Usually, a test-time speed increase in SVMs is achieved by somehow reducing the number of support vectors, which allows a faster evaluation of the decision function. In this paper a novel approach is described for fast classification in a PCA+SVM scenario. In the proposed approach, classification of an unseen sample is performed incrementally in increasingly larger feature spaces. As soon as the classification confidence is above a threshold the process stops and the class label is retrieved. Easy samples will thus be classified using less features, thus producing a faster decision. Experiments in a gender recognition problem show that the method is by itself able to give good speed-error tradeoffs, and that it can also be used in conjunction with other SV-reduction algorithms to produce tradeoffs that are better than with either approach alone.

Keywords: gender recognition, Support Vector Machines, Principal Component Analysis, Eigenfaces.

1 Introduction

One of the most frequent classification systems encountered in research is the combination of PCA (Principal Component Analysis) and SVM (Support Vector Machines). PCA is frequently used because of its simplicity and relative effectiveness, while SVM have already demonstrated impressing classification capabilities. The two techniques have been used together for face recognition and verification, face detection, biosignal (i.e. EEG, ECG, EMG, CT scans...) classification, operations research, part inspection, biochemistry, anomaly detection, text categorization, medicine composition analysis, etc. For a comprehensive list of SVM applications the reader is referred to [1].

Despite the power of SVMs, they are orders of magnitude more costly at query-time than other popular machine learning alternatives such as decision trees and neural networks [2]. Classification speed is crucial for learning problems

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 305–312, 2011.

that use a large number of samples, like in emerging data mining applications. In some domains the amount of data available is growing at exponential rates, especially with the advent of global networks and the possibility of ubiquitous generation of data. Human-computer and human-robot interaction applications also need to produce fast responses, as for example in phoneme classification. Low computational complexity is also required for embedded and mobile systems, where available resources are rather limited.

Most of the research carried out in fast classification kernel machines has involved reducing the number of support vectors [3]. Such reduction can be achieved by approximating the discriminating hypersurface to a user-specified accuracy. In [4] the approach taken was to reduce the complexity of the generated hypothesis by excluding some training samples, specifically subsets of the support vectors obtained in the first place. In a similar fashion, [5] is based on stopping the evaluation of support vectors of the hypothesis when the confidence of the result (measured by the partial classification result) is above a threshold. This requires the support vectors of the hypothesis to be ordered by decreasing importance. In [6] pairs of close support vectors are iteratively substituted by a new one. Similarly, in [7] the decision function is simplified by removing support vectors that contribute less to the decision.

As shown above, most research in fast kernel machines has involved selecting subsets of support vectors (or training samples in general). This paper describes a framework for fast classification in PCA+SVM systems in which classification is performed incrementally in increasingly larger feature spaces. As soon as the classification confidence is above a threshold the process stops and the class label is retrieved. Fast classification is not achieved by using less support vectors, but by classifying in simpler spaces, which reduces the number of computations. Section 2 explains the common PCA+SVM setting encountered in supervised learning problems and describes the method proposed. Experimental results are shown in Section 3. Finally, the main conclusions and lines of future work are outlined.

2 Fast Classification in PCA+SVM Settings

PCA is often used to project input samples to a (generally lower dimensional) space where classification is carried out. This is specially useful when the input samples are images. Basically, PCA gives a set of orthogonal dimensions that maximize the variance of the input samples. In face recognition, this set is called *eigenfaces*, see [8]. Not all of these dimensions (eigenfaces) are useful for classification. Only the first n eigenfaces are appropriate for classification, with the last eigenfaces typically encoding noise.

When a test sample X is to be projected with PCA, the operation to perform is: $Y = XW$, where W is the transform matrix. When working with vectorized images in the rows of X, the columns of W are the eigenfaces. As mentioned above, usually only the first n columns of W are used in the multiplication. This is thought to avoid the noise of the last eigenfaces. What should be a good

value for n? There are reasons to believe that a large dimensional input space would be needed to separate difficult samples, while 'easy' samples could be separated in simpler spaces (i.e. lower value for n). In the method proposed, a different K value may be used for each specific test sample, instead of a fixed dimension n. *Easy* samples can be classified in a PCA space of a low number K of dimensions. The necessary number of dimensions to use will be ultimately given by the classifier output. For each test sample the system would classify it in a low dimensional space first. If the classifier output is large enough (i.e. above a fixed threshold) then classification will end and a class label will be retrieved. Otherwise the process should be repeated in a more informative space of a larger dimension, see Fig. 1.

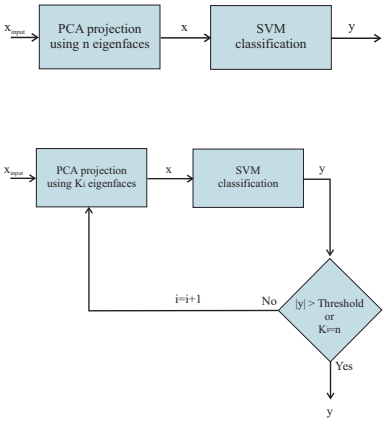


Fig. 1. Top: Typical PCA+SVM classification procedure for a test sample. Bottom: Fast PCA+SVM classification method for a test sample.

The loop of Figure 1 would have to be incremental in terms of computational cost. Otherwise there would not be any speed gain over the use of a fixed dimension. The PCA projection of the input sample can be done incrementally, since it is a matrix multiplication (see above).

It can be shown that SVM classification can be also made incremental in the input space dimension as long as the new dimensions at each step are orthogonal to the previous ones, which is the case when using PCA. Kernels typically used (like polynomial, RBF and sigmoid) are functions either of a dot product or a norm of samples. When classifying a sample, the cost of the kernel evaluations is therefore dependent on the space dimension. For a given input sample, let us suppose that classification has been already made in a space of dimension K_{i-1}. Therefore, we have already evaluated the kernel values $\kappa(\mathbf{x}, \mathbf{x_i})$. Let us suppose that we set to classify the same sample in a space of dimension $K_i > K_{i-1}$. Here the input and training samples can be respectively represented as $\mathbf{x} + \Delta\mathbf{x}$ and

$\mathbf{x_i} + \Delta\mathbf{x_i}$, where vectors $\mathbf{x}$ and $\mathbf{x_i}$ are augmented with zeros in order to have K_i components. $\Delta\mathbf{x}$ and $\Delta\mathbf{x_i}$ represent the values of the new $\Delta K = K_i - K_{i-1}$ dimensions, with the other components set to zero.

With the assumption imposed on the input space, in this space of K_i dimensions the following orthogonality relations hold: $\mathbf{x} \perp \Delta\mathbf{x}$, $\mathbf{x} \perp \Delta\mathbf{x_i}$, $\mathbf{x_i} \perp \Delta\mathbf{x}$, $\mathbf{x_i} \perp \Delta\mathbf{x_i}$. Using these orthogonality relations two cases are now possible:

- Dot product-based kernels (for example the polynomial kernel $\kappa(\mathbf{x}, \mathbf{x_i}) = (\mathbf{x} \cdot \mathbf{x_i} + 1)^p$):

$$(\mathbf{x} + \Delta\mathbf{x}) \cdot (\mathbf{x_i} + \Delta\mathbf{x_i}) = \mathbf{x} \cdot \mathbf{x_i} + \mathbf{x} \cdot \Delta\mathbf{x_i} + \Delta\mathbf{x} \cdot \mathbf{x_i} + \Delta\mathbf{x} \cdot \Delta\mathbf{x_i} = \\ = \mathbf{x} \cdot \mathbf{x_i} + \Delta\mathbf{x} \cdot \Delta\mathbf{x_i} \quad (1)$$

- Norm-based kernels (for example the RBF kernel: $\kappa(\mathbf{x}, \mathbf{x_i}) = \exp(-||\mathbf{x} - \mathbf{x_i}||^2/2p^2)$):

$$||(\mathbf{x} + \Delta\mathbf{x}) - (\mathbf{x_i} + \Delta\mathbf{x_i})||^2 = ||\mathbf{x} + \Delta\mathbf{x}||^2 + ||\mathbf{x_i} + \Delta\mathbf{x_i}||^2 - 2(\mathbf{x} + \Delta\mathbf{x}) \cdot (\mathbf{x_i} + \Delta\mathbf{x_i}) = \\ = ||\mathbf{x}||^2 + ||\Delta\mathbf{x}||^2 + 2\mathbf{x} \cdot \Delta\mathbf{x} + ||\mathbf{x_i}||^2 + ||\Delta\mathbf{x_i}||^2 + 2\mathbf{x_i} \cdot \Delta\mathbf{x_i} - 2(\mathbf{x} + \Delta\mathbf{x}) \cdot (\mathbf{x_i} + \Delta\mathbf{x_i}) = \\ = ||\mathbf{x}||^2 + ||\Delta\mathbf{x}||^2 + ||\mathbf{x_i}||^2 + ||\Delta\mathbf{x_i}||^2 - 2(\mathbf{x} \cdot \mathbf{x_i} + \mathbf{x} \cdot \Delta\mathbf{x_i} + \Delta\mathbf{x} \cdot \mathbf{x_i} + \Delta\mathbf{x} \cdot \Delta\mathbf{x_i}) = \\ = ||\mathbf{x} - \mathbf{x_i}||^2 + ||\Delta\mathbf{x}||^2 + ||\Delta\mathbf{x_i}||^2 - 2\Delta\mathbf{x} \cdot \Delta\mathbf{x_i} \quad (2)$$

It can be seen that the computations are based on the dot product or norm of the previous step plus some terms that can be computed with a constant cost proportional to ΔK. Thus, in both cases the computation can be done incrementally.

Note that the training cost of the proposed method is the same as in a non-incremental classifier, only one training stage is carried out using a space of whatever dimension n. Once we have a trained classifier, the proposed method only works at test time, where we have the incremental classification of samples.

3 Experiments

Since the speed gain in the proposed method is based on classifier confidence, we will have a trade-off between classification speed and error. In this respect, the main performance indicator that will be used here is the error-speedup curve, which represents test error as a function of classification speed gains. This curve is obtained by varying the classifier confidence threshold (see Figure 1), with values ranging from 0 to 1. An RBF kernel was used in all the experiments.

The question arises whether the proposed dimensionality reduction strategy can be compared with SV-reduction. Note that there are cases in which one reduction strategy will always be superior to the other and vice versa. For the dimensionality reduction approach the results will depend on the number n of dimensions used (i.e. the size of the feature space). On the other hand, the performance of SV-reduction methods depends on the number of support vectors which in turn depends on the parameters used for training the classifier (i.e. the kernel parameter 'p'). The best values for these parameters depend on the

problem at hand (and also on the number of training samples available). For large values of n, for example, the proposed dimensionality reduction method should give better error-speedup curves than the SV reduction method. For small values of n the reverse will be true. For these reasons, a direct comparison is not appropriate. Instead, we focused on combining the two strategies to test whether better net results can be obtained.

The combination implies progressively reducing both the number of support vectors and dimensions, following some order (i.e. choosing at each step between SV-reduction or dimensionality reduction). Searching for the ordering that gives the optimal error-speedup curve is not feasible, since there is a factorial number of orderings.

Assuming independence between both reduction methods, approximations can be obtained. In our case, a simple greedy search was carried out in the validation set. The search involves choosing between reducing the dimension or reducing the number of support vectors, at each step of the curve. The selection is made according to the error decrease in the validation set produced by each option.

The proposed strategy was used in conjunction with the SV-reduction method described in [5], in which classification speed is improved by using only the most important support vectors in the classification function evaluation. In order to achieve this, the support vectors are ordered by the absolute value of the associated coefficient. With that algorithm, important computational savings can be achieved without significant degradation in terms of recognition accuracy.

A gender recognition scenario was used in the experiments, using the typical PCA+SVM combination. A number of face images were captured in our laboratory. These included talking and changes in facial expression, illumination and pose. A total of 7256 male+7567 female images were gathered, and later normalized in rotation and size to 39x43 pixels. In each run of the experiment, 300 of these images were randomly selected and randomly partitioned in a training set of 120 images (60+60), a validation set of 80 images (40+40) and a test set of 100 images (50+50).

PCA was previously applied over an independent set of 4000 face images taken from the Internet. The eyes in each image were located manually and then the image was normalized to 39x43. PCA was computed over this set of 4000 normalized images, retaining a number n of coefficients, see Figure 2. The collected images were all projected onto this PCA space previous to training and classifying.

Even though our goal in this work was not to obtain better absolute recognition values, we wanted to test the algorithms in an independent database, a subset of frontal faces of the FERET [9] data set was also considered. In this case the working set was made up of a total of 177 male+177 female faces, normalized to 52x60 pixels. In each experiment a random ordering of the samples of each class was performed. PCA was applied to the first 144 of them (77+77). The training set had 120 samples (60+60), the validation set 40 (20+20) and the test set the other 40 (20+20).

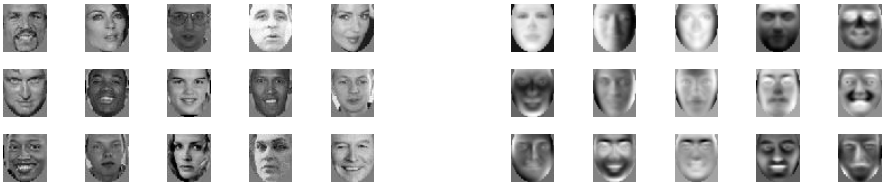


Fig. 2. Left: some Internet faces, as detected and normalized. Right: the first 36 eigenfaces obtained with PCA.

The experiments described here show the effects of the combination in two different cases: a) dimensionality reduction performing better than SV-reduction and b) SV-reduction performing better than dimensionality reduction. These cases were achieved by adjusting the parameter p of the support vector classifier and the value of n. The speedups were calculated as 100 minus the percentage of numerical operations used for classification, with respect to the initial case of dimensionality n and number of support vectors. Figure 3 shows the performance for case a), i.e. dimensionality-reduction method better than SV-reduction.

Note that for large values of n (as was the case in Figure 3) the dimensionality reduction curve is more horizontal, which makes it difficult to obtain significant improvements with the combination. This occurs because with the validation set the greedy algorithm could choose to reduce support vectors at a given point when in fact the best option is to keep on reducing dimensionality (thus keeping the error constant most of the time). This causes the performance of the combination to be worse than with either of the two methods, especially if the validation set is small. Since we have a validation set available, in such cases it

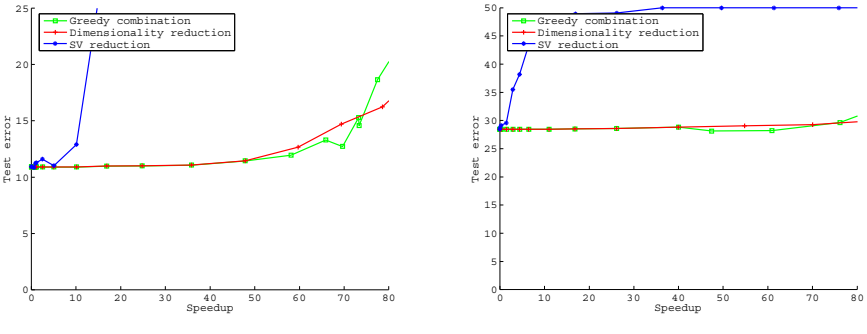


Fig. 3. Left: Error-speedup curves for the three methods considered, using our laboratory database. 40 runs, random distributions of the samples in training, validation and test sets. Kernel parameter $p = 3000$. The (initial) dimensionality n was calculated as that which accounted for a 90% of the total variance. Right: Error-speedup curves for the three methods considered, using the FERET database. Same conditions except for kernel parameter $p = 10000$.

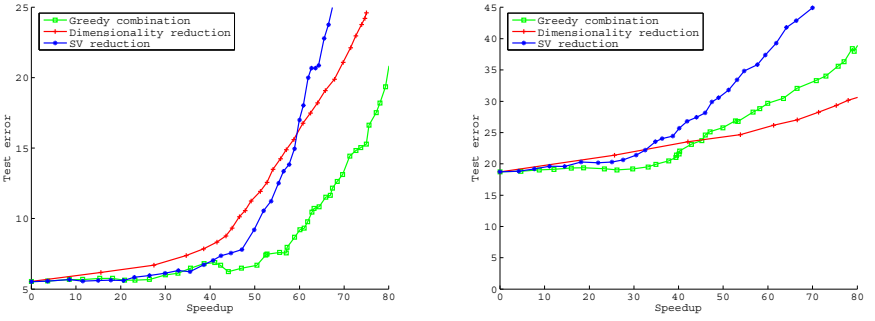


Fig. 4. Left: Error-speedup curves for the three methods considered, using our laboratory database. 25 runs, random distributions of the samples in training, validation and test sets. Kernel parameter $p = 500$. The (initial) dimensionality n was 4. Right: Error-speedup curves for the three methods considered, using the FERET database. 50 runs, random distributions of the samples in training, validation and test sets. Kernel parameter $p = 1500$. The (initial) dimensionality n was 25.

may be useful to set a threshold in the greedy search so that SV-reduction is used only when large errors begin to appear with dimensionality reduction. Alternatively, SV-reduction could be made to proceed only after the speedup gain has reached a given point, which can be estimated (manually) with the validation set. The latter option was used in Figure 3, where SV-reduction only acted after a speedup of 60% and 40% was reached using dimensionality reduction.

Figure 4 shows the performances for case b), i.e. SV-reduction better than dimensionality-reduction. For the FERET images it was very difficult to find a set of parameter values that made the SV-reduction method be clearly better than dimensionality reduction. We postulate that this was due to the fact that this data set was considerably more difficult (the images were larger, the PCA space was obtained with fewer samples, many races were present, more significant illumination variations, ...), which would have made the obtained support vectors more critical for classification. Still, the figure shows how the greedy algorithm allows to obtain an improvement for speedups between 10-40%, although after that point the performance of the combination obviously turns worse than with dimensionality reduction alone. Overall, the results shown above suggest that even with a simple greedy combination a better net performance can be achieved. With more computational effort better combinations could be used that take advantage of the (in)dependence between feature space size and classifier size.

4 Conclusions

The test speed of state-of-the-art classifiers such as SVM is an important aspect to be considered for certain applications. Usually, the reduction in classification complexity in SVMs is achieved by reducing the number of support vectors used

in the decision function. In this paper a novel approach has been described in which the computational reduction is achieved by classifying each sample with the minimum number of features necessary (note that the typical setting is to use a fixed dimension for the input space). Experiments in a gender recognition problem show that the method is by itself able to give good speed-error trade-offs, and that it can also be used in conjunction with support vector-reduction algorithms to produce trade-offs that are better than with either approach alone.

Acknowledgments

This work was partially supported by project PII2I09-0043-3364 of Castilla-La Mancha Regional Government and the Spanish Ministry of Science and Innovation funds (TIN2008-06068).

References

1. Guyon, I.: SVM application list (2010), http://www.clopinet.com/isabelle/Projects/SVM/applist.html
2. Decoste, D., Mazzoni, D.: Fast query-optimized kernel machine classification via incremental approximate nearest support vectors. In: International Conference on Machine Learning, pp. 115–122 (2003)
3. Fehr, J., Zapien, K., Burkhardt, H.: Fast support vector machine classification of very large datasets. In: Proceedings of the GfKl Conference, Data Analysis, Machine Learning, and Applications. LNCS. Springer, University of Freiburg, Germany (2007)
4. Zhana, Y., Shen, D.: Design efficient support vector machine for fast classification. Pattern Recognition 38(1), 157–161 (2005)
5. Arenas-García, J., Gómez-Verdejo, V., Figueiras-Vidal, A.R.: Fast evaluation of neural networks via confidence rating. Neurocomput 70(16-18), 2775–2782 (2007), http://dx.doi.org/10.1016/j.neucom.2006.04.014
6. Nguyen, D., Ho, T.: An efficient method for simplifying support vector machines. In: Procs. of the 22nd Int. Conf. on Machine Learning, pp. 617–624 (2005)
7. Guo, J., Takahashi, N., Nishi, T.: An efficient method for simplifying decision functions of support vector machines. IEICE Transactions 89-A(10), 2795–2802 (2006)
8. Turk, M.A., Pentland, A.: Eigenfaces for Recognition. Cognitive Neuroscience 3(1), 71–86 (1991), ftp://whitechapel.media.mit.edu/pub/images/
9. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. TPAMI 22(10), 1090–1104 (2000)

Null Space Based Image Recognition Using Incremental Eigendecomposition*

Katerine Diaz-Chito, Francesc J. Ferri, and Wladimiro Díaz-Villanueva

Dept. Informàtica, Universitat de València, Spain
{Katerine.Diaz,Francesc.Ferri,Wladimiro.Diaz}@uv.es

Abstract. An incremental approach to the discriminative common vector (DCV) method for image recognition is considered. Discriminative projections are tackled in the particular context in which new training data becomes available and learned subspaces may need continuous updating. Starting from incremental eigendecomposition of scatter matrices, an efficient updating rule based on projections and orthogonalization is given. The corresponding algorithm has been empirically assessed and compared to its batch counterpart. The same good properties and performance results of the original method are kept but with a dramatic decrease in the computation needed.

Keywords: Incremental learning, discriminative common vector, subespaces, discriminative projections.

1 Introduction

Representing images in subspaces in order to reduce computational burden and also to improve their discriminability constitutes a very general goal in many image recognition practical problems [1, 2]. When applied to very large images, these methods imply relatively high time and space requirements as they usually need non trivial numerical operations on large matrices computed from a previously given training set.

In particular dynamic or highly interactive scenarios, image recognition algorithms may require retraining as new information becomes available. New (labeled) data may be then added to the previous training set so that the original (batch) algorithm can be used but this involves prohibitive computational burden for most practical applications. Instead, incremental subspace learning algorithms have been proposed for basic algorithms such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in order to alleviate these requirements while keeping most of the performance properties of its batch counterpart[3–7].

* Work partially funded by FEDER and Spanish and Valencian Governments through projects TIN2009-14205-C04-03, ACOMP/2010/287, GV/2010/086 and Consolider Ingenio 2010 CSD07-00018.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 313–320, 2011.

Subspace learning methods based on Null Space or Discriminative Common Vectors (DCV) have been recently proposed for face recognition [2]. The rationale behind DCV is close to LDA but is particularly appealing because its good performance behavior and flexibility of implementation specially in the case of very large dimensionalities [2, 8].

In this paper, incremental formulations corresponding to basic (batch) implementations of the DCV method are given. The derived algorithms follow previously published ideas about (incrementally) modifying subspaces [9, 10] but in the particular context of Null Spaces associated to the within-class scatter matrix. A preliminar implementation of this idea, has already been presented in [11] along with limited experimentation. In the present work, both subspace projections and explicit common vectors are efficiently recomputed allowing the application of these algorithms in a wide range of situations in interactive and dynamic problems. Extensive experimentation for different real databases and using several parameter settings has been included to demonstrate the benefits of the proposed approach.

2 Image Recognition through Discriminant Common Vectors

The DCV method was proposed for face recognition problems in which input data dimension is much higher than the training set size [2]. This case is referred to in the specialized literature as the small sample size case [1]. In particular, the method looks for a linear projection that maximizes class separability by considering a criterion very similar to the one used for LDA-like algorithms and also uses the within-class scatter matrix. In short, the method consists of constructing a linear mapping onto the null space of this matrix in which all training data gets collapsed into the so-called *discriminant common vectors*. Classification of new data can be then accomplished by first projecting it and then measuring similarity to DCVs of each class with an appropriate distance measure.

Let $\mathcal{X} \in \mathbb{R}^{d \times M}$ be a given training set consisting of M d-dimensional (column) vector-shaped images, $x_j^i \in \mathbb{R}^d$, where $i = 1, \ldots, M_j$ refers to images of any of the c given classes, $j = 1, \ldots, c$ and $M = \sum_{j=1}^{c} M_j$. Let S_X^w be their corresponding within-class scatter matrix and let x_j be the j-th class mean vector from $\mathcal{X}$.

Let $U \in \mathbb{R}^{d \times r}$ and $\overline{U} \in \mathbb{R}^{d \times n}$ be matrices formed with the eigenvectors corresponding to non zero and zero eigenvalues, computed from the eigenvalue decomposition (EVD) of S_X^w where r and $n = d - r$ are the dimensions of its range and null spaces, respectively. The range and null spaces generated by U and $\overline{U}$, respectively, are complementary subspaces, so that their direct sum is all $\mathbb{R}^d$. Each sample $x_j^i \in \mathbb{R}^d$ admits a unique decomposition of the form $x_j^i = UU^T x_j^i + \overline{U}\,\overline{U}^T x_j^i$, where UU^T and $\overline{U}\,\overline{U}^T$ are orthogonal projection operators onto range and null spaces, respectively.

The j-th class common vector can be computed as the orthogonal projection of the j-th class mean vector onto this null space, $\overline{U}\,\overline{U}^T x_j$ or, equivalently as:

$$x_{com}^j = \overline{U}\,\overline{U}^T x_j = x_j - UU^T x_j \tag{1}$$

In both expressions, the mean vector x_j may in fact be substituted by any other j-class training vector [2]. Note that it is much easier and convenient to use U rather than $\overline{U}$, partially because in the context of image recognition usually $r \ll n$. These d-dimensional common vectors constitute a set of size c to which standard PCA can be applied. The combination of this with the previous mapping gives rise to a linear mapping onto a reduced space, $\Theta \in \mathbb{R}^{d\times(c-1)}$. Reduced dimensionality discriminative common vectors (DCVs) can be then computed as $\Theta^T x_j$. When new (test) data, x, is to be classified, it can get projected as $\Theta^T x$ and then appropriately compared to $\Theta^T x_j$ in order to be recognized.

Even after several improvements that can be applied [2, 12], the computational burden associated to this procedure is dominated by the eigendecomposition of S_X^w and leads to a cost in $O(\ell M^3 + dM^2)$, where ℓ is a constant related to the iterative methods used for EVD.

3 Incremental Null Space Characterization

Let $\mathcal{X}$ be as defined in Section 2 and let $\mathcal{Y} \in \mathbb{R}^{d\times N}$ be the currently available training set (new incoming data) consisting of N_j vectors from each of the c classes. And let $\mathcal{Z} = [\mathcal{X}\ \ \mathcal{Y}] \in \mathbb{R}^{d\times(M+N)}$.

Basic DCV method on $\mathcal{Z}$ would require eigendecomposing S_Z^w first in order to use Equation 1 to compute the common vectors and go forward. To avoid this, S_Z^w must be decomposed into simpler parts that can be put in terms of eigendecompositions $S_X^w = U\Lambda U^T$ (from the previous iteration) and $S_Y^w = V\Delta V^T$ (that can be done straightaway as $N \ll M$) along with their corresponding mean vectors x_j and y_j. From the standard within-class scatter matrix definition we can arrive at

$$S_Z^w = S_X^w + S_Y^w + S_{XY} = U\Lambda U^T + V\Delta V^T + WW^T \tag{2}$$

which could be seen as a generalization of the decomposition in [9] for the overall scatter matrix. In this expression, $W \in \mathbb{R}^{d\times c}$ is a (non orthonormal) generator set of the subspace spanned by the difference mean vectors whose columns are defined from $\mathcal{X}$ and $\mathcal{Y}$ as $\sqrt{\frac{M_j N_j}{(M_j+N_j)}}(x_j - y_j)$ for each class j.

In the particular case in which only one sample per class is added (that is, $N_j = 1$ for all j), the above expression becomes simpler as $S_Y^w = 0$. The problem can be posed then as updating the basis of the range subspace, U, by adding new (orthogonal) components from subspaces represented by V (if it is not empty) and W. This situation is illustrated in Figure 1.

As only directions complementary to the ones in U are needed, all column vectors in V and W must be projected onto its complementary as $V - UU^T V$ and $W - UU^T W$, respectively. These (at most $(N - c) + c$) projected vectors

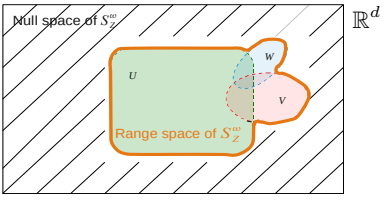


Fig. 1. Range subspaces corresponding to data and basis involved in the incremental updating of DCV

can be now orthonormalized and then simply added to U in order to obtain the sought update, $[U\ v]$ where v is the result of the orthonormalization procedure. The discriminative common vectors, z_{com}^j, can be also incrementally updated as

$$z_{com}^j = (I - [U\ v][U\ v]^T)z_j = (I - UU^T - vv^T)z_j = x_{com}^j - vv^T x_j$$

Note that the basis $[U\ v]$ is in general different to the one that will be obtained from the eigendecomposition, U', up to a rotation, R. Therefore, we can write

$$S_Z^w = U'\Lambda' U'^T = S_X^w + S_Y^w + S_{XY} = [U\ v]R\ \Lambda' R^T [U\ v]^T$$

and then multiplying by the left and right inverses of the new basis to obtain

$$R\Lambda' R^T = \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} U^T V \Delta V^T U & U^T V \Delta V^T v \\ v^T V \Delta V^T U & v^T V \Delta V^T v \end{bmatrix} + \begin{bmatrix} U^T W W^T U & U^T W W^T v \\ v^T W W^T U & v^T W W^T v \end{bmatrix}$$

The right hand side of the above equation can be easily computed from the incremental data and, consequently, the rotation, if needed, can be obtained from the corresponding diagonalization.

To implement the plain DCV method, any orthogonal basis for the corresponding range subspace will suffice and the above rotation need not be computed. In this particular case, the full update of this basis can be done in $O(\ell N^3 + dMN)$ time. This constitutes an improvement with respect to the corresponding basic algorithm which would imply a computation time in $O(\ell(M+N)^3 + d(M+N)^2)$.

In the asymptotic expression for the computation time corresponding to the incremental algorithm, the cubic term will vanish if $N_j = 1$ $(N = c)$ and would go up to $(M + N)^3$ if R would need to be computed.

4 Experiments and Discussion

An in depth comparative experimentation has been carried out using a wide range of the involved parameters of the incremental algorithm over 4 different publicly available databases. In all cases, images were previously normalized in intensity, scaled and aligned using the position of eyes and mouth. Figure 2 shows some sample images and their basic characteristics as dimensionality

name	**size**	c	**size**/c	M_j	N_j	**type**
AR [13]	222×299	50	14	3 ... 12	1, 2, 3, 4	expression & lighting
Jaffe [14]	256×256	10	20	4 ... 16	1, 2, 3, 4	expression
Cmu-Pie [15]	120×160	68	56	11 ... 47	1, 2, 3, 4	pose & lighting
Coil-20 [16]	128×128	20	72	13 ... 62	1, 2, 3, 4	pose

Fig. 2. Examples and details about the four databases considered in the experiments

(image size), number of classes (c), number of objects (per class), and type of variability. More details about these databases can be found in the corresponding references shown also in the Figure 2.

In particular, an experimental setup in which more training data becomes available to the algorithm has been designed. For each database, the available data has been split into 3 disjoint sets that will be referred to initial, incremental and test sets, respectively. The initial dataset contains about 15% of the available data (the number of samples per class for each database is the smallest number in the M_j column in Figure 2) and the test set is formed by approximately 20%. The remaining incremental set is divided into blocks of N_j samples per class (ranging from 1 to 4) that will be made progressively available for the algorithm in each experiment.

The learning process was repeated 10 times for each database, where training and test sets have been randomly permuted after each shift to remove any kind of dependence on the order in which data is made available to the algorithm. The results presented correspond then to an average across the whole database along with corresponding standard deviations.

At each iteration, N_j new images per class are available. The incremental algorithm is then run using the previous M_j images per class. The basic DCV algorithm is also run from scratch using the current $M + N$ images. In this way, M values range approximately from 15% to 80% while the value of N has been fixed for each database and experiment according to its global size.

The accuracy of the minimum distance classifier using DCVs in the projected subspace has been considered [2]. According to expectation, accuracy results obtained do not lead to any significant difference between incremental and batch approaches. Moreover, the accuracy rate of the incremental algorithm does not

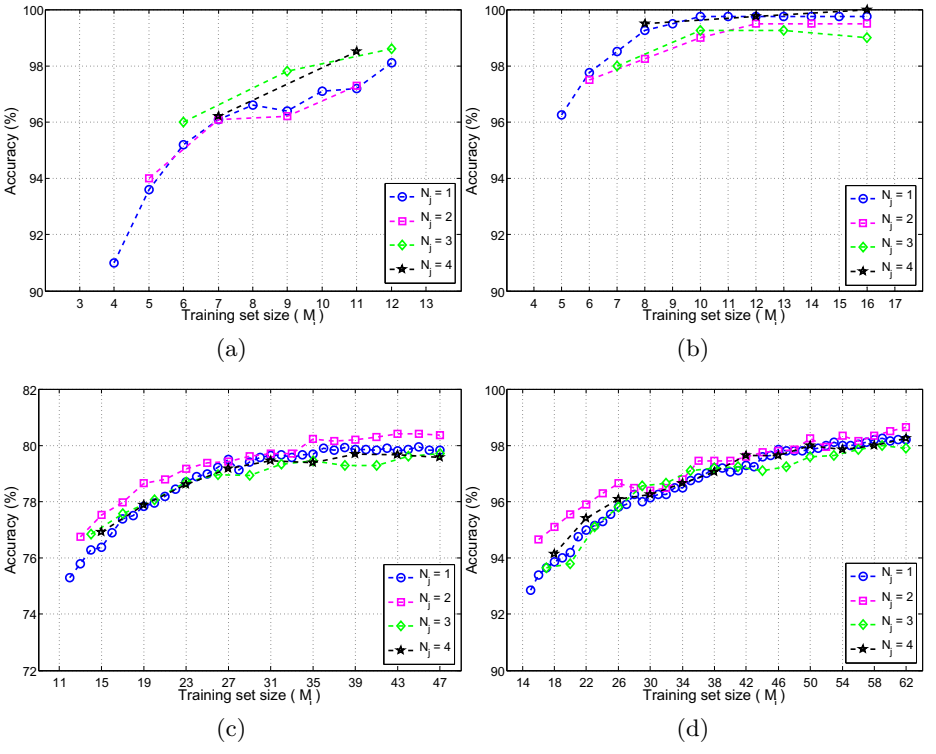


Fig. 3. Averaged accuracy vs accumulated training set size for both algorithms considered on the 4 databases, (a) AR, (b) Jaffe, (c) Cmu-Pie and (d) Coil-20

present significant changes for the different incremental training set sizes (shown in Figure 3), where the small differences are due to the randomness of the training set at each iteration.

Figure 4 shows the relative CPU time for each incremental algorithm configuration at each iteration with regard to the corresponding basic (batch) DCV algorithm.

The general trend illustrated in the figure shows that the relative CPU time decreases with M for a fixed N. This tendency follows approximately the theoretical $O(1/M)$ that comes from the corresponding asymptotic results for fixed N and d. So when the number of total training samples is greater than the number of samples added incrementally, the differences in computational cost are more significant.

When the size of the incremental training set is about the size of the total training, the computational cost of incremental algorithm is not comparable with respect to the batch algorithm, as is the case of AR and Jaffe database, for values of $M \approx N$.

In the case $M >> N$, several interesting facts can be put forward from the results obtained. First, incremental approaches for conveniently small N values

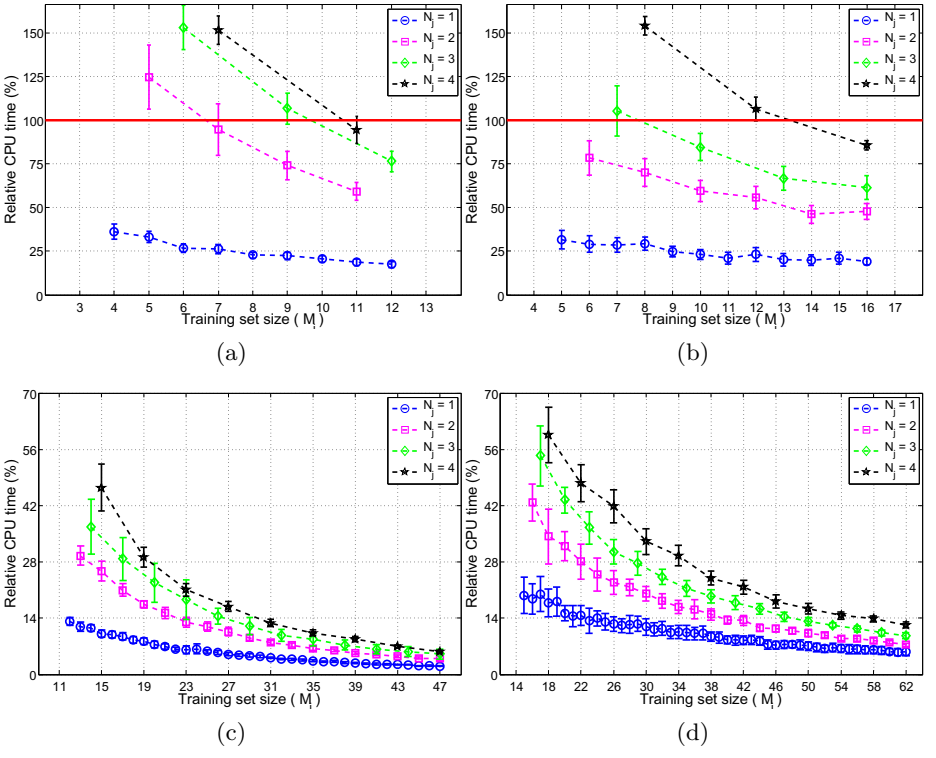


Fig. 4. Relative CPU time of the incremental method with regard to corresponding batch DCV method on the databases, (a) AR, (b) Jaffe, (c) Cmu-Pie and (d) Coil-20

will cut down computation to very small percentages of the computation needed by the whole retraining using the batch algorithm. Second, the computation decreases as N becomes smaller. And this decrease with N compensates even taking into account that you will need a higher number of updates as N gets smaller.

5 Concluding Remarks and Further Work

An incremental algorithm to compute DCVs and corresponding subspaces has been considered. The algorithm use incremental eigendecomposition and orthonormalization as in the original (batch) algorithm. Dramatic computational savings are observed while performance behavior of DCV is preserved in the experimentation carried out on four publicly available image databases.

Further work is driven towards the implementation of more general common vector based subspace algorithms, using extended null space and kernels, in an incremental way along with extending the experimentation to other, more challenging and truly dynamic scenarios.

References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
2. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 27(1), 4–13 (2005)
3. Murakami, H., Kumar, B.: Efficient calculation of primary images from a set of images. IEEE Trans. Patt. Analysis and Machine Intell. 4(5), 511–515 (1982)
4. Chandrasekaran, S., Manjunath, B., Wang, Y., Winkler, J., Zhang, H.: An eigenspace update algorithm for image analysis. Graphical Models and Image Processing 59(5), 321–332 (1997)
5. Hall, P.M., Marshall, D., Martin, R.R.: Incremental eigenanalysis for classification. In: British Machine Vision Conference, pp. 286–295 (1998)
6. Ozawa, S., Toh, S.L., Abe, S., Pang, S., Kasabov, N.: Incremental learning of feature space and classifier for face recognition. Neural Netw. 18(5-6), 575–584 (2005)
7. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vision 77(1-3), 125–141 (2008)
8. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: A novel method for face recognition. In: Proceedings of the IEEE 12th Signal Processing and Communications Applications Conference, pp. 579–582 (2004)
9. Hall, P., Marshall, D., Martin, R.: Merging and splitting eigenspace models. IEEE Trans on Pattern Analysis and Machine Intelligence 22(9), 1042–1049 (2000)
10. Hall, P., Marshall, D., Martin, R.: Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. Image and Vision Computing 20(13-14), 1009–1016 (2002)
11. Díaz-Chito, K., Ferri, F.J., Díaz-Villanueva, W.: Image recognition through incremental discriminative common vectors. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010, Part II. LNCS, vol. 6475, pp. 304–311. Springer, Heidelberg (2010)
12. Gulmezoglu, M., Dzhafarov, V., Keskin, M., Barkana, A.: A novel approach to isolated word recognition. IEEE Trans. Speech and Audio Processing 7(6), 618–620 (1999)
13. Martinez, A., Benavente, R.: The ar face database. Technical Report 24, Computer Vision Center CVC (1998)
14. Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998)
15. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression (PIE) database. In: Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition (2002)
16. Nene, S., Nayar, S.K., Murase, H.: Columbia object image library (coil-20). Technical report (1996)

Multi-sensor People Counting*

Daniel Hernández-Sosa, Modesto Castrillón-Santana,
and Javier Lorenzo-Navarro

SIANI
Universidad de Las Palmas de Gran Canaria

Abstract. An accurate estimation of the number of people entering / leaving a controlled area is an interesting capability for automatic surveillance systems. Potential applications where this technology can be applied include those related to security, safety, energy saving or fraud control. In this paper we present a novel configuration of a multi-sensor system combining both visual and range data specially suited for troublesome scenarios such as public transportation. The approach applies probabilistic estimation filters on raw sensor data to create intermediate level hypothesis that are later fused using a certainty-based integration stage. Promising results have been obtained in several tests performed on a realistic test bed scenario under variable lightning conditions.

Keywords: people counting, EKF, MHI, laser sensors.

1 Introduction

Automatic surveillance systems are becoming more and more frequent nowadays. People counting constitutes a relevant component of those for many applications. For example, the number of passengers getting in/out of a public transport is necessary for control and management. In pubs and discos the evacuation protocols are designed according to the building capacity and it must not be exceeded. Another example is the presence control for implementing energy saving politics.

Two main technologies have been used to solve the people counting problem: Computer Vision and light beams. On one hand, Computer Vision techniques has been successfully applied to more and more areas in the recent years. This process is favored by the introduction of lower-cost higher-performance hardware and the improvements in the reliability of detection methods. On the other hand, laser sensors have also evolved in the same directions, so smaller and lighter units are available at a reasonable cost vs. precision ratio.

1.1 Computer Vision Methods

In the literature, we can find many examples of Computer Vision based systems with cameras located both in zenithal and non zenithal position. However for some applications where privacy preserving is a crucial matter, the use of vision-based systems with non zenithal cameras is not permitted.

* This work was partially supported by the Spanish MICIIN funds (TIN2008-06068).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 321–328, 2011.

Chant et al. [3] proposed a method based on analysing a crowd and making use of mixture of dynamic textures to segment the crowd into different directions; after a perspective correction, some features are computed on each segment and with a Gaussian Process the number of people per segment is obtained. Bozzoli et al. [2] introduced a method for people counting in crowded environments as bus or train gates. The proposal is based on the computation of a running average-like background model applied to edge images in order to avoid the influence of sudden lighting condition changes. Foreground edges are filtered and with the remaining one the optical flow image is computed. Finally each movement vector is assigned to a segment and all the movement vectors assigned to the same segment can be used to estimate the people passing in each direction.

Vision based techniques are well suited for large, wide and open areas, like train platforms or commercial areas besides gates or corridors, provided that lightning conditions are kept under control.

1.2 Range Laser Methods

Katabira et al. [5] proposed a system based on a sensor mounted on the ceiling of a passage. From the range data acquired by the sensor human shapes can be obtained by transforming the data to $X - Z$ plane. The proposed method detects a passing pedestrian when a prominent object is detected.

Mathews and Poigné [8] introduced a system based on a set of passive infrared beacons. The detection of people is done with an Echo State Network which is trained with a set of motion patterns obtained with a simulator.

Light beams based systems have the advantage of privacy preserving, and are best suited for small areas.

1.3 Hybrid Methods

In order to come together the advantages of light beam and vision based systems, some authors have proposed to fusion laser and camera data [9].

Gwang et al. [7] make use of a laser beam as a structured light source. In this way, 3D estimation can be done in an area by means of the integration of consecutive images. When people cross the area, the obtained pattern allows to count the number of people and also the direction of the movement.

Cui et al. [4] describe a method that fuses data from a laser and a visual tracker. The laser module is based on the integration of several laser readings to detect pair of legs and later tracked using a Kalman filter to estimate the position, velocity and acceleration of both feet. A calibrated camera allows to perform visual tracking with color information which feed a mean-shift tracker. Finally, the results of both tracking process are fused with a Bayesian approach.

1.4 The Proposal

In this paper, we propose a fast processing multi-sensor method for counting people getting in/out through a controlled area, using low-cost infrastructure

and preserving privacy. The system is specially well suited for troublesome scenarios with space limitations and changing lightning conditions, such as public transportation applications. Laser and visual based detectors run asynchronously generating hypothesis of crossing people. An upper level combines these hypothesis in order to accept or reject them. The laser process basically extracts relevant peaks from a dynamically generated 3D surface, while the vision process makes use of motion history images to obtain direction and location of people.

The paper is organized as follows: Section 2 gives a brief description of the system. Section 3 presents the results achieved in the experiments. Finally, in the conclusions, some remarks and future work are presented.

2 System Description

The main purpose of our system is to count the number of persons leaving and entering a space. For this work we have considered a specially challenging problem, the monitoring of access to public transportation. In this scenario, an automatic detection system must cope with adverse factors such as severe space limitations and varying lightning conditions. Additionally, the global system cost should be kept reasonably low and subject's privacy should be guaranteed.

The combination of the aforementioned factors has implications on the processing system, as low cost hardware translates into poor data quality and slow acquisition rate. For example, depending on people crossing speed a basic laser sensor can only obtain 3 to 4 scans per individual. Also, height limitation generates important occlusions when a tall person enters/leaves the controlled area, both in a camera or laser data. Besides, due to normally under-illuminated conditions, a standard camera auto-iris is generally wide open, making the depth focus thinner and producing blurring while adjusting to different height people.

The proposed counting people system is composed of a standard webcam and a low cost laser based range sensor. This seems to be an interesting configuration, as lasers provide precise range data but on a small area, while cameras cover a wider area but with a worse signal/noise ratio. Unlike previous works based on fusion of camera and laser readings [4,9], the range sensor is placed zenithally next to the camera. This configuration is better suited for narrow areas as public transports where the horizontal configuration of the laser is not recommended due to maintenance problems. Additionally, the zenithal location of the camera avoids the privacy matter because faces are not grabbed. The use of low cost sensors allows for a wider economically affordable deployment.

The software architecture is based on a fast pre-attentive visual motion detection and range processing stage, an intermediate data fusion and filtering layer and a final certainty-based decision module. In a previous phase, both laser and camera need to be calibrated, and a region of interest is defined for each sensor. As a result of this calibration process, two coordinate transformation matrices, M_l and M_c, are obtained for laser and camera respectively.

2.1 Visual Processing

Two main elements are involved in the visual processing: motion detection and data filtering. The detection uses a motion-energy/motion-history framework to identify module and direction of displacement on images. The data filtering uses an Extended Kalman Filtering (EKF) estimator to integrate motion measures.

Motion detection
Temporal templates have been used as a view-specific representation of movement in the last decade [1]. Image differencing is adequate for some applications, but in our case, as we need to identify if an individual is entering or leaving the room/space, the information contained in a set of consecutive frames is more useful. Motion templates seem to be applicable to this task, and have been used in the past for layering successively the silhouettes obtained after differencing one frame with the previous one.

The motion-energy image (MEI) is a binary cumulative image based on image binary differencing ($D(x, y, t)$) indicating where a motion takes place. Considering a cumulative interval τ, we have the following expression for the MEI image:

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i).$$

On the other hand, the motion-history image (MHI) indicates also how the motion was produced using each pixel intensity to represent the time elapsed since the motion occurred. In our implementation, we have considered a scalar image, where brighter pixels (max value v_{max}) refer to most recent moving pixels, decaying in intensity (v_{dec} factor) when no difference is present (see Figure 1).

MEI and MHI have been frequently used for gesture recognition [1]. Recently those representations have also been combined with texture descriptors to improve the recognition robustness [6]. However, in our approach, the gestures to recognize are simpler, but different situations can be presented to the sensors due to the various behavioral possibilities that can take place in a door when multiple people are present.

Data filtering
Camera frames are processed to extract motion blobs b_k and filtered out comparing blobs area with a minimum threshold to discard less significant ones ($area(b_k) < area_{min}$). As the camera does not provide depth information, several height possible values (z_1, ..., z_{n_h}) are tested in parallel, using a function that back-projects the blob center coordinates into the corresponding world coordinates via the matrix calibration camera: $b3D_{k,j} = MapXYZ(b_k, M_c, z_j)$ for $j = 1 \dots n_h$.

The set of $b3D_{k,j}$ are used to generate camera-based hypothesis for object trajectories $OTC_{c,j}(t)$, using an EKF framework. The object hypothesis are updated on the basis of the $k{-}th$ detected blob b_k on the current frame, according to the following rule:

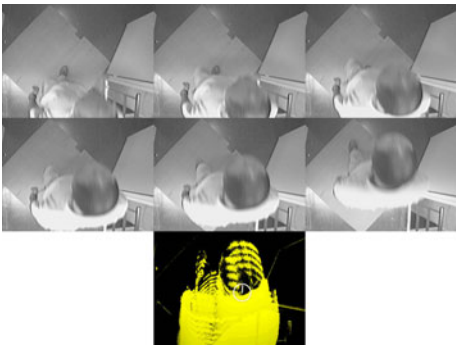


Fig. 1. MHI example

$$\begin{cases} \text{if } \min_{i=1\ldots n_c} Dist3D(k,i) < \mu_c & \text{EKF update(all j)}, \\ & OTC_{i,j}(t) \\ \text{otherwise} & n_c = n_c + 1, \\ & \text{EKF init(all j)}, \\ & OTC_{n_c,j}(t) = b3D_{k,j} \end{cases}$$

where $Dist3D(k,i) = \sum_{j=1\ldots n_h} \|b3D_{k,j} - OTC_{i,j}(t-1)\|$, n_c is the number of current active objects and μ_c is a distance threshold.

EKF filters operate on a three state vector representing the objects X and X coordinates and the angle of motion. See image on Figure 2 for an example of blob detection and the corresponding EKF trajectory estimation.

The visual filter keeps integrating data until an object trajectory is detected to intersect the door line, activating then a verification test including the analysis of trajectory length and filter covariance ellipses. On success, the object hypothesis is added to a set H_c of camera-based hypothesis with its spatial and temporal references.

2.2 Laser Scan Processing

Laser sensors are specially convenient to this problem due to their precision and relative invariance to lightning conditions. In our approach laser readings are integrated over time to generate a kind of 3D topographical surface, $s(x,y,z)$, which exhibits a peak for each person crossing under the sensor ($\nabla(s) = 0$). The $s(x,y,z)$ function is processed to extract relevant peaks which are then fed into a multi-modal tracking filter that keeps laser-based 3D trajectory for hypothetical person objects. These hypothesis, $OTL_i(t)$, are updated on the basis of the $k-th$ detected peak p_k in the current scan according to the following gate-reject/gate-augment rule:

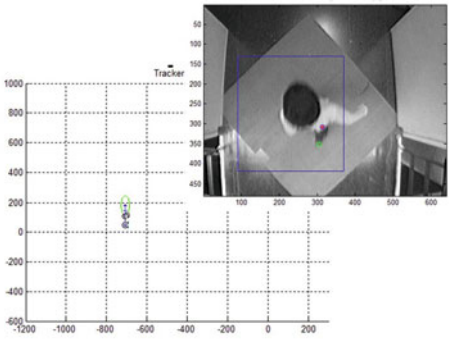


Fig. 2. EKF blobs tracking example

$$\begin{cases} \text{if } \min_{i=1\ldots n_l} Dist(k,i) < \mu_{rej} & \text{Update } OTL_i(t) \\ \text{if } \min_{i=1\ldots n_l} Dist(k,i) > \mu_{aug} & n_l = n_l + 1, \\ & OTL_{n_l}(t) = p_k \\ \text{otherwise} & discard \end{cases}$$

where $Dist(k,i) = \|ProjXY(p_k, M_l) - ProjXY(OTL_i(t-1), M_l)\|$, and $ProjXY$ is the transformation of the 3D peak coordinates into the XY, n_l is the number of current active object trajectories, μ_{rej} is the gate reject value and μ_{aug} is the gate augment value.

Each time a peak is associated to a trajectory, the area under that peak is computed and integrated to obtain a volume estimation of the object, OV_i, assuming constant velocity.

A trajectory is thus defined in a time interval, starting at its creation $OTL_{i_{ti}}$, and finishing when no peak is associated to the trajectory in the current laser acquisition, $OTL_{i_{tf}}$. Once the trajectory is completed, it is processed to estimate if it could correspond to a crossing person, according to a set of conditions: persistence ($OTL_{i_{tf}} - OTL_{i_{ti}} > t_{min}$), max height ($Max_z(OTL_i) > height_{min}$) and volume ($OV_i > vol_{min}$); where t_{min}, $height_{min}$ and vol_{min} are lower thresholds for trajectory duration, maximum height and volume, respectively. These conditions are defined to try to reduce false positive detections. As a result of this process, a set H_l of laser-based hypothesis about the number and location of people crossing is generated.

2.3 Integration

The high-level heuristic module fuses evidences and sensor data from the detection and filtering layers to generate the final decisions on people crossing events, assigning global certainty values. The two hypothesis sets H_l and H_c are cross validated to get a more robust estimation of people get in/out counting.

Basically, the temporal references (event initial/final time) and spatial references (trajectory coordinates) are used to identify a given object from one set into the other.

The global detection system can label and assign certainty to the crossings depending on they have only range or visual confirmation or both. In general terms, high certainty values are produced when pairs of visual and range data evidences are found to be temporally and spatially coherent. In case of discrepancies, certainty values based on special rules are applied, giving more credibility to the source that has integrated more measures at a reasonable certainty level.

3 Experiments and Results

Several tests have been performed on an experimental setup simulating the conditions of a public transportation conditions, with both visual and range sensors installed at 2.5 meters on a door frame. Data collected include diverse cases of individual and paired in/out door traversing, and simultaneous opposite direction trajectories. Illumination changes have been also introduced artificially switching lights to observe system response. Camera frames were captured from a firewire web-cam in 640x480 format at 30 Hz, while the laser provided 10 Hz scans over a 180 degrees sector with 1 degree angular resolution.

Although sensor configuration was not especially suited due to space and, mainly, height limitations, promising results have been achieved. So, in a sequence of approximately 150 crossing events, around a 90% were correctly detected with the integrated system.

Table 1. Summary of detection test results

Constant illumination				Illumination changes				Troublesome			
Real		Detected		Real		Detected		Real		Detected	
In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
31	33	29	30	29	28	26	25	12	14	7	6

Table 1 summarizes the results of real vs. detected events obtained in three different scenarios: constant illumination, illumination changes and troublesome. In the first one the illumination was kept constant during the test. In the second one the illumination level was changed artificially several times during the test. The third scenario included a collection of adverse situations such as: crowded group crossing, erratic trajectories, runners, etc.

A small bias can be observed in the table when comparing in and out events. This is due to a vertical sensor misalignment that caused a slightly larger detection area for in events.

Individually considered, laser detection suffers some problems with false positives due to arm movement and poor detection on fast walking people. On the other hand camera detection showed a lower performance when analyzing tall

people crossings and experiences some problems due to auto-iris adjust during illumination changes. The combined detection compensates for these conditions yielding a better global result.

Some situations are clearly not solved by this system configuration, for example children walking close to their parents, people collisions or the use of umbrellas.

4 Conclusions

A low-cost solution to people counting applications in public transportation have been proposed and tested. The combination of range detection and visual detection mechanisms contributes to compensate some specific problems of each method, exhibiting more robust results in getting in/out counting. Regarding more specifically illumination changes, the system is able to discard artificial motion blobs, either on filtering or fusion stages.

Future work includes specific experiments in crowded groups environments and more intensive testing and comparison with range cameras.

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movem. IEEE Transactions on Intelligent Transportation Systems Transactions on Pattern Analysis and Machine Intelligence 23(257-267), 3 (2001)
2. Bozzoli, M., Cinque, L., Sangineto, E.: A statistical method for people counting in crowded environments. In: 14th International Conference on Image Analysis and Processing (2007)
3. Chan, A.B., Liang, Z.-S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Computer Vision and Pattern Recognition, pp. 1–7 (2008)
4. Cui, J., Zha, H., Zhao, H., Shibasaki, R.: Multi-modal tracking of people using laser scanners and video camera. Image and Vision Computing 26(2), 240–252 (2008)
5. Katabira, K., Nakamura, K., Zhao, H., Shibasaki, R.: A method for counting pedestrians using a laser range scanner. In: 25th Asian Conference on Remote Sensing (ACRS 2004), Thailand, November 22-26 (2004)
6. Kellokumpu, V., Zhao, G., Pietikinen, M.: Recognition of human actions using texture descriptors. Machine Vision and Applications (2010) (in press)
7. Lee, G.G., Ki Kim, H., Yoon, J.Y., Kim, J.J., Kim, W.Y.: Pedestrian counting using an IR line laser. In: International Conference on Convergence and Hybrid Information Technology 2008 (2008)
8. Mathews, E., Poigné, A.: Evaluation of a "smart" pedestrian counting system based on echo state networks. EURASIP Journal on Embedded Systems, 1–9 (2009)
9. Scheutz, M., McRaven, J., Cserey, G.: Fast, reliable, adaptive, bimodal people tracking for indoor environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), vol. 2, pp. 1347–1352 (2004)

Lossless Compression of Polar Iris Image Data

Kurt Horvath[1], Herbert Stögner[1], Andreas Uhl[1,2,⋆], and Georg Weinhandel[1]

[1] School of CEIT, Carinthian University of Applied Sciences, Austria
[2] Department of Computer Sciences, University of Salzburg, Austria
uhl@cosy.sbg.ac.at

Abstract. The impact of using different lossless compression algorithms when compressing biometric iris sample data from several public iris databases is investigated. In particular, the application of dedicated lossless image codecs (lossless JPEG, JPEG-LS, PNG, and GIF), lossless variants of lossy codecs (JPEG2000, JPEG XR, and SPIHT), and a few general purpose file compression schemes is compared. We specifically focus on polar iris images (as a result after iris detection, iris extraction, and mapping to polar coordinates). The results are discussed in the light of the recent ISO/IEC FDIS 19794-6 standard and IREX recommendations.

1 Introduction

With the increasing usage of biometric systems the question arises naturally how to store and handle the acquired sensor data (denoted as sample data subsequently). In this context, the compression of these data may become imperative under certain circumstances due to the large amounts of data involved. For example, in distributed biometric systems, the data acquisition stage is often dislocated from the feature extraction and matching stage (this is true for the enrolment phase as well as for authentication). In such environments the sample data have to be transferred via a network link to the respective location, often over wireless channels with low bandwidth and high latency. Therefore, a minimisation of the amount of data to be transferred is highly desirable, which is achieved by compressing the data before transmission.

Having found that compression of the raw sensor data can be advantageous or even required in certain applications, we have to identify techniques suited to accomplish this task in an optimal manner. In order to maximise the benefit in terms of data reduction, lossy compression techniques have to be applied. However, the distortions introduced by compression artifacts may interfere with subsequent feature extraction and may degrade the matching results. As an alternative, lossless compression techniques can be applied which avoid any impact on recognition performance but are generally known to deliver much lower compression rates. An additional advantage of lossless compression algorithms is that these are often less demanding in terms of required computations as compared to lossy compression technology.

In this work, we experimentally assess the the application of several lossless compression schemes to iris image sample data as contained in several public

⋆ Corresponding author.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 329–337, 2011.

iris databases. In Section 2, we briefly review related work on biometric sample data compression. Section 3 is the experimental study where we first describe the applied algorithms / software and biometric data sets. Subsequently, results with respect to achieved compression ratios for polar iris image sets and selected recangular image data sets are discussed. Section 4 concludes this work.

2 Biometric Sample Compression

During the last decade, several algorithms and standards for compressing image data relevant in biometric systems have evolved. The certainly most relevant one is the ISO/IEC 19794 standard on Biometric Data Interchange Formats, where in its former version (ISO/IEC 19794-6:2005), JPEG and JPEG2000 (and WSQ for fingerprints) were defined as admissible formats for lossy compression, whereas for lossless and nearly lossless compression JPEG-LS as defined in ISO/IEC 14495 was suggested. In the most recently published version (ISO/IEC FDIS 19794-6 as of August 2010), only JPEG2000 is included for lossy compression while the PNG format serves as lossless compressor. These formats have also been recommended for various application scenarios and standardized iris images (IREX records) by the NIST Iris Exchange (IREX `http://iris.nist.gov/irex/`) program.

A significant amount of work exists on using compression schemes in biometric systems. However, the attention is almost exclusively focussed on lossy techniques since in this context the impact of compression to recognition accuracy needs to be investigated. One of the few results on applying lossless compression techniques exploits the strong directional features in fingerprint images caused by ridges and valleys. A scanning procedure following dominant ridge direction has shown to improve lossless coding results as compared to JPEG-LS and PNG [1]. In recent work [2] a set of lossless compression schemes has been compared when applied to image data from several biometric modalities like fingerprints, hand data, face imagery, retina, and iris.

In the subsequent experimental study we will apply an extended set of lossless compression algorithms to image data from different public iris image databases. Extensive results with respect to achieved compression ratio are shown. Specifically, we focus on polar iris images (as a result after iris detection, iris extraction, and mapping to polar coordinates, corresponding to KIND16 IREX records). While in the former version of the corresponding standard this type of imagery has been covered (ISO/IEC 19794-6:2005), the most recently published version (ISO/IEC FDIS 19794-6 as of August 2010) does no longer include this data type (based on the IREX recommendations). However, for applications not focussing on data exchange with other systems this data type can still be an option due to the extremely low data volume. In addition, employing this data type in a distributed biometric system shifts iris detection, extraction, and rectangular warping away from the feature extraction / matching device to the acquisition device since these operations are performed **before** compression and transmission. This can be of advantage in situations where the feature extraction / matching device is highly busy due to identification-mode operations (e.g. consider a scenario where numerous surveillance cameras submit data to the

feature extraction and matching device for identification) and therefore can lead to higher throughput of the entire system. Also in applications where reference data is stored in encrypted manner in databases and decrypted for each matching procedure a small data amount is favourable to minimize the effort required for repeated decryption operations.

3 Experimental Study

3.1 Setting and Methods

Compression Algorithms. We employ 4 dedicated lossless image compression algorithms (lossless JPEG – PNG), 3 lossy image compression algorithms with their respective lossless settings (JPEG2000 – JPEG XR), and 5 general purpose lossless data compression algorithms:

Lossless JPEG Image Converter Plus[1] is used to apply lossless JPEG, the best performing predictor (compression strength 7) of the DPCM scheme is employed.

JPEG-LS IrfanView[2] is used to apply JPEG-LS which is based on using Median edge detection and subsequent predictive and Golumb encoding (in two modes: run and regular modes) [3].

GIF is used from the XN-View software[3] employing LZW encoding.

PNG is also used from the XN-View implementation using an LZSS encoding variant setting compression strength to 6.

JPEG2000 Imagemagick[4] is used to apply JPEG2000 Part 1, a wavelet-based lossy-to-lossless transform coder.

SPIHT lossy-to-lossless zerotree-based wavelet transform codec[5].

JPEG XR FuturixImager[6] is used to apply this most recent ISO still image coding standard, which is based on the Microsoft HD format.

7z uses LZMA as compression procedure which includes an improved LZ77 and range encoder. We use the 7ZIP software[7].

BZip2 concatenates RLE, Burrows-Wheeler transform and Huffman coding, also the 7ZIP software is used.

Gzip uses a combination of LZ77 and Huffman encoding, also the 7ZIP software is used.

ZIP uses the DEFLATE algorithm, similar to Gzip, also the 7ZIP software is used.

UHA supports several algorithms out of which ALZ-2 has been used. ALZ is optimised LZ77 with an arithmetic entropy encoder. The WinUHA software is employed[8].

[1] http://www.imageconverterplus.com/
[2] http://irfanview.tuwien.ac.at
[3] http://www.xnview.com/
[4] http://www.imagemagick.org/script/download.php
[5] http://www.cipr.rpi.edu/research/SPIHT
[6] http://fximage.com/downloads/
[7] http://www.7-zip.org/download.html
[8] http://www.klaimsoft.com/winuha/download.php

Sample Data. For all our experiments we used the images in 8-bit grayscale information per pixel in .bmp format since all software can handle this format (except for SPIHT which requires a RAW format with removed .pgm headers). Database imagery has been converted into this format if not already given so, colour images have been converted to the YUV format using the Y channel as grayscale image. Only images that could be compressed with all codecs have been included into the testset as specified below. We use the images in their respective original resolutions (as rectangular iris images) and in form of polar iris images, which correspond to iris texture patches in polar coordinates which are obtained after iris segmentation and log-polar mapping. For generating these latter type of images, we use an open-source MatLAB iris-recognition implementation which applies a 1D Gabor-filter version of the Daugman iris code strategy [4] for iris recognition[9]. Depending on size and contrast of the rectangular iris images, several parameters for iris texture segmentation had to be adjusted accordingly (functions `Segmentiris.m`, `findcircle.m`, and `findline.m` are affected, e.g. the parameters `lpupilradius`, `Upupilradius`, `Hithresh`, `Lowthresh`, etc.) and the size of the resulting polar iris images has been fixed to 240×20 pixels for all databases. Nevertheless, iris segmentation was not successful in all cases, so we also provide the number of polar iris images per database used subsequently in compression experiments.

CASIA V1 database[10] consists of 756 images with 320×280 pixels in 8 bit grayscale .bmp format, 756 polar iris images have been extracted.

CASIA V3 Interval database (same URL as above) consists of 2639 images with 320×280 pixels in 8 bit grayscale .jpeg format, 2638 polar iris images have been extracted.

MMU database[11] consists of 457 images with 320×240 pixels in 24 bit grayscale .bmp format, 439 polar iris images have been extracted.

MMU2 database (same URL as above) consists of 996 images with 320×238 pixels in 24 bit colour .bmp format, 981 polar iris images have been extracted.

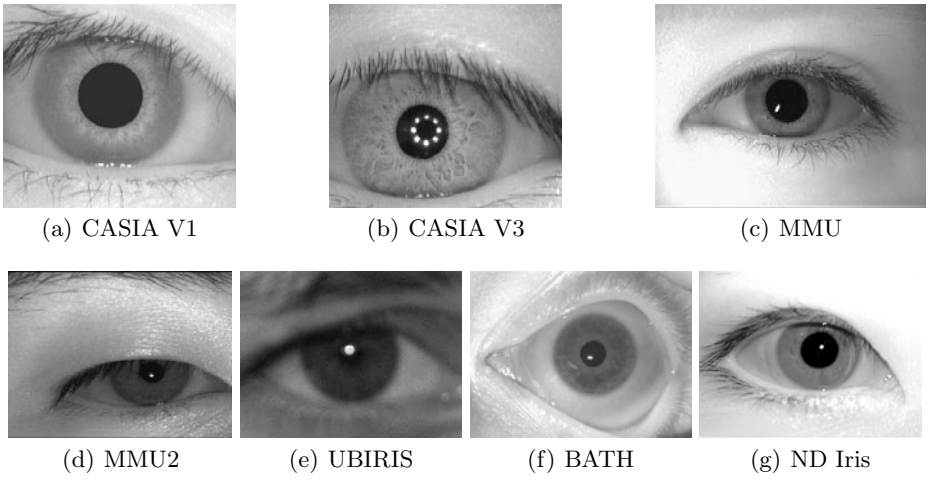
UBIRIS database[12] consists of 1876 images with 200×150 pixels in 24 bit colour .jpeg format, 614 polar iris images have been extracted.

BATH database[13] consists of 1000 images with 1280×960 pixels in 8 bit grayscale .jp2 (JPEG2000) format, 734 polar iris images have been extracted.

ND Iris database[14] consists of 1801 images with 640×480 pixels in 8 bit grayscale .tiff format, 1626 polar iris images have been extracted.

Figures 1 and 2 provide one example image from each database, the former a rectangular iris image, the latter an extracted polar iris image.

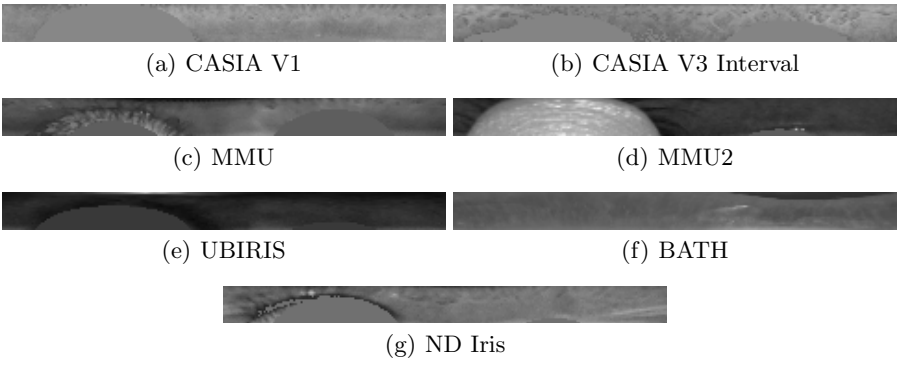
[9] `http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html`
[10] `http://http://www.cbsr.ia.ac.cn/IrisDatabase.htm/`
[11] `http://pesona.mmu.edu.my/~ccteo/`
[12] `http://www.di.ubi.pt/~hugomcp/investigacao.htm`
[13] `http://www.irisbase.com/`
[14] `http://www.nd.edu/~cvrl/CVRL/Data_Sets.html`



(a) CASIA V1 (b) CASIA V3 (c) MMU

(d) MMU2 (e) UBIRIS (f) BATH (g) ND Iris

Fig. 1. Example rectangular iris images from the used databases



(a) CASIA V1 (b) CASIA V3 Interval

(c) MMU (d) MMU2

(e) UBIRIS (f) BATH

(g) ND Iris

Fig. 2. Example iris polar images from the used databases

3.2 Results

In the subsequent plots, we display the achieved averaged compression ratio on the y-axis, while giving results for different databases or compression algorithms on the x-axis. The small black "error" bars indicate result standard deviation.

When comparing all databases under the compression of a single algorithm, JPEG-LS and PNG provide prototypical results shown in fig. 3 which are very similar to that of most other compression schemes in that there are no significant differences among different databases. Please note that we cannot provide results for SPIHT since the software does not support the low resolution of the polar iris images in y-direction.

For most databases, we result in a compression ratio of about 2.5 or slightly above for JPEG-LS. PNG on the other hand exhibits even less result variability,

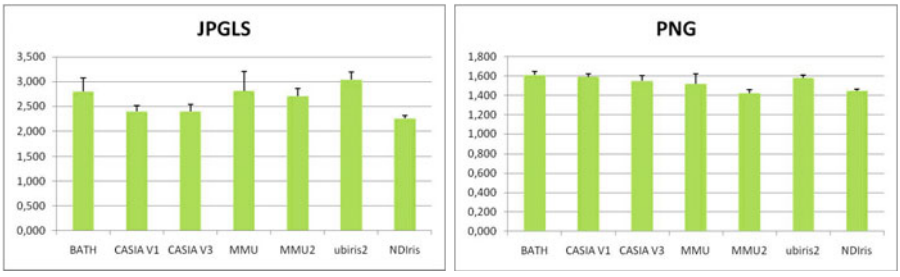


Fig. 3. Compression ratios achieved by JPEG-LS and PNG

however, compression ratio does not exceed 1.6 for all databases considered. In the light of the change from JPEG-LS to PNG in the recent ISO/IEC FDIS 19794-6 standard this is a surprising result.

In the following, we provide results for the different databases considered. Fig. 4 shows the results for the CASIA databases. We notice some interesting effects. First, JPEG-LS is the best algorithm overall. Second, for CASIA V1, ZIP is by far the best performing general purpose compressor while UHA is the best of its group for CASIA V3. Third, we observe suprisingly good results for lossless JPEG while fourth, the results for JPEG XR are almost as poor as those for GIF and PNG.

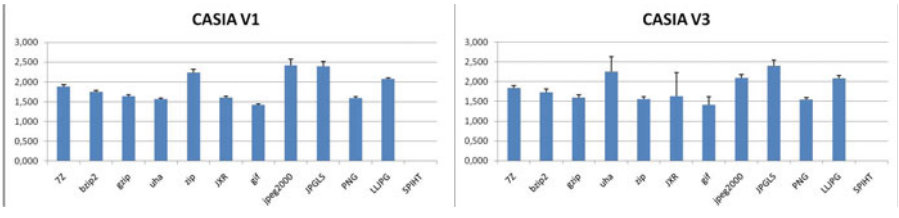


Fig. 4. Compression ratios achieved for polar iris images from the CASIA datasets

As shown in fig. 5, for the MMU (and MMU2 which gives almost identical results) and the ND Iris databases we obtain similar results as for CASIA V1. ZIP is the best general purpose algorithm and JPEG-LS is the best algorithm overall. Also, lossless JPEG performs well. There is an interesting fact to notice. In [2], JPEG2000 has been applied to the MMU dataset in lossless mode with surprisingly good results, however, in this work rectangular iris image data was considered. Here, we do not at all observe specific behaviour of JPEG2000 when applied to the MMU dataset, the results are perfectly in line with those for other datasets.

Similarly, for the BATH and UBIRIS databases JPEG-LS is the best algorithm as shown in fig. 6, JPEG2000 and lossless JPEG perform well. Main

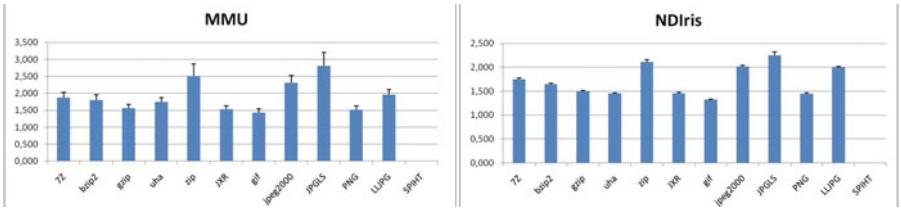


Fig. 5. Compression ratios achieved for polar iris images of the MMU and ND Iris datasets

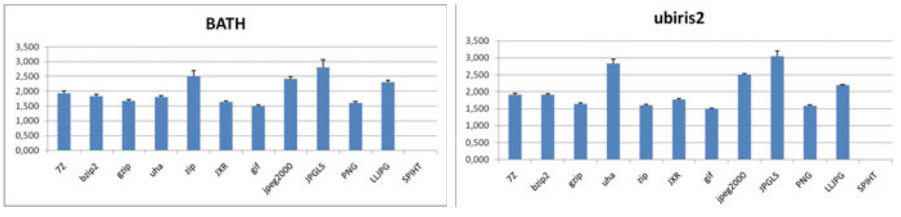


Fig. 6. Compression ratios achieved for polar iris images of the BATH and UBIRIS datasets

difference is again the performance of ZIP and UHA – while for BATH ZIP is the best general purpose algorithm, for the UBIRIS dataset UHA is the second best algorithm overall.

Table 1 displays an overview of all databases. For the polar iris images the situation is clear: JPEG-LS is the best algorithm for all datasets (except for CASIA V1 with JPEG2000 ranked first) whereas GIF is always worst. Considering the overall compression ratio achieved, we observe a range of 2.25 - 3.04 for the best techniques. This result taken together with the already small data amount for uncompressed polar iris images makes the required overall data rate very small for this configuration. It is also worth noticing that despite not being specifically designed for image compression purposes, ZIP and UHA excel for several databases, however results vary among different datasets in a non-predictable manner as opposed to the top-performing dedicated image compression schemes.

Table 1. Best and worst compression algorithm for each database (polar iris images) with corresponding achieved compression ratio

	Best	Ratio	Worst	Ratio
CASIA V1	JPEG2000	2.42	GIF	1.42
CASIA V3 Int.	JPEG-LS	2.40	GIF	1.41
MMU	JPEG-LS	2.81	GIF	1.43
MMU2	JPEG-LS	2.71	GIF	1.29
UBIRIS	JPEG-LS	3.04	GIF	1.50
BATH	JPEG-LS	2.80	GIF	1.50
ND Iris	JPEG-LS	2.25	GIF	1.32

When comparing the compression ratios to those which can be achieved with lossy techniques (e.g. [5]) we found the relation to be acceptable (considering the advantages of lossless techniques in terms of speed and non-impact on recognition). Polar iris images cannot be compressed that severely using lossy techniques due to the much lower resolution. Therefore, the achieved compression ratios of lossless and lossy schemes differ not too much, such that lossless compression techniques can be a realistic alternative. This is specifically the case for JPEG-LS which exhibits the best compression results and very low computational demands [2,3].

4 Conclusion and Future Work

Overall, JPEG-LS is the best performing algorithm for almost all datasets for polar iris images. Therefore, the employment of JPEG-LS in biometric systems can be recommended for most scenarios which confirms the earlier standardisation done in ISO/IEC 19794. The current choice for a lossless compression scheme in the recent ISO/IEC FDIS 19794-6 standard relying on the PNG format on the other hand seems to be questionable based on the results of this study, at least for polar iris image data. Moreover, as shown in [2], JPEG-LS turns out to be also significantly faster compared to PNG.

We observe compression ratios about 3 and additionally, the ratios found when applying lossy compression schemes to those kind of data are much lower compared to the rectangular iris case due to the much lower resolution. As a consequence, for polar iris images lossless compression schemes can be considered a sensible alternative to lossy schemes in certain scenarios where it is important to limit the computational effort invested for compression.

Acknowledgements

Most of the work described in this paper has been done in the scope of the ILV "Compression Technologies and Data Formats" (winter term 2009/2010) in the master program on "Communication Engineering for IT" at Carinthia Tech Institute. The artificial name "Georg Weinhandel" represents the following group of students working on this project: Ecker Sebastian, Lercher Markus, Montagnana Emiglio, Pollak David, Pölz Florian, and Rieger Thomas. This work has been partially supported by the Austrian Science Fund, project no. L554-N15.

References

1. Thärna, J., Nilsson, K., Bigun, J.: Orientation scanning to improve lossless compression of fingerprint images. In: Kittler, J., Nixon, M. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 343–350. Springer, Heidelberg (2003)
2. Weinhandel, G., Stögner, H., Uhl, A.: Experimental study on lossless compression of biometric sample data. In: Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis, ISPA 2009, Salzburg, Austria (September 2009)

3. Weinberger, M., Seroussi, G., Sapiro, G.: Lossless image compression algorithm: Principles and standardization into JPEG-LS. IEEE Transactions on Image Processing 9(8), 1309–1324 (2000)
4. Daugman, J.: How iris recognition works. IEEE Transactions on Circiuts and Systems for Video Technology 14(1), 21–30 (2004)
5. Konrad, M., Stögner, H., Uhl, A.: Custom design of JPEG quantisation tables for compressing iris polar images to improve recognition accuracy. In: Tistarelli, M., Nixon, M. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1091–1101. Springer, Heidelberg (2009)

Learning Features for Human Action Recognition Using Multilayer Architectures

Manuel Jesús Marín-Jiménez[1], Nicolás Pérez de la Blanca[2], and María Ángeles Mendoza[2]

[1] University of Córdoba, Córdoba, Spain
mjmarin@uco.es
[2] University of Granada, Granada, Spain
nicolas@ugr.es, nines@decsai.ugr.es

Abstract. This paper presents an evaluation of two multilevel architectures in the human action recognition (HAR) task. By combining low level features with multi-layer learning architectures, we infer discriminative semantic features that highly improve the classification performance. This approach eliminates the difficult process of selecting good mid-level feature descriptors, changing the feature selection and extraction process by a learning stage. The data probability distribution is modeled by a multi-layer graphical model. In this way, this approach is different to the standard ones. Experiments on KTH and Weizmann video sequence databases are carried out in order to evaluate the performance of the proposal. The results show that the new learnt features offer a classification performance comparable to the state-of-the-art on these databases.

1 Introduction

Most of the current approaches to motion recognition share the same two stages architecture: feature selection from image information (optical flow, contours, gradients, texture, etc) and feeding a good classifier with them. Although good results have been achieved, all these approaches share the same weakness of having to guess what functions of the raw data represent good features for the classification problem. Recently, Hinton in [7] introduced new models to learn high level semantic features from raw data: Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBN). These two algorithms have been shown to be very successful in some image classification problems [17]. However, in our case, the motion is not explicitly represented in the raw image and one representation must be introduced.

Our interest in this paper is to evaluate how discriminative are the high-level features provided by the multilayer architectures and, on the other hand, to evaluate the probability representations given by the multi-layer architectures when used in HAR tasks. The main contribution of this paper is a comparative study between extracted features and learned features from three different classifiers (SVM, GentleBoost, SoftMax) in a supervised framework.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 338–346, 2011.

The paper is organized as follows: in section 2 the related works are presented and discussed. Section 3 summarizes the main properties of the multilayer architectures. Section 4 shows the experimental setup with the obtained results. In section 5 the discussion, conclusions and future work are presented.

2 Related Works

HAR using middle-level feature extraction plus a classification stage has received a lot of attention in the last years. A good summary of early works can be found in [14,10] and references therein. Important to all of them is the discriminative level of the selected features. However our focus here is different, since we try to make explicit the discriminative information implicit in the simple features.

Training a RBM or DBN model is a very difficult optimization problem due mainly to the difficulty of getting good estimators of the gradients on the hidden layer parameters. Contrastive Divergence (CD) is an approximation to the data log-likelihood gradient estimation, introduced by Hinton in [5], that has shown to work very efficiently in training RBMs and DBN (see [1]).

In order to define the input information to the multilayer model we use the descriptor (aHOF) proposed for HAR in [13]. It is assumed that the subject performing the action has been detected and/or tracked and the bounding box (BB) enclosing such image regions are known. From the BB sequence the optical flow (OF) for each pair of consecutive frames is computed.

The main steps used to compute aHOF descriptor are summarized as follows: *i)* compute OF; *ii)* split the image window in $M \times N$ non-overlapping cells; *iii)* compute a Histogram of Optical Flow (HOF) at each spatial cell over OF orientations and magnitudes; *iv)* accumulate HOFs along time axis; and, *v)* normalize the accumulated HOFs. In general, the full sequence is split in several overlapping subsequences of a prefixed length computing a descriptor on each one.

3 Multilayer Models

A Restricted Boltzmann Machine (RBM) is a Boltzmann Machine with a bipartite connectivity graph between observed x and hidden h variables [1] (see Fig.1.a). In the case of binary variables the conditional distribution can be written as $P(\mathbf{h}_i|\mathbf{x}) = sigm(c_i + \mathbf{W}_i\mathbf{x})$, $P(\mathbf{x}_j|\mathbf{h}) = sigm(b_j + \mathbf{W}_j h)$ where $sigm(x) = (1 + e^{-x})^{-1}$ is the logistic sigmoidal function, $\mathbf{W}_i$ and $\mathbf{W}_j$ represent the i-th row and the j-th column of the $\mathbf{W}$-matrix respectively, and b_j and c_i are biases.

An important property of this model is that the joint model distribution on (x, h) can be sampled using the Gibbs sampling on the conditional distribution [7]. A Monte-Carlo Markov Chain (MCMC) sampling is started from the empirical data distribution, denoted it by $\mathbf{P_0}$. After sampling $\mathbf{h}$ from $P(\mathbf{h}|\mathbf{x})$, we sample $\tilde{\mathbf{x}}$ from $P(\tilde{\mathbf{x}}|\mathbf{h})$, denoting this new distribution by $\mathbf{P_1}$.

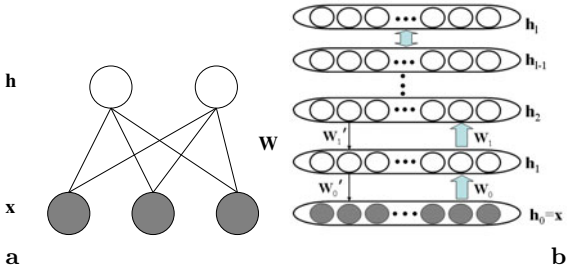


Fig. 1. RBM and Deep Belief Network. (a) Example of a RBM with 3 observed and 2 hidden units. (b) Example of a DBN with l hidden layers. The upward arrows only play a role in the training phase. W_i' is W_i^T (W_i transpose) when a RBM is trained. The number of units per layer can be different.

A RBM is trained from data by maximizing the log-likelihood gradient on the vector of parameters $\theta = (\mathbf{b}, \mathbf{c}, \mathbf{W})$ by using CD. In [6], it is shown that in the binary case the increment updating equations are as follows:

$$\begin{aligned} \Delta W_{ij}^t &\leftarrow \alpha * \Delta W_{ij}^{t-1} + \tau \left(< \mathbf{x}_i \mathbf{h}_j >_{\mathbf{P}_0} - < \tilde{\mathbf{x}}_i \mathbf{h}_j >_{\mathbf{P}_1} - \gamma * W_{ij}^{t-1} \right) \\ \Delta b_i^t &\leftarrow \alpha * \Delta b_i^{t-1} + \tau \left(< \mathbf{x}_i >_{\mathbf{P}_0} - < \tilde{\mathbf{x}}_i >_{\mathbf{P}_1} \right) \\ \Delta c_j^t &\leftarrow \alpha * \Delta c_j^{t-1} + \tau \left(< \mathbf{h}_j >_{\mathbf{P}_0} - < \tilde{\mathbf{h}}_j >_{\mathbf{P}_1} \right) \end{aligned} \tag{1}$$

τ being the learning rate, α the momentum and $\gamma = 0.0002$ a regularization constant. P_0 and P_1 being the distribution after 0 and 1 sampling steps.

By adding new layers to a RBM, a generalized multilayer model can be obtained. A Deep Belief Network (DBN) with l hidden layers is a mixed graphical model representing the joint distribution between the observed values $\mathbf{x}$ and the l hidden layers $\mathbf{h}_k$, by

$$P(\mathbf{x}, \mathbf{h}_1, \cdots, \mathbf{h}_l) = \prod_{k=0}^{l-2} P(\mathbf{h}_k | \mathbf{h}_{k+1}) P(\mathbf{h}_{l-1}, \mathbf{h}_l)$$

(see fig.1) where $\mathbf{x} = \mathbf{h}_0$ and each conditional distribution $P(\mathbf{h}_{k-1}|\mathbf{h}_k)$ can be seen as the conditional distribution of the visible units of a RBM associated with the $(k-1, k)$ layers in the DBN hierarchy. In [7] a strategy based on training a RBM on every two layers using CD is proposed to obtain the initial solution. From the initial solution, different fine tuning criteria for supervised and non-supervised experiments can be used. In the supervised case, a backpropagation algorithm from the classification error is applied fixing $W_i' = W_i^T$ (transpose).

4 Experiments and Results

We present an experimental study to evaluate the goodness of the descriptors generated by RBM and DBN models, in contrast to the descriptors built up from raw features. We mainly focus on supervised experiments.

The training of the models is performed in two stages: pre-training and fine-tuning. The RBM pre-training consist in a non-supervised training between both

layers where the learning weights try to fit the input data from P_1. For the DBN case every two consecutive layers are considered a RBM model. Equations 1 with learning-rate $\tau = 0.1$ and momentum $\alpha = 0.9$ on sample batches of size 100 have been used. The batch average value is used as the update. From 120 to 200 epochs are run for the full training. From the 120-th epoch, training is stopped if variation of the update gradient magnitude from iteration $t-1$ to t is lower than 0.01. A different number of batches are used depending on the length of the sequences in the database. For KTH, 14, 74, 16 and 28 batches, for scenarios 1-4, respectively. For Weizmann, we use 15. The $\mathbf{W}_{ij}$-parameters are initialized to small random numbers (<0.1) and the others parameters to 0.

The fine-tuning stage is carried out using a standard backpropagation algorithm using the label classification error measured on a new output layer. A layer with as many units as classes (from now on, *short-code*), is added to the 1024 top sigmoidal-layer (from now on, *long-code*). The connection between these two layers uses a SoftMax criteria to generate the short-code (label) from the long one, while the reverse connection remains sigmoidal. Here we fix the width of all the hidden layers to the width of the visible one (1024). Therefore, the number of training parameters for each RBM is (1024x1024)$\mathbf{W}$ + (1024)$\mathbf{b}$ + (1024)$\mathbf{c}$.

We assign a class label to a full video sequence by classifying multiple subsequences (same length) of the video, with SVM or GentleBoost (see [4]), and taking a final decision by *majority voting* on the subsequences. We convert the binary classfiers in multiclass ones by using the *one-vs-all* approach. Both classifiers are also compared with KNN and the *SoftMax* classifier. In this context, SoftMax classifier assigns to each sample the index of the maximum value in its *short-code*.

The pipeline of the processing stages used in our experiments is as follows: *i)* aHOF computation on the input video; *ii)* higher level feature learning/extraction by using RBM/DBN; and, *iii)* training/classification with discriminative classifiers on the extracted features.

Databases. We test our approach on two publicly available databases that have been widely used in action recognition: KTH human motion dataset [16] and Weizmann human action dataset [2].

KTH: this database contains a total of 2391 sequences, where 25 actors performs 6 classes of actions (walking, running, jogging, boxing, hand-clapping and hand-waving). The sequences were taken in 4 different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors (s4). As in [16], we split the database in 16 actors for training and 9 for test. In our experiments, we consider KTH as 5 different datasets: each one of the 4 scenario is a different dataset, and the mixture of the 4 scenarios is the fifth one. In this way we make our results comparable with others appeared in the literature.

Weizmann: this database consists of 93 videos, where 9 people perform 10 different actions: walking, running, jumping, jumping in place, galloping sideways, jumping jack, bending, skipping, one-hand waving and two-hands waving.

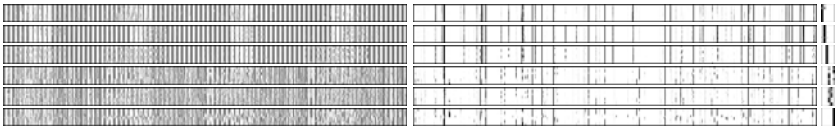


Fig. 2. RBM codes. Stacked vector of features for the 6 different actions in KTH. (a) aHOF data, (b) 1024-codes, (c) 6-codes. The darker the pixel, the greater the probability of taking value 1. Note in (b) the sparsity gained by encoding aHOF features in (a). For clarity, vectors in (c) have been scaled in width.

aHOF parameters. In order to carry out our experiments, we have chosen the same aHOF parameters proposed in [13]: 8x4 ($rows \times cols$) cells; 8-bins for orientation and 4-bins for magnitude are used in the histograms. Therefore, the full aHOF descriptor for each image is a 1024-vector with values in $(0, 1)$. We consider the descriptor a random vector where each bin value is a binary variable with probability equal to its value.

Results on KTH. All the results we show in this subsection, come from averaging the results of 10 repetitions of the experiment with different pairs of training/test sets.

We firstly evaluated the performance of the raw aHOF features in the classification task. Subsequences were extracted every other frame from the full length sequence. Empirically, we chose a length of 20 frames for the subsequences. For this setup, 94.6% of correct classification of the full senquences is achieved with GentleBoost. For the following experiments, we will always use subsequences of length 20 frames to compute the aHOF descriptors.

Table 1. Classification performance on KTH with high-level features. Mean performance is reported for the four scenarios mixed in a single dataset. Different classifiers and codes are compared. Subsequences have length 20.

L	SMax	1NN	5NN	SVM	SVM-6	GB	GB-6
1024	95.4	94.3	94.5	**96.0**	95.7	95.5	95.8

Regarding multilevel architectures, we begin by learning high-level features from the 1024-aHOF vectors with a 1024-1024-6 (input-hidden-top) architecture. In table 1, five classifiers are compared on short and long codes: (*i*) SoftMax, (*ii*) KNN on long-codes, (*iii*) SVM (radial basis) on long-codes, (*iv*) GentleBoost on long-codes, (*v*) SVM on short-codes (SVM-6), and (*vi*) GentleBoost on short-codes (GB-6). Notice that none of these results are on raw aHOF features.

Table 2 shows the confusion matrix on KTH for our best result (SVM-code-1024). Note that the highest confusions are *running* with *jogging* and *handclapping* with *boxing*.

Table 2. Confusion matrix for KTH. Scenarios 1+2+3+4. SVM on 1024 codes (from 1024-1024-6). Rows are the true classes, and columns the predicted ones.

	box	*hclap*	*hwave*	*jog*	*run*	*walk*
box	**99.6**	0.3	0.1	0.0	0.0	0.0
hclap	4.4	**92.8**	2.8	0.0	0.0	0.0
hwave	0.1	0.5	**99.4**	0.0	0.0	0.0
jog	0.0	0.5	0.0	**94.0**	3.0	2.5
run	0.0	0.0	0.0	7.5	**92.0**	0.5
walk	0.1	0.5	0.0	0.8	0.3	**98.4**

Table 3. One-layer VS multilayer. Different number of intermediate hidden layers are compared by using various classifiers. Row *aHOF* refers to the raw input data, therefore, SoftMax (*SMax* column) classification cannot be applied.

	Scenario 1			**Scenario 2**			**Scenario 3**			**Scenario 4**		
L	SMax	1NN	SVM	SMax	1NN	SVM	SMax	1NN	SVM	SMax	1NN	SVM
aHOF	-	94.8	95.1	-	93.3	96.3	-	90.5	88.2	-	**96.4**	**97.6**
1024	95.0	95.7	95.0	96.6	92.9	97.0	91.7	91.7	91.3	**95.7**	94.4	96.2
1024-1024	**95.2**	**95.8**	**95.5**	**96.7**	93.3	**97.5**	92.0	91.9	**92.4**	95.0	93.8	96.1
1024-1024-1024	94.9	95.1	95.2	96.2	**93.6**	96.3	**92.7**	**92.6**	92.3	94.5	93.8	94.8

One-layer VS multilayer. In this experiment, we are interested in studying the effect of the number of intermediate hidden layers. In particular, we carry out experiments with the following hidden layer architectures: 1024-6, 1024-1024-6 and 1024-1024-1024-6. The first layer is always the visible one (input data) and the last one (6 hidden units) is the SoftMax one. In table 3 we show a comparative of the classification performance for each separate scenario.

Comparison with the state-of-the-art. A comparison of our method with the state-of-the-art performance, on KTH database, can be seen in table 4.

Note that the learnt features, just based on optical flow, offers a classification performance comparable to the best result published up to our knowledge [12], with the same experimental setup.

We report results for each scenario trained and tested independently, as well as the results for the mixed scenarios dataset (i.e. *s1234*). The result reported by Lui *et al.* [12] corresponds to the mixed scenarios dataset, directly comparable with our *s1234*. Unfortunately, only Jhuang *et al.* [8] publish the individual results per scenario (here their *Avg.* score is the mean of the separate scenarios). In our case the average of the 4 separate scenarios is 94.9%.

The three bottom rows ([18,15,11]) of the table contain results on this database but using a different experimental setup. In particular, one of the best published result [11] (93.4% on *s1234*) uses both shape and motion features and, also, they use more actors for training (i.e. *leave-one-out*) than us.

Table 4. Comparison with the state-of-the-art on KTH. Column *s1234* corresponds to 'all-in-one' dataset, and columns *s1-s4* show the results per scenario. *Avg.* column shows the averaged result on the four scenarios. Symbol '-' indicates that such result is not available.

Method	**s1234**	**s1**	**s2**	**s3**	**s4**	**Avg**
aHOF+RBM+SVM	**96.0**	95.0	97.0	91.3	96.2	**94.9**
Laptev *et al.* [10]	91.8	-	-	-	-	-
Jhuang *et al.* [8]	-	96.0	86.1	88.7	95.7	91.6
Kovashka&Grauman [9]	94.5	-	-	-	-	-
Lui *et al.* [12]	96.0	-	-	-	-	-
Zhang *et al.* [18]	91.3	-	-	-	-	-
Schindler&Van Gool [15]	92.7	-	-	-	-	-
Lin *et al.* [11]	93.4	98.8	94	94.8	95.5	95.8

Figure 2 shows a graphical comparative of the amount of information provided by each one of the codes used: raw, 1024-vector and 6-vector. The classification results shown in table 1 emphasize that a very high percentage of this information is redundant and can be coded in few bits when adequate algorithms are used.

Results on Weizmann. On this database we perform a short classification experiment with the best configurations obtained for KTH database. In particular, we use a 1024-1024-10 architecture on 8x4 aHOF features from subsequences of length 20 frames. Now the top layer has 10 hidden units, as much as action categories.

Table 5 contains results obtained with different classifiers. The results we show come from averaging on a *leave-one-out* evaluation: 8 actors for training and 1 for testing. On average, for our best result, the system fails 3 sequences out of 93. With the same evaluation criterion, some authors have reported perfect classification on this dataset (e.g. [3,11]) but at the cost of extracting a very specific feature vector.

Table 5. Classification on Weizmann. Results on sequence classification by using different classifiers and codes.

L	**SMax**	**1NN**	**SVM**	**SVM-10**	**GB**	**GB-10**
1024	96.3	89.6	94.1	96.3	92.6	96.3

5 Discussion and Conclusions

Tables 1 and 3 show that the proposed scheme obtains state-of-the-art results on the KTH database. From Table 3 we conclude that the number of layers is the most important factor, increasing the semantic level of the features with the number of layers. However the proper training of a high number of layers still remains a hard problem.

Table 4 shows that our proposal reaches the state-of-the-art performance both globally and in two of the four scenarios. In our approach, it is important the idea of generating discriminative high-level features from low level information. Note that in three of the scenarios, the classification performance is above 95%. The lowest result in scenario 3 is due to the loose clothes (e.g. raincoat) used by the actors, what highly corruptes the quality of the computed optical flow in the person boundaries.

It is also interesting to remark the score obtained by the short-codes and long-codes respectively, see Table 1. Although short-codes have a slight loss in performance, the result is encouraging. It shows that all the information represented in a sequence could be encoded into very few numbers.

In conclusion, these new multilayer architectures show a big potential on complex computer vision classification problems. Our research focus for the future is on algorithms for training large multilayer architectures.

Acknowledgments. This work has been granted by the project CSD2007-00018 (MIPRCV) from the Spanish Minister of Science and Technology.

References

1. Bengio, Y.: Learning deep architectures for AI. Tech. Rep. 1312, Dept. IRO, Universite de Montreal (2007), `http://www.iro.umontreal.ca/~lisa/pointeurs/TR1312.pdf`
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Int. Conf. Comp. Vision., vol. 2, pp. 1395–1402 (2005)
3. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: CVPR (2008)
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Heidelberg (2001)
5. Hinton, G.: Training product of experts by minimizing contrastive divergence. Neural Computation 14(8), 1711–1800 (2002)
6. Hinton, G.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
7. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for Deep Belief Nets. Neural Computation 18, 1527–1554 (2006)
8. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proc. ICCV 2007, pp. 1–8 (2007)
9. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: CVPR (2010)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proc. CVPR 2008 (2008)
11. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: Int. Conf. Comp. Vision (2009)
12. Lui, Y.M., Beveridge, J., Kirby, M.: Action classification on product manifolds. In: Comp. Vision and Patt. Rec., pp. 833–839 (2010)
13. Marín-Jiménez, M., de la Blanca, N.P., Mendoza, M., Lucena, M., Fuertes, J.: Learning action descriptors for recognition. In: WIAMIS 2009, pp. 5–8. IEEE Computer Society, London (2009)

14. Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. In: CVIU, vol. 104, pp. 90–126 (2006)
15. Schindler, K., van Gool, L.: Combining densely sampled form and motion for human action recognition. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 122–131. Springer, Heidelberg (2008)
16. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local SVM approach. In: Int. Conf. Patt. Rec., Cambridge, U.K, vol. 3, pp. 32–36 (2004)
17. Torralba, A., Fergus, R., Weiss, Y.: Small codes and large database for recognition. In: Comp. Vision and Patt. Rec. (2008)
18. Zhang, Z., Hu, Y., Chan, S., Chia, L.: Motion context: A new representation for human action recognition. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 817–829. Springer, Heidelberg (2008)

Human Recognition Based on Gait Poses*

Raúl Martín-Félez, Ramón A. Mollineda, and J. Salvador Sánchez

Institute of New Imaging Technologies and Dept. Llenguatges i Sistemes Informàtics
Universitat Jaume I. Av. Sos Baynat s/n, 12071, Castelló de la Plana, Spain
{martinr,mollined,sanchez}@uji.es

Abstract. This paper introduces a new approach for gait analysis based on the Gait Energy Image (GEI). The main idea is to segment the gait cycle into some biomechanical poses, and to compute a particular GEI for each pose. Pose-based GEIs can better represent body parts and dynamics descriptors with respect to the usually blurred depiction provided by a general GEI. Gait classification is carried out by fusing separated pose-based decisions. Experiments on human identification prove the benefits of this new approach when compared to the original GEI method.

Keywords: gait recognition, GEI, pose estimation, decision level fusion.

1 Introduction

Gait can be defined as a manner or style of walking. Interestingly, there are studies asserting that every individual has a unique gait pattern [3], what has lead gait to be considered as a new biometric feature. When compared to other biometric features such as face, voice or fingerprint, gait has several attractive properties. It can be reliably perceived at a greater distance with simple instrumentation, and it does not require the cooperation or awareness of the individual.

There exist many applications that could benefit from gait analysis, including surveillance, diagnosis and treatment of gait-related disorders, motion capture in computer graphics and games, and so on.

However, there are also several factors that hinder the use of gait as a biometric feature. For instance, gait analysis is very sensitive to segmentation of the subject's silhouette, but also footwear, clothing, carrying conditions and walking speed may affect gait by reducing its discriminative power as a biometric. Even so, it is still useful to complement other biometric features under certain conditions (e.g. uncooperative subject, low quality images, etc.).

In literature, two main approaches have been proposed to obtain gait patterns from video sequences [1]: model-based and model-free methods. Proposals in the first group aim at recovering a structural model of human motion [8,11] by

* Partially supported by projects CSD2007-00018 and CICYT TIN2009-14205-C04-04 from the Spanish Ministry of Innovation and Science, P1-1B2009-04 from Fundació Bancaixa and PREDOC/2008/04 grant from Universitat Jaume I. Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 347–354, 2011.

matching the joint locations with a robust kinematic model of the human body. However, this is a hard task because of some problems such as occlusion of body parts, joint angle singularities, etc. On the other hand, model-free methods [4,7] do not use any model, and they are based on changes of the subject's appearance, which implicitly contain information about body movements.

Probably, the best-known model-free method is *gait energy image* (GEI) [4], which obtains an average silhouette image to represent both body shape and movements over a gait cycle. Although several potentially discriminative body parts usually appear blurred in GEI images (chest and back regions caused by movements of arms, and shape of arms and legs because of their motion), this method has proved to be effective in many tasks, such as human identification [13] and gender classification [12].

Biomechanical studies [9] assert that several recognizable poses happen successively in a gait cycle, e.g., those in which legs are spread, legs are closest together, etc. A first attempt to segment recognizable poses from gait video sequences was carried out by selecting a set of key frames representing some poses [2], but results on a human recognition task mostly depended on quality of the chosen frames.

In this work, a new way of using GEI is proposed and its effectiveness is compared with that of the original method. The main hypothesis supporting the proposal is that shape of body parts could be better noticed if different GEIs were separately obtained for some predefined key poses within a gait cycle [9]. Furthermore, they could also provide a few dynamic descriptors usually blurred in original GEI. By following this idea, several silhouettes are used to represent each pose instead of using a unique key frame as in [2], what is expected to be more robust to noise from individual frames. In the classification stage, each pose-based GEI is individually classified, and these decisions are fused to produce the subject recognition. Experiments prove the higher performance of this combined solution with respect to results obtained from the original GEI and from the particular key poses.

2 Background

This section describes the original GEI method and gives some information about the biomechanical phases and poses involved in a gait cycle.

2.1 Gait Energy Image (GEI)

As already said, GEI basically generates an *average silhouette* for a gait cycle, which reflects shape of the body parts and in some extent, their changes over time (gait dynamics). In this way, it reduces storage and time requirements and it is also more robust to noise of individual frames.

Before computing the GEI of a given gait sequence, its frames must be preprocessed as follows:

- **Foreground segmentation.** Given a frame, the aim is to obtain an image in which foreground is segmented from background and a silhouette is highlighted. It could be simply done, for instance, by background subtraction.
- **Silhouette extraction.** From each foreground frame, a cropped image is extracted from the bounding box that encloses all silhouette pixels.
- **Size normalization and horizontal alignment.** The silhouette image is scaled to a new one having a pre-fixed common height and a variable width to keep its original aspect ratio. Then this normalized silhouette is horizontally centered in a template of fixed sizes from the horizontal centroid value of its upper-half, since this part of the body involves fewer changes than the lower-half when a person walks.

Afterwards, given the set of preprocessed silhouettes of a gait video sequence $\{I_t(x, y)\}$ with $1 \leq t \leq N$, N being the number of silhouettes, and (x, y) referring to a specific position in the 2D image space, each gray-level pixel of a GEI is computed as $GEI(x, y) = \frac{1}{N} \sum_{t=1}^{N} I_t(x, y)$.

2.2 Phases and Key Poses within a Gait Cycle

According to biomechanical studies [9], walking is a cyclic process of limb motion to move the body forward while balance is simultaneously kept. Therefore, the relevant information of a gait pattern can be captured from a whole gait cycle (or stride), which is the interval between two sequential foot contacts by the same limb.

Limbs goes through two phases while walking. The first one is *Double Limb Stance*. It comprises the period of time in which both feet are on the ground for the transfer of body weight from the support limb to the other. The second phase is *Single Limb Stance* that consists of a larger period of time after the first phase in which one limb serves as a mobile source of support while the other limb is advancing or swinging to a new support place. Then limbs interchange their roles and a new two-phase succession completes the gait cycle.

More specifically, eight main poses can be distinguished during a gait cycle when both limbs are separately considered, but they can be reduced to four poses if the expected symmetry of both halves of a gait cycle is taken into account.

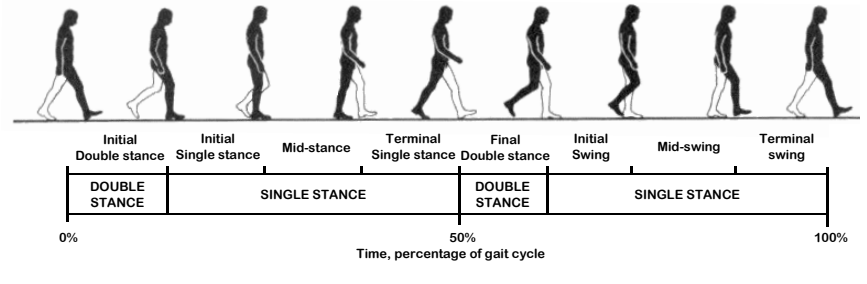


Fig. 1. Phases and poses within the gait cycle

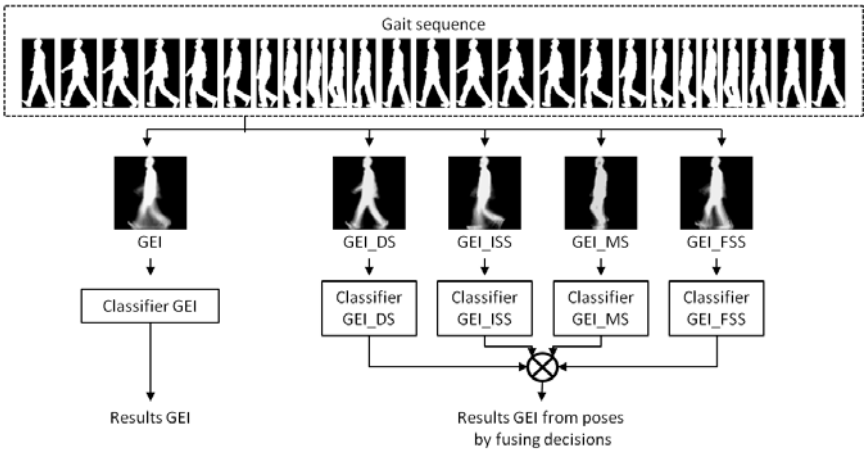


Fig. 2. General solution scheme

Figure 1 illustrates a full gait cycle showing the eight poses and the different phases.

3 Methodology

This section provides the basis of the method here proposed along with the details to generate GEIs for the different poses.

3.1 Fundamentals of the Proposed Method

The new way of using GEI here proposed is based on three main assumptions:

Assumption 1. Focusing on Figure 1, poses of the two halves of a gait cycle are expected to be symmetric if information about what limb is closer to the camera is unavailable. Therefore the step could be chosen as the gait cycle measure instead of the stride (unlike most of related works [4,13,12]).

Assumption 2. By using several pose-based GEIs, each one associated to a particular key pose, the shape of all body parts could be reflected in a more accurate way than in the representation given by a unique GEI averaging all poses. In addition, they could better show some dynamic features such as the length of the stride and extent of arm swing.

Assumption 3. Given a gait sequence with multiple gait cycles, most of works compute a different GEI for each gait cycle. However, a better representation could be obtained by comprising all cycles in a unique GEI, since noise of individual silhouettes could not affect so much, and it could lead to more efficient algorithms. If assumptions 2 and 3 are satisfied, pose-based GEIs could be generated from a higher number of frames.

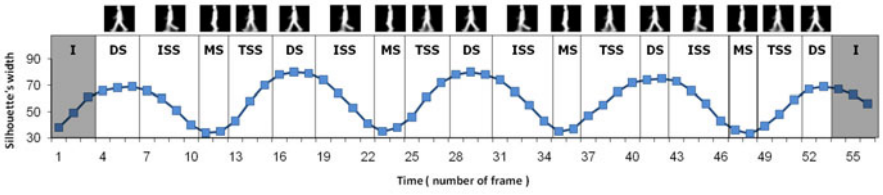


Fig. 3. Classifying frames in poses by their width

Under these assumptions, given a test gait sequence, the plain GEI of the whole sequence (image called GEI in Figure 2) and four pose-based GEIs during the complete sequence (images called GEI_DS, GEI_ISS, GEI_MS and GEI_FSS in Figure 2) are computed.

A general overview of the method, which has been illustrated in Figure 2, is now introduced. Firstly, individual frames are classified in one of four poses following the procedure detailed in Section 3.2. Then, for the whole sequence, all silhouettes belonging to a given pose are averaged to compute its corresponding GEI. From each one of the four pose-based GEIs, an individual decision about the identity of the subject is given according to the label of the most similar training GEI among those of the same pose. Finally, these four decisions are fused by majority voting to produce a unique decision more robust and probably more reliable than the individual decision obtained from the plain use of GEI. In case of a tie, the system rejects the classification of the corresponding sample.

3.2 Creating GEIs of Each Pose

In order to classify each silhouette in a landmark pose, the periodic signal provided by the silhouette width as a function of time for the whole sequence is used in a similar way to the work by Collins *et al.* [2]. As can be seen in Figure 3, the sequence of side-view silhouettes of a person walking defines a periodic function with peaks and valleys. The silhouette width alternatively expands (peaks) and contracts (valleys) over time as the person's legs spread and come back together again during the gait cycle. Therefore, each step comprises from peak to peak in the periodic signal.

It conducts to the following classification of frames into gait poses:

1. **Double limb Stance (DS)**, the frame at the current peak and those frames in the neighbourhood with a very close width value (both legs spread and touching the ground).
2. **Initial Single Stance (ISS)**, from the last frame of DS to a frame that surrounds the next valley (the front leg is on the ground and the rear leg is swinging towards it).
3. **Mid-Stance (MS)**, the frame at the current valley and those frames in the neighbourhood with a very close width value (legs are closest together with the swinging leg just passing the static one).

4. **Terminal Single Stance (TSS)**, from the last frame of MS to a frame at the surroundings of the following peak (the supporting leg is now the rear one, and the swinging leg appears as the front leg).

This simple and robust method is able to accurately determine to which pose each frame belongs to. Figure 3 shows an example of how silhouettes are labelled.

4 Experiments and Results

Experiments aim at assessing the effectiveness of combining pose-based decisions with respect to a plain classification using the original GEI method.

4.1 Database Description and Preprocessing

Experiments are carried out on CASIA Gait Database [5] - Database B, which consists of videos from 124 subjects. Only the six gait sequences in which each subject appears walking in their side-view are considered, what gives a total of 744 sequences to be used in experiments. In addition, this database provides well-segmented foreground images that have been used in this work as inputs to the silhouette extraction step (see Section 2.1).

Once silhouettes have been obtained, they are scaled and horizontally aligned to images with two different prefixed sizes (32x32 and 64x64 pixels) in order to measure the effect of image size. The gray-level value of each pixel is considered a different feature, thus a high dimensionality should be managed (more than 1000 features). In order to avoid this problem, dimensionality is reduced by using the well-known PCA technique [6]. In this way, the original GEI features are projected onto a smaller number of new uncorrelated ones that accounts for a given percentage of the variance (95% in this work).

4.2 Experimental Setup

Apart from the main objective previously described, experiments also aim at measuring effects of dimensionality reduction with PCA. Due to the limited number of samples available in the database (744), the *leaving-one-out* method is used to estimate the performance of each strategy. Since the CASIA Database contains several gait sequences for each subject, at least one gait sequence belonging to the tested subject is in the gallery set.

The *nearest neighbor classifier (1NN)* was here selected because of its simplicity and common good performance. With respect to the distance metric, the *Euclidean distance* was used for experiments without PCA because all original attributes share the same domain, but the *Mahalanobis distance* was used when PCA was applied since it is able to remove the dominance of features with large variance in the transformed space.

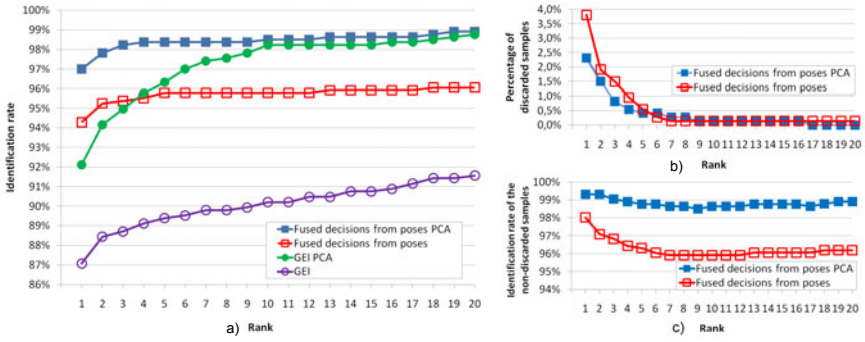


Fig. 4. a) Comparison of the methods with CMS. b) Percentage of discarded samples of the new method. c) CMS over non-discarded samples with the new method.

4.3 Analysis of Results

In this work, performance is assessed by *Cumulative Match Scores* (as in the FERET scheme [10]) and it is shown in Figure 4 a). This chart depicts the percentage of test sequences whose same-class nearest neighbour is among the top x matches. In this way, Rank 1 value is the classification correct rate (CCR), i.e., the percentage of subjects correctly classified at the first match.

Figure 4 a) reflects results of both aims of this work. It compares the performance of the new method with that of the usual GEI and, at the same time, it shows the effect of reducing dimensionality by PCA in both methods. It is very important to remark that results of the new method are considering the worst case (all ties are counted as errors). By analysing this figure, a first conclusion is that the new method performs better than the original GEI. It can be easily seen by comparing their CCRs (Rank 1). They are 97% and 94% for the new method using PCA or not respectively, and 92% and 87% in the same cases for the original GEI method. Such experimental results are probably supported by the fact that the combination of heterogeneous decisions from different poses provides more and better information than the original method. Finally, another conclusion is that both methods benefit from the use of PCA, not only in terms of performance improvement (3-7% higher for all ranks in both methods), but also in terms of reducing the computational complexity and dimensionality.

Figure 4 b) shows the percentage of samples rejected due to ties, while the Figure 4 c) depicts the identification rate over the non-discarded samples. From the analysis of these two charts together, very encouraging results are found. As the rank increases, the number of discarded samples (ties) drops drastically, while the success rate of the new method using PCA keeps almost steady and very close to 99%. It means that if the low percentage of ties is accepted (lower than 2.5%), the identification rate increases from about 97% (when ties are counted as errors, see Figure 4 a)) to about the 99% previously commented. When PCA is not applied, the classifier performance is slightly affected.

Results of methods using GEIs of size 64x64 are not shown in this paper because they follow a similar trend and they are more computationally demanding.

5 Conclusion

This paper proposes a new method for characterizing key poses of the gait pattern by individual GEIs. These pose-based representations implicitly capture biometric shape features (e.g. body part proportions) and dynamic descriptors (e.g. amount of arm swing) more accurately than usual GEI. Classification results obtained from combining individual pose-based decisions were better than those from the plain use of the original GEI method for the whole gait sequence. Besides, the new method is robust to noisy data and it is easy to understand.

However, it suffers from view dependence and it is limited to classify test sequences taken from roughly the same viewing angle as the training sequences. Our future research could be addressed to extend this method to mitigate the impact of such a problem.

References

1. Boulgouris, N., Hatzinakos, D., Plataniotis, K.: Gait recognition: a challenging signal processing technology for biometric identification. IEEE Signal Processing Mag. 22(6), 78–90 (2005)
2. Collins, R.T., Gross, R., Shi, J.: Silhouette-based human identification from body shape and gait. In: Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition, pp. 366–371 (2002)
3. Cutting, J., Kozlowski, L.: Recognizing friends by their walk: Gait perception without familiarity cues. Bull. Psychonomic Soc. 9(5), 353–356 (1977)
4. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. 28(2), 316–322 (2006)
5. Institute of Automation, Chinese Academy of Sciences: CASIA Gait Database (2005), `http://www.sinobiometrics.com`
6. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, Heidelberg (2002)
7. Kale, A., Rajagopalan, A.N., Sundaresan, A., Cuntoor, N., Roychowdhury, A., Krueger, V.: Identification of humans using gait. IEEE Trans. Image Process. 13(9), 1163–1173 (2004)
8. Lee, L., Grimson, W.: Gait analysis for recognition and classification. In: Proc. 5th IEEE Int'l. Conf. Automatic Face and Gesture Recognition, pp. 155–162 (2002)
9. Perry, J.: Gait Analysis: Normal and Pathological Function. SLACK Inc. (1992)
10. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 22(10), 1090–1104 (2000)
11. Yam, C., Nixon, M., Carter, J.: Automated person recognition by walking and running via model-based approaches. Pattern Recognit. 37(5), 1057–1072 (2004)
12. Yu, S., Tan, T., Huang, K., Jia, K., Wu, X.: A study on gait-based gender classification. IEEE Trans. Image Process. 18(8), 1905–1910 (2009)
13. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Proc. 18th Int'l. Conf. Pattern Recognition, pp. 441–444 (2006)

On-Line Classification of Data Streams with Missing Values Based on Reinforcement Learning*

Mónica Millán-Giraldo, Vicente Javier Traver, and J. Salvador Sánchez

Institute of New Imaging Technologies
http://init.uji.es
Universitat Jaume I, 12071, Castellón, Spain
{mmillan,vtraver,sanchez}@uji.es

Abstract. In some applications, data arrive sequentially and they are not available in batch form, what makes difficult the use of traditional classification systems. In addition, some attributes may lack due to some real-world conditions. For this problem, a number of decisions have to be made regarding how to proceed with the incomplete and unlabeled incoming objects, how to guess its missing attributes values, how to classify it, whether to include it in the training set, or when to ask for the class label to an expert. Unfortunately, no decision works well for all data sets. This data dependency motivates our formulation of the problem in terms of elements of reinforcement learning. The application of this learning paradigm for this problem is, to the best of our knowledge, novel. The empirical results are encouraging since the proposed framework behaves better and more generally than many strategies used isolatedly, and makes an efficient use of human effort (requests for the class label to an expert) and computer memory (the increase of size of the training set).

Keywords: Reinforcement learning; Active learning; Adaptive learning; Streaming data; Incomplete data; Imputation techniques; On-line classification.

1 Introduction

In many streaming data applications, where objects arrive one at a time, data may come with one or more missing attributes. Usually, the lack of some attributes causes statistical distortions in the data, which might degrade the predictive model and possibly lead to considerable reductions in the classification accuracy. Conventional techniques of dealing with missing data usually solve this problem by ignoring the incomplete objects (*projection*) or filling the missing attributes with values estimated from complete objects using statistical measures (*imputation*). Several approaches have been proposed in the literature [1–3].

In general, these results show that the application of methods of handling missing attributes might be helpful to improving the classification accuracy of incomplete data. However, in our experience [4], no single method or strategy is generally suitable for all

* This work has been supported in part by the Spanish Ministry of Education and Science under grants CSD2007–00018 (Consolider Ingenio 2010) and TIN2009–14205, and by Bancaixa under grant P1–1B2009–04.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 355–362, 2011.

the different data sets. Therefore, in this paper we developed a framework for solving the classification of incomplete data streams, which is inspired in part on reinforcement learning (RL) in order to combine different approaches of handling missing attributes. Besides, since in RL, the learning process is done by interacting with the environment, we developed the method with the intention that it could adapt to the peculiarities of each data set and even each object with respect to their missing values.

The theory of RL has been broadly used for applications of engineering control, robotics and planning. Few works of machine learning (ML) based on RL have been proposed in the literature; however, its interest has increased in the last decade, [5, 6]. Nevertheless, in spite of these few efforts to the best of our knowledge RL has not been explored in the classification of data with missing attributes. The reason behind of this may be that there is not an immediate formulation in RL under this learning context. In this paper, the motivation for the use of RL as a framework for classification of incomplete data, is to find a general solution which is adaptive to the classification problem so that it is independent of the peculiarities of the data set. In the proposed framework the algorithm, based on ideas from RL, learns a prediction model in the sense that it handles incomplete objects and classify them. For this purpose, the algorithm includes five techniques of handling missing attributes, which one of them is the projection approach and the remaining ones correspond to support vector regression, nearest neighbor, random and mean imputation methods. Additionally, we introduce an action which the algorithm may ask the expert for the class label of a given object.

2 Reinforcement Learning

In this paper, the proposed algorithm is based partially on Q-learning and, more specifically, on *temporal difference learning* (TD). This method was chosen because it is a simple technique and its structure is adequate for solving general optimization problems of learning [7]. Besides, it does not require a priori model of the environment or for the actions selection, then it must learn an action-value function (Q-function) as it interacts with the environment. These fit very well to our problem, where the characteristics of each data set are completely unknown.

2.1 Formulating the Problem of Classification of Incomplete Data Using RL

We propose an RL-based framework to handle the problem of data streams with incomplete and unlabeled objects [8]. In this work, each object arrives with only one missing attribute, being the most relevant. The attribute relevance is measured using the Jeffries-Matusita distance [9]. Here, states are defined for the objects and not for the data stream. This is not the single way to formulate the problem, since different states and actions definitions may lead to different solutions. In this work, each state represents the current state of the incoming object, that is, if the object is complete x or incomplete x^j (without the attribute j), and if it was classified, considering the obtained confidence in its classification. Four states were defined:

State 1. *Incomplete object:* the incoming object, x^j, has one missing attribute in j.

State 2. *Complete object:* the missing attribute of x^j has been filled with either imputation approach.
State 3. *Classified complete object:* x is classified with a confidence p higher or equal than a minimum predefined threshold, $th_{\min}$ (more on this later).
State 4. *Classified incomplete object:* the incomplete object x^j is classified, using the projection method, with confidence $p \geq th_{\min}$.

Actions defined in the algorithm are applied to the incoming object. They are related basically with (1) the methods of dealing with incomplete data, (2) classification, (3) whether to incorporate or not the incoming object into the training set and (4) returning to the initial state. Nine actions were included in the algorithm:

Action 1. *Projection:* it classifies the incomplete object using the projection approach, which objects from the training set, T, are projected to one dimension less.
Action 2. *SVR:* the missing attribute j is imputed using support vector regression (SVR) [4]. To this end, objects from T are used to build a new space using SVR. Then, x^j is mapped to this space to estimate the value for the missing attribute.
Action 3. 1*NN:* the missing attribute of x^j is filled with the attribute j of its nearest neighbor (1NN).
Action 4. *Random:* the missing attribute of x^j is imputed with a random value estimated in the range between the minimum and maximum values of the attribute j from the training set objects.
Action 5. *Mean:* the missing attribute of x^j is filled with the mean value obtained from attributes j of objects from T.
Action 6. *Expert intervention:* the expert provides the class label of the incoming object when the algorithm asks for it. The aid of the expert may be requested either for complete or incomplete objects, but it is only for the true class label and never for the missing attribute.
Action 7. *Classification:* it classifies x using the confidence based on its k-NN [4].
Action 8. *Insert x into T:* objects classified with $th_{\min} \leq p \leq th_{\max}$, where $th_{\max}$ is a maximum predefined threshold, are incorporated into T with the aim to insert only objects which may provide useful information.
Action 9. *Return to the initial state:* if in the classification $p < th_{\min}$ the algorithm returns to the initial state to repeat the process using a different method of dealing with missing attributes.

Actions may be applied depending on the state where the algorithm is. Allowed actions in each state are shown in Table 1. When the algorithm tries to perform any invalid action it is penalized. Additionally, whether the algorithm reaches the final goal, which is to classify the object with confidence $p \geq th_{\min}$, it is rewarded. Thus, the algorithm keeps adapting its Q-values as it goes learning. Details of parameter values, such as rewards and penalizations, were not here included due to the lack of space.

3 Experiments and Results

In this section we describe the experiments carried out for evaluating empirically the performance of the algorithm. Experiments were conducted as follows:

Table 1. Allowed actions in each state

State:	1	2	3	4
Actions:	1 to 6	6 and 7	8	9

Data sets: 18 real data sets were used in the experiment (see Table 2). We normalized the data sets in the range $[0, +1]$, in order to avoid the possible influence of the difference between attribute scales on the results.

Partitions: Data sets were split into two sets. One of them is for the training set with a size of $d \times c$, where d and c are the number of features and classes [10], respectively. The remaining objects are employed for the streaming data set. Objects for data stream were shuffled before each run with the aim to simulate an independent and identically-distributed sequence. Data partitions for the initial training and streaming data sets were the same for all methods, and objects were presented in the same order with the aim to avoid that the order of data affects in the comparison of performances.

Repetitions: Experiments were carried out ten times for all data sets, with different training and streaming data sets each time, this with the aim to have a general behavior for each data set.

Incomplete objects: An object with one missing attribute arrived to the system one at a time. The missing attribute chosen for experiments was the most relevant attribute, that means, the attribute with the most discriminative information.

Expert intervention: To prevent the algorithm from constantly asking the expert for a class label, a constraint is imposed so that at least n objects have to be processed before asking the expert again. In real applications n should be set according to the actual cost/benefit ratio. The arbitrary value $n = 7$ was used in these experiments.

Classification: It is performed according to the probabilities of its k-nearest neighbors, from the training set, of belonging to each class [4]. Objects classified with $th_{\min} \leq p \leq th_{\max}$ are incorporated into the training set with the aim to insert only objects which may provide useful information. Thresholds values were automatically estimated from training set objects for each run. The classification error is estimated as the number of misclassifications divided by the numbers of objects seen until the moment by the algorithm. Besides, the classification error was averaged over the 10 runs in order to have a single performance for each data set.

Table 2. Data sets used in the experiments

Repository	Ref.	Data sets
UCI	[11]	iris, wine, sonar, thyroid, heart, liver, voice9, wbc, australian, pima, vowel, german
Ripley	[12]	crabs
Library	[13]	laryngeal1, intubation, spect, laryngeal2
Private		Images of pieces of kerogen extracted from microscope images of palynomorphs

3.1 Analysis of Results

Results obtained with the proposed RL-based framework were compared with five techniques of dealing with incomplete data: projection, SVR, 1NN, random and mean imputations techniques. As the baseline, we employed results obtained by classifying the same incoming objects but with all their attributes, which is called *Complete*.

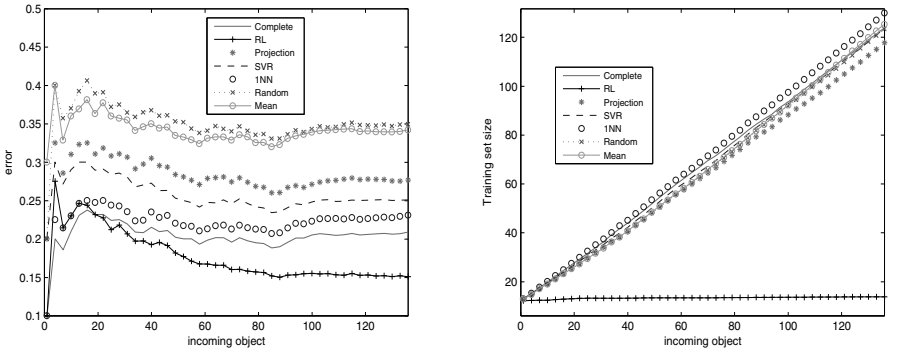
Table 3 shows the final classification error (i.e. the accumulated error after all objects in the stream arrived), for each method (columns) and for the 18 real data sets (rows). Error values highlighted in bold indicate the best final error obtained for each data set. Interestingly, the method based on RL outperforms the other techniques (including the baseline case which has no missing attribute!) in 16 out of the 18 data sets. This result points to the general good behavior of the algorithm: whereas the other methods work the best for *some* data sets but not for some others, the proposed algorithm works the best for *almost all* the data sets. On the other hand, the percentage of objects (last column) for which the expert provided the true class label is relatively low (around 4–11%, depending on the data set). This suggests the ability of the algorithm of being sparing with the expert knowledge while benefiting from it.

Table 3. Final classification error

Data set	Complete	RL	Projection	SVR	1NN	Random	Mean	Expert [%]
iris	0.209	**0.150**	0.277	0.251	0.231	0.350	0.344	6.23
wine	0.050	**0.024**	0.043	0.046	0.049	0.045	0.044	9.64
crabs	0.425	**0.403**	0.435	0.441	0.444	0.460	0.476	7.02
sonar	0.232	**0.199**	0.238	0.235	0.231	0.239	0.241	10.57
laryngeal1	0.193	**0.162**	0.193	0.194	0.193	0.211	0.192	11.38
thyroid	0.244	**0.198**	0.258	0.253	0.250	0.325	0.291	5.05
intubation	0.359	**0.273**	0.405	0.381	0.402	0.397	0.405	10.49
heart	0.465	**0.415**	0.464	0.461	0.467	0.463	0.463	9.37
liver	0.452	**0.432**	0.453	0.455	0.452	0.456	0.456	8.14
spect	0.291	**0.251**	0.302	0.293	0.298	0.283	0.305	8.05
voice9	0.629	**0.603**	0.627	0.630	0.629	0.659	0.636	6.89
wbc	0.057	**0.038**	0.061	0.060	0.061	0.073	0.072	7.35
palynomorphs	0.186	**0.160**	0.201	0.187	0.198	0.222	0.206	10.85
australian	0.165	**0.142**	0.166	0.166	0.166	0.166	0.166	10.98
laryngeal2	0.067	**0.060**	0.067	0.067	0.066	0.068	0.065	3.80
pima	**0.336**	0.357	0.345	0.346	0.347	0.353	0.349	10.09
vowel	**0.444**	0.478	0.490	0.471	0.477	0.599	0.559	6.52
german	0.378	**0.343**	0.395	0.395	0.408	0.401	0.400	10.48

To present an example of the results, we choose one case when the algorithm works well (iris data set) which represent the general behavior for the most data sets (16 out of 18) and another one which show the worst case (pima data set). As it can be seen in Fig. 1(a), the classification error for iris data set has a decreasing trend; which could suggest that the algorithm is learning which methods of handling missing attributes are more appropriate for this data set. Additionally, the algorithm is incorporating only few objects to the training set, Fig. 1(b), unlike what occurs with the other methods. The cause of the good behavior is probably due to that only those objects that contain useful information could be being inserted into the training set, and they could be being used in subsequent classifications.

The percentage of objects that used the actions of handling incomplete objects is presented in Table 4. As it can be seen, the technique more employed for iris data set



(a) (b)

Fig. 1. Results for iris data set. (a) Classification error computed for ten runs, when the missing attribute was the most relevant. (b) Training set sizes of each technique.

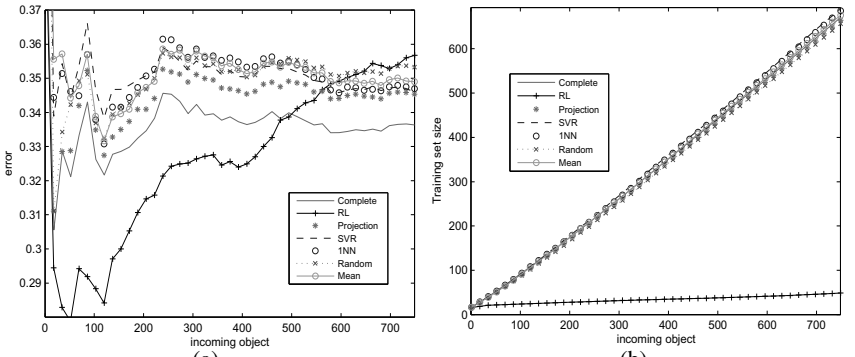
was the 1NN imputation. On the other hand, in Fig. 1(a), among the projection and imputation methods, the 1NN is the curve with the best performance. Therefore, it may be that the algorithm has identified this technique as the most suitable for this data set. Probably, the use of the expert aid is influencing the results for iris data set. However, the percentage of objects that used the expert action was only about 6.23% (Table 3).

Table 4. Percentage of objects that employs each action of handling incomplete data

	Projection	**SVR**	**1NN**	**Random**	**Mean**
iris	2.75	16.96	**55.00**	10.58	14.57
pima	**28.21**	20.92	19.60	16.28	14.91

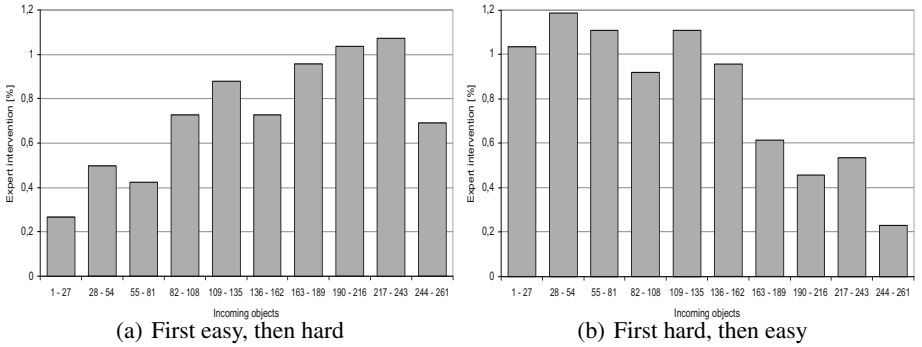
For the pima data set, the best performance is obtained by projecting the objects (Fig. 2(a)). However, by design, objects which are projected are decided not to be included in the training set (just because they are incomplete). As a result, even if the true class label is provided by the expert, this information cannot be exploited for the classification of subsequent objects, which may explain the poor performance of the system in this scenario. Therefore, this design issue should be reconsidered in our future work.

A controlled experiment is carried out with the aim to verify that the RL algorithm used the expert action when it is supposed to, i.e. when an object is harder to classify. For that purpose, we sort the objects by difficulty of classification before they are inserted into the system. To make such a sorting, we organized them according to their posterior probabilities obtained by classifying them using the remaining objects and the confidence of their 5 nearest neighbors. Figure 3 shows, for spect data set, how much the expert was asked for the class label for the spect data set. In this figure, input objects were grouped in ten blocks by intervals of time. Figure 3(a) shows the results when the easiest objects for classifying arrived first and then the hardest, (b) otherwise. As can be seen in Fig. 3(a), the percentage of objects that make use of the expert action is



(a) (b)

Fig. 2. Results for pima data set. (a) Classification error computed for ten runs, when the missing attribute was the most relevant. (b) Training set sizes of each technique.



(a) First easy, then hard (b) First hard, then easy

Fig. 3. Histogram of the percentage of objects that the expert has classified for spect data set, when the easiest (a) or hardest (b) objects arrived first. Objects are grouped by intervals of time.

increasing, whereas in Fig. 3(b) this percentage is decreasing. This is because the more difficult to classify objects are the more the algorithm requests the expert, and vice versa. However, this interesting behavior is not observed for most of the other data sets.

4 Conclusions

A reinforcement learning-based approach has been devised for on-line learning and classification of incomplete objects in data streams. The proposed algorithm has some interesting properties. First, it is quite *general*: by automatically exploring a number of imputation/classification techniques which are available, the approach is able of outperforming each of these techniques being used isolatedly. Second, it is *memory effective*: by selectively choosing (guessing) the most representative objects, the size of the training set may be kept stable along time, in contrast to more naïve procedures

which blindly include all incoming samples. Third, it makes a *modest usage of the expert knowledge*: by tapping into expert-provided class labels only on a few objects (less than 11%), the classification performance is significantly boosted.

Despite these good qualities, the algorithm has still some limitations that deserve further work. For instance, the exploratory component of the reinforcement learning implies trying many costly actions (such as imputation and classification) for each incoming object. Another interesting goal is designing and testing alternative formulations of the problem in terms of reinforcement learning.

References

1. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, Chichester (1987)
2. Ding, Y., Simonoff, J.S.: An investigation of missing data methods for classification trees applied to binary response data. J. of Machine Learning Res. 11, 131–170 (2010)
3. Farhangfar, A., Kurgan, L., Dy, J.: Impact of imputation of missing values on classification error for discrete data. Pattern Recognitions 41(12), 3692–3705 (2008)
4. Millán-Giraldo, M., Sánchez, J.S., Traver, V.J.: Exploring early classification strategies of streaming data with delayed attributes. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009. LNCS, vol. 5863, pp. 875–883. Springer, Heidelberg (2009)
5. Vogiatzis, D., Stafylopatis, A.: Reinforcement learning for rule extraction from a labeled dataset. Cognitive Systems Research 3(2), 237–253 (2002)
6. Langford, J., Zadrozny, B.: Relating reinforcement learning performance to classification performance. In: Proc. of the Intl. Conference on Machine Learning, pp. 473–480 (2005)
7. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
8. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge (2006)
9. Bruzzone, L., Roli, F., Serpico, S.B.: An extension of the Jeffreys Matusita distance to multiclass cases for feature selection. IEEE Transactions on Geoscience and Remote Sensing 33(6), 1318–1321 (1995)
10. Nagy, G.: Classifiers that improve with use. In: In Proc. Conf. on Pattern Recognition and Multimedia, pp. 79–86 (2004)
11. Frank, A., Asuncion, A.: UCI Machine Learning Repository
12. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
13. Library: Real medical data sets, `http://www.bangor.ac.uk/~mas00a/activities/real_data.htm`

Opponent Colors for Human Detection

Rao Muhammad Anwer, David Vázquez, and Antonio M. López

Computer Vision Center and Computer Science Dpt.,
Universitat Autònoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{muhammad,david.vazquez,antonio}@cvc.uab.es -- www.cvc.uab.es/adas

Abstract. Human detection is a key component in fields such as advanced driving assistance and video surveillance. However, even detecting non-occluded standing humans remains a challenge of intensive research. Finding good features to build human models for further detection is probably one of the most important issues to face. Currently, shape, texture and motion features have deserve extensive attention in the literature. However, color-based features, which are important in other domains (*e.g.*, image categorization), have received much less attention. In fact, the use of RGB color space has become a kind of choice *by default*. The focus has been put in developing first and second order features on top of RGB space (*e.g.*, HOG and co-occurrence matrices, resp.). In this paper we evaluate the opponent colors (OPP) space as a biologically inspired alternative for human detection. In particular, by feeding OPP space in the baseline framework of Dalal *et al.* for human detection (based on RGB, HOG and linear SVM), we will obtain better detection performance than by using RGB space. This is a relevant result since, up to the best of our knowledge, OPP space has not been previously used for human detection. This suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., also on top of OPP space, *i.e.*, as we have done with HOG in this paper.

1 Introduction

Human detection is a key component in fields such as advanced driving assistance [1–3] and video surveillance [4–6]. Detecting humans in images is quite challenging because of their intra-class variability, the diversity of backgrounds and the different image acquisition conditions. Even detecting non-occluded humans that are standing, is still a hot topic of research. In order to improve human detection results we can focus on *classification*, *i.e.*, on building a classifier that given an image window decides if it contains a human or not. Nowadays, most successful classification processes for human detection follow the learning-from-examples paradigm [1, 2]. For instance, Dalal *et al.* [7] proposed a holistic classifier that relies on histograms of oriented gradients (HOG) as features and linear support vector machines (linear SVM) as learning algorithm, which still remains as a competitive baseline method for comparison with new human classifiers [2, 8].

Finding good features for developing a human classifier is a major key for its success. Focusing on human appearance, different sets of features try to exploit

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 363–370, 2011.

(combinations of) cues such of shape and texture [4]. However, although color information deserves special attention in domains such as segmentation and category recognition [9, 10], it has not been explored in deep for human detection. In fact, the baseline classifier of Dalal *et al.* [7] uses standard RGB. In particular, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. Dalal *et al.* reported that similar results were obtained using LAB space. This approach has been the common way of using color for human detection since then [4, 11–15] and, as a matter of fact, it has been considered as pretty similar to the use of the image intensity in cases where color information was not available [2].

Human beings do not rely on long (L), middle (M) and short (S) wavelength channels (RGB-like) separately for color perception. In order to increase subsistence, evolution provided the human retina with ganglion cells that combine L, M and S channels to work in *opponent-colors-space* mode for enhancing the visual detection of events of interest as well as compressing the color information of L, M and S *acquisition cells* [16–18]. Such compressed color information is sent through the optical nerve to the brain for later decompression and interpretation. Accordingly, in this paper we evaluate the opponent colors (OPP) space as a biologically inspired alternative for human detection. In particular, by feeding OPP space in the baseline framework of Dalal *et al.*, we will obtain better detection performance than by using RGB space. Besides, this finding is reinforced by the work in [10], where K. van de Sande *et al.* show that applying a scale invariant feature transform (SIFT [19]) to OPP space is the best *a priori* option in the context of image category recognition. Note, that HOG is a SIFT inspired descriptor.

For our current work, as Dalal *et al.*, we have used the so-called INRIA human dataset. This dataset contains color images and still is widely used for benchmarking. To support our claim we not only present so-called *per window* evaluation on INRIA human dataset, but also *per image* evaluation as highly recommended in [8].

We argue that altogether is a relevant result since, up to the best of our knowledge, OPP space was not previously used for human detection. Thus, with the aim of enriching feature space for human classifiers, our work suggests that in the future it could be worth to compute co-occurrence matrices, self-similarity features, etc., on top of OPP space, *i.e.*, as we have done here with HOG.

The rest of the paper is organized as follows. In section 2 we define the OPP space. In section 3 we summarize the details of the human detector developed for our experiments. In section 4 we draw the experiments and discuss the corresponding results. Finally, section 5 summarizes the main conclusions.

2 Opponent Colors Space

In the late 19th century, E. Hering noted that the four hues red, green, yellow and blue are fundamental in the sense that they cannot be described as mixtures

of other hues. Then, he stated that there were three types of photo receptors: white-black, yellow-blue and red-green [16]. Nowadays we know that there are not such *image acquisition cells* in human vision. However, Hering was right in postulating the *computation of opponent colors* (*i.e.*, red *vs* green and yellow *vs* blue) in human color vision.

Contemporary science of human vision states that color photo receptors at the retina (*i.e.*, cones) are sensitive to long (L-cone), middle (M-cone) and short (S-cone) wavelengths. A single cone is color blind since its activation depends on both the wavelengths and intensity of the stimulus. A comparison of the signals from different classes of photo receptors is therefore the most basic computational requirement of a color vision system. The existence of cone-opponent retinal ganglion cells that perform such comparisons is well established for human vision.

In particular, opponent process theory postulates that yellow-blue and red-green information is represented by two parallel channels in the visual system that combine cone signals differently. It is now accepted that at an early stage in the red-green opponent pathway, signals from L and M cones are opposed, and in the yellow-blue pathway signals from S cones oppose a combined signal from L and M cones [17]. In addition, there is a third luminance or achromatic mechanisms in which retinal ganglion cells receive L- and M- cone input. Thus, L, M and S belong to a first layer of the retina whereas luminance and opponent colors belong to a second layer of it, forming the basis of chromatic input to the primary visual cortex. Note also that this mechanism is not random since human color vision evolved for increasing the probability of subsistence [18].

Seeing the RGB space used for codifying color in digital images as the LMS color space of the first layer of human retina, we can also compute an opponent colors (OPP) space as follows [10]:

$$\begin{aligned} \text{red-green} &: O_1 = (R-G)/\sqrt{2} \ , \\ \text{yellow-blue} &: O_2 = ((R+G)-2B)/\sqrt{6} \ , \\ \text{luminance} &: O_3 = (R+G+B)/\sqrt{3} \ , \end{aligned} \tag{1}$$

for R, G and B running on values in $[0, 1]$.

3 Human Detector

A *human detector* is composed of a *human classifier* learnt from a training set by using specific *features* and a *learning machine*. With this classifier we *scan a given image* looking for humans. Since multiple detections can be produced by a single human, we also need a mechanism to *select the best detection*. The procedures we use for feature extraction, machine learning, scanning the images, as well as selecting the best detection from a cluster of them, are briefly reviewed in this section.

Human classifier. We follow the settings suggested by Dalal *et al.* for computing HOG features and learning the human classifier using a linear SVM. Such approach remains competitive [2, 8] and, in fact, is the core from which many

new proposals are developed [4, 14]. However, Dalal *et al.* as well as in many following works [4, 11–15], compute HOG on top of RGB space. More specifically, gradient information is computed individually for each color channel and then, at each pixel, only the gradient information corresponding to the maximum magnitude among the RGB channels is used for computing the HOG. We argue that the *max* operation basically is throwing away the color information, *i.e.*, only some sort of luminance contrast is captured by HOG. Accordingly, we propose to replace the features considered by Dalal *et al.* so that color information is also captured.

Our proposal is twofold. First, we remove the *max* operation, *i.e.*, HOG are applied to each color channel separately and, then, the corresponding feature vectors are concatenated to form a single feature vector. Such three-channels HOG are then the input that the linear SVM will use to learn the human classifier. Second, we propose the use of OPP space instead of RGB one. We will see that both ideas are essential to improve human classification performance.

Image scanning. In order to perform multi-scale human detection we use the extended *pyramidal sliding window* strategy as proposed in Dalal's PhD [20]. The original image is scaled by a factor s^i to obtain the image corresponding to the pyramid level i. Then, given a pyramid level, we must shift the search window along the horizontal and vertical directions with a given stride $\Delta = (\delta_x, \delta_y)$ pixels. The smaller the s and Δ parameters, the finer the sliding window search. Using a finer search we can expect better detection performance. However, this is to the expense of a higher processing time. Dalal set $s = 1.2$ and $\Delta = (8, 8)$. In our work we found $s = 1.05$ and $\Delta = (4, 4)$ pixels a better tradeoff between processing time and detection performance.

Select the best detection. In multi-scale human detection a single person can be detected several times at slightly different positions and scales. Since a unique detection per human is desired, multiple overlapped detections should be grouped by a clustering or *non-maximum-suppression* procedure. In this case, we don't follow the Dalal's proposal in [20]. Instead, we rely on the iterative confidence- and overlapping clustering approach of Laptev [21], which is a simpler and faster technique than Dalal's proposal and yields similar results.

4 Experiments

4.1 Human Dataset

We rely on the widely used INRIA person dataset of color images for our experiments. This dataset shows a wide range of human variations in pose, clothing, occlusions as well as complex backgrounds. Moreover, the dataset is divided in separated sets of null intersection for training and testing.

The training set contains 2,416 *positive* samples consisting in image windows (original and vertical mirror), each one containing a person framed by certain

Fig. 1. Positive (humans) and negative (background) windows from INRIA dataset

amount of background. Positives are of the same size (*canonical detection window*), although many of them come from an isotropic down scaling. We term this set of windows as $\mathcal{W}_+^{\text{train}}$. For collecting *negative* samples, *i.e.*, image windows that do not contain persons, there are available 1,218 human-free images. We term this set of images as $\mathcal{I}_-^{\text{train}}$. The testing set consists of: (1) $\mathcal{I}_-^{\text{test}}$: 453 human-free images; (2) $\mathcal{I}_+^{\text{test}}$: 288 images containing labelled persons (ground truth); (3) $\mathcal{W}_+^{\text{test}}$: 1,126 positives analogous to the ones in $\mathcal{W}_+^{\text{train}}$ after cropping and mirroring the ground truth of $\mathcal{I}_+^{\text{test}}$.

4.2 Training

We use the standard training procedure for the INRIA dataset [7, 20]. First, we collect random negative windows from the images in $\mathcal{I}_-^{\text{train}}$ (10 windows per image to have 12,180 negatives) and down scale them to the size of the canonical detection window; let's call this set of windows $\mathcal{W}_-^{\text{train}}$. Then, given the sets $\mathcal{W}_+^{\text{train}}$ and $\mathcal{W}_-^{\text{train}}$, we compute the HOG of such labelled windows on top of the desired color space, and learn the human classifier using the linear SVM. Finally, we run the corresponding human detector on $\mathcal{I}_-^{\text{train}}$ in order to follow the recommended *bootstrapping* technique, *i.e.*, to append the set $\mathcal{W}_-^{\text{train}}$ with *hard negative windows* and re-train the human classifier. We apply two bootstrapping iterations. Figure 1 shows positive and negative training samples.

4.3 Evaluation

In our experiments we use two widely extended methods of evaluation: *per window* and *per image*. In per window evaluation we asses the results of the human classifier when applied to the $\mathcal{W}_+^{\text{test}}$ and the images in $\mathcal{I}_-^{\text{test}}$. Let $\text{P}^{\#}$ be the cardinality of $\mathcal{W}_+^{\text{test}}$, and let's term as P^{TP} the number of elements in $\mathcal{W}_+^{\text{test}}$ classified as *humans* (*i.e.*, total of so-called true positives). Let $\text{N}^{\#}$ be the total number of windows processed by applying the pyramidal sliding window technique to the images in $\mathcal{I}_-^{\text{test}}$ (for each image more than one million of windows are usually processed), and let's term as N^{FP} the number of such windows classified as *humans* (*i.e.*, total of so-called false positives). Then, we define the per window detection rate as $\text{DR}^{\text{pw}} = \text{P}^{\text{TP}}/\text{P}^{\#}$, $\text{DR}^{\text{pw}} \in [0, 1]$. Corresponding miss rate is defined as $\text{MR}^{\text{pw}} = 1 - \text{DR}^{\text{pw}}$. Analogously, we define the false positives per window as $\text{FP}^{\text{pw}} = \text{N}^{\text{FP}}/\text{N}^{\#}$, $\text{FP}^{\text{pw}} \in [0, 1]$. We remark that for any given image window, the human classifier returns a real value that we threshold with a fixed value t in order to classify the window as of type *human* or *non-human*.

Thus, $\mathrm{DR}^{\mathrm{pw}}$ and $\mathrm{FP}^{\mathrm{pw}}$ are functions of t. This allows to plot evaluation curves $\mathrm{E}^{\mathrm{pw}}(t) = (\mathrm{FP}^{\mathrm{pw}}(t), \mathrm{MR}^{\mathrm{pw}}(t))$ (so-called ROCs) that show the tradeoff between the miss rate and the false positives per window for each t.

However, some researchers show that it may be more realistic to follow per image evaluation [8]. In this case, not only the human classifier is evaluated but the whole human detector. In particular, the sets $\mathcal{I}_+^{\mathrm{test}}$ and $\mathcal{I}_-^{\mathrm{test}}$ are seen as a single set of images, $\mathcal{I}^{\mathrm{test}}$, where the human detector is run. Then, the set of detections is compared with the ground truth for counting how many of such detections are true positives (T^{TP}) and how many are false positives (T^{FP}). If $\mathrm{I}^{\#}$ is the cardinality of $\mathcal{I}^{\mathrm{test}}$ and $\mathrm{H}^{\#}$ the number of labelled humans in $\mathcal{I}_+^{\mathrm{test}}$, then we can define the per image detection rate as $\mathrm{DR}^{\mathrm{pi}} = \mathrm{T}^{\mathrm{TP}}/\mathrm{H}^{\#}$ ($\mathrm{DR}^{\mathrm{pi}} \in [0, 1]$; per image miss rate $\mathrm{MR}^{\mathrm{pi}} = 1 - \mathrm{DR}^{\mathrm{pi}}$) and the false positives per image as $\mathrm{FP}^{\mathrm{pi}} = \mathrm{T}^{\mathrm{FP}}/\mathrm{I}^{\#}$. In order to determine if a detection overlaps sufficiently with a labelled human of $\mathcal{I}_+^{\mathrm{test}}$ we follow the so-called PASCAL criteria [8] (also for bootstrapping during training). Now, analogously to $\mathrm{E}^{\mathrm{pw}}(t)$ we can define the evaluation curve $\mathrm{E}^{\mathrm{pi}}(t) = (\mathrm{FP}^{\mathrm{pi}}(t), \mathrm{MR}^{\mathrm{pi}}(t))$; $\mathrm{FP}^{\mathrm{pi}}(t)$ can be greater than one.

4.4 Devised Experiments

We train four types of classifiers: RGB-*max*; RGB-*3ch*; OPP-*max* and OPP-*3ch*. The OPP *vs* RGB refers to the used color space. The *3ch* stands for computing HOG for each color channel separately and then concatenate the three feature vectors into a single one. The *max* stands for computing HOG by taking into account, at each pixel, only the gradient of highest magnitude among the color channels, *i.e.*, the usual approach introduced by Dalal *et al.*

Since collecting negatives during training involves a random selection, obtained classifiers can vary from train to train. Therefore, for each type of classifier we repeat the training and further evaluation five times. This gives five curves per classifier (20 curves), thus, we condense the results for each classifier in the respective mean ± standard deviation curves for both per window ($\mathrm{E}^{\mathrm{pw}}(t)$) and per image ($\mathrm{E}^{\mathrm{pi}}(t)$) evaluation. Figure 2 summarizes the obtained results.

4.5 Discussion

We point out two main observations: (1) OPP-*3ch* clearly outperforms RGB-*3ch/max*; (2) the *max* operation throws away the color information. Let us argue these observations. Per image and per window evaluation show that OPP-*3ch* outperforms RGB-*3ch/max*, especially at the usual points of interest, *i.e.*, $\mathrm{FP}^{\mathrm{pw}} = 10^{-4}$ and $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$. At $\mathrm{FP}^{\mathrm{pw}} = 10^{-4}$ OPP-*3ch* has an average miss rate of 0.0358, while for RGB-*3ch* is 0.0806 and for RGB-*max* 0.1111. At $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$ OPP-*3ch* has an average miss rate of 0.1391, while for RGB-*3ch* is 0.1981 and for RGB-*max* 0.1889. Moreover, the *max* operation removes the difference between RGB and OPP spaces. Besides, per image evaluation shows

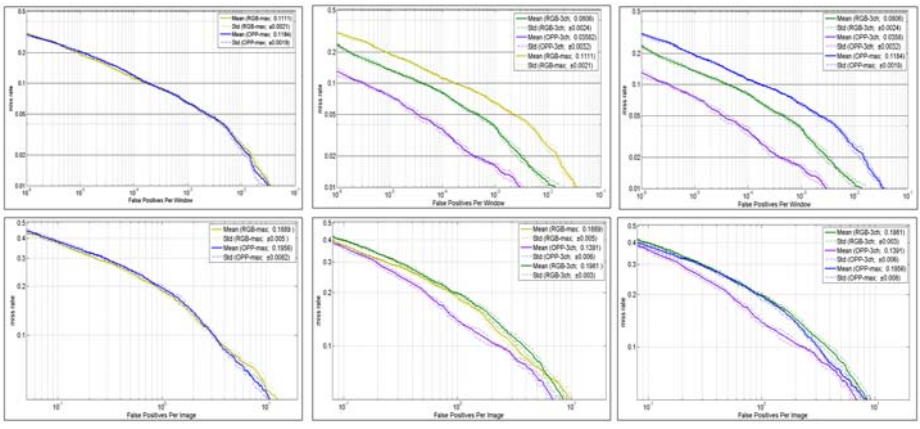


Fig. 2. Per window (top) and per image (bottom) evaluation using logarithmic scales. Values at usual points of interest are included, *i.e.*, 10^{-4} FPPW and 10^{0} FPPI, resp.

that both the *3ch* and the *max* configurations are similar for the RGB case, but quite different for OPP, where *3ch* clearly wins. For instance, the average miss rate of OPP-*3ch* at $\mathrm{FP}^{\mathrm{pi}} = 10^{0}$ is 0.1391 while for OPP-*max* it is 0.1956.

5 Conclusions

In this paper we have explored the use of the biologically inspired opponent color space as the basis to obtain better features for human detection. In particular, we have seen that by feeding such a color space in the HOG+LinearSVM baseline classifier, we obtain better results than by following the common practice of using RGB color space. This conclusion is based on per window and per image evaluation over the widely used INRIA dataset. We think that this is a relevant finding, because, up to the best of our knowledge, opponent color was not previously used for human detection. Moreover, co-occurrence matrices, self-similarity features, etc., could be computed in the future on top of opponent color space in order to further improve human detection results.

Acknowledgments. This work was supported by the Spanish Government (projects TRA2007-62526/AUT, TRA2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018)).

References

1. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(7), 1239–1258 (2010)
2. Enzweiler, M., Gavrila, D.: Monocular pedestrian detection: survey and experiments. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(12), 2179–2195 (2009)

3. Gandhi, T., Trivedi, M.M.: Pedestrian protection systems: issues, survey, and challenges. IEEE Trans. on Intelligence Transportation Systems 8(3), 413–430 (2007)
4. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: Int. Conf. on Computer Vision, Kyoto, Japan (2009)
5. Jones, M., Snow, D.: Pedestrian detection using boosted features over many frames. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
6. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. Int. Journal on Computer Vision 63(2), 153–161 (2005)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA (2005)
8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: IEEE Conf. on Computer Vision and Pattern Recognition, Miami Beach, FL, USA (2009)
9. Álvarez, J.M., Gevers, T., López, A.M.: Learning photometric invariance for object detection. Int. Journal on Computer Vision 90(1), 45–61 (2008)
10. van de Sande, K., Gevers, T., Snoek, C.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
11. Oliveira, L., Nunes, U., Peixoto, P.: On exploration of classifier ensemble synergism in pedestrian detection. IEEE Trans. on Intelligence Transportation Systems 11(1), 16–27 (2010)
12. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: Int. Conf. on Computer Vision, Kyoto, Japan (2009)
13. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, Miami Beach, FL, USA (2009)
14. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
15. Leibe, B., Cornelis, N., Cornelis, K., Gool, L.V.: Dynamic 3D scene analysis from a moving vehicle. In: IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, USA (2007)
16. Hering, E.: Outlines of a theory of the light sense (translated by L.M. Hurvich and D. Jameson). Harvard University Press, Cambridge (1964)
17. Krauskopf, J., Williams, D.R., Heeley, D.W.: Cardinal directions of color space. Vision Research 22(9), 1123–1132 (1982)
18. Mollon, J.D.: "tho' she kneel'd in that place where they grew... " the uses and origins of primate colour vision. Journal of Experimental Biology 146(1), 21–38 (1989)
19. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision, Kerkyra, Greece (1999)
20. Dalal, N.: Finding people in images and videos. PhD Thesis, Institut National Polytechnique de Grenoble / INRIA Rhône-Alpes (2006)
21. Laptev, I.: Improving object detection with boosted histograms. Image and Vision Computing 27(5), 535–544 (2009)

Automatic Detection of Facial Feature Points via HOGs and Geometric Prior Models

Mario Rojas Quiñones[1], David Masip[1,2], and Jordi Vitrià[1,3]

[1] Computer Vision Center, Universitat Autònoma de Barcelona
[2] Universitat Oberta de Catalunya
[3] Dept. de Matemàtica Aplicada i Anàlisi
Universitat de Barcelona
mrojas@cvc.uab.es, dmasipr@uoc.edu, jordi.vitria@ub.edu

Abstract. Most applications dealing with problems involving the face require a robust estimation of the facial salient points. Nevertheless, this estimation is not usually an automated preprocessing step in applications dealing with facial expression recognition. In this paper we present a simple method to detect facial salient points in the face. It is based on a prior Point Distribution Model and a robust object descriptor. The model learns the distribution of the points from the training data, as well as the amount of variation in location each point exhibits. Using this model, we reduce the search areas to look for each point. In addition, we also exploit the global consistency of the points constellation, increasing the detection accuracy. The method was tested on two separate data sets and the results, in some cases, outperform the state of the art.

Keywords: Salient Point Detection, Histogram of Oriented Gradients, Ensemble learning.

1 Introduction

In the context of human computer interfaces, the analysis and processing of human faces is a key aspect of the performance of such systems. While automatic face detection has been rather successfully accomplished by algorithms like the presented by Viola & Jones [12], the same may not necessarily be said about detecting automatically specific points of interest in the face. Although this task has been largely studied in the object recognition literature [8] is not unusual that works on facial expression recognition neglect the automatic the detection of salient points, therefore algorithms are usually experimentally validated using manually annotated images [1]. Applications dealing with problems involving the face require a robust estimation of the facial salient points (such as the tip of the eyebrows). Recent approaches consider the detection of facial elements such as eyes or mouth as references in the face [9], yet the variation on the shape and appearance of these makes it difficult to learn models for applications involving expression analysis for example. Furthermore the variance in appearance of many of the points (e.g. eyebrows) results in a high variability when it comes to labeling

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 371–378, 2011.

(even by human annotators) exact locations, thus making the problem even more complex.

In [11,13] a classification of the methods for facial feature point detection is proposed: (i) texture-based methods, where the local neighborhood is used to locate points [10], and (ii) shape-based methods that use all feature points in the face [2] to model the shape. In the context of texture, Vukadinovik and Pantic [13] proposed to learn each salient point as a category in a multi-class classification framework. The authors used a set of Gabor jet filters to extract a feature vector from each salient point. A GentleBoost classifier is trained with the extracted samples, resulting in a robust point detection strategy, yet the extraction of the Gabor descriptors can be computationally expensive. In [11] Valstar et al. presented a method based on a combination of support vector regressors and Markov Random Fields to detect facial salient points using Haar-based descriptors over local patches, aiming to reduce the search time and make it robust to variations on appearance and rotations. In [7] Kozakaya et al. presented a method that estimates facial feature points by a concentration of directional vectors from sampled points. The vectors are weighted according to the outcome of a nearest neighbor search between the pattern learned from the sample point and the model, both of which use the HOG algorithm as a descriptor. Although their approach yields good results, it still requires a rather large look-up table to determine the nearest neighbor.

In this paper we propose a simple yet promising approximation to the task of facial salient point detection. Our system intends to view the detection problem as a classification one. It learns the model for each fiducial point using the HOG algorithm [3] to compute the descriptor over a local neighborhood and trains a GentleBoost classifier [4]. Subsequently, it learns the distribution of points in the reference frame of the face bounding box from the training set. During test it uses as a basis the face localization and a pair of reference points. This information is used to adjust the model of the spatial location and center the search areas, reducing the computational cost by limiting the amount of points to test. Inspite of the simplicity of the strategy used, the results obtained outperformed -in many cases, the ones obtained by state of the art methods.

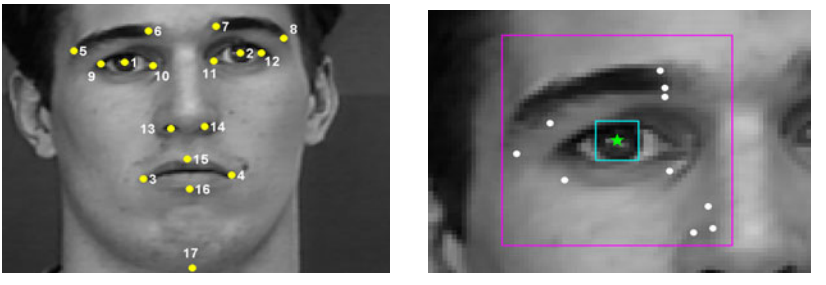
2 Point Detection

2.1 Learning the Facial Point Models

The presented method exploits the strengths of a state of the art algorithm for object description and takes into account the global distribution consistency to locate more reliably the facial feature points. Given that the aim is to formulate the detection problem into a binary classification task, a supervised learning scheme is adopted consisting of a training and a evaluation phase. During training the annotations of the locations of the feature points are used to center windows from which the models will be learned. Positive and negative examples are extracted by the following procedure: once the window around the feature point is centered, a patch around the coordinates of the current feature point is

passed to the descriptor generator to render the positive example. The negative patch examples are generated by randomly sampling the window region with the constraint that the adjacent neighborhood around the target feature point coordinates is left empty. This reduces the distortion effect/noise (with respect to the descriptor generated) that similar patches near the true interest point can create in the training data.

Figure 1 shows the 17 points extracted and a training region with the location of the interest point, the void area from which no negative examples are extracted and the locations of points that represent the center of negative patch examples.



(a) (b)

Fig. 1. 1(a): Point model of the 17 salient points, 1(b): Example of training region. The green star represents the center of the positive example and the white circles the centers of negative examples. The cyan square limits the area inside of which no negative examples are extracted.

The algorithm used to generate the descriptor is the Histogram of Oriented Gradients - HOG, proposed by Dalal and Trigs [3], where the authors highlight that their method is well suited to robustly extract features for visual object recognition.

The HOG descriptor can be computed as follows. The gradient of the image is computed and the phase is quantized according to a predefined number of orientation intervals, which will represent the bins in the histogram. Thereafter the image is divided in small regions called *cells* from which the orientation histogram is built by votes of the quantized orientation of each pixel. These votes are weighted by the magnitude of the gradient for each pixel. Subsequently cells are grouped in *blocks* which are the normalization units of the algorithm. This normalization constitutes an important part of the algorithm, because it represents a smoothing factor and limits the effect of the variations of the gradient in local areas due to illumination and object/background contrast. Finally the descriptor is created by the concatenation of the block-normalized histograms of all the cells.

The descriptor is tuned mainly by four parameters namely the number of orientation bins, the size of the cells, the size of the blocks and the overlap factor. These parameters change the resolution of the grid thus changing the

quality of the descriptor, making the selection of their values important to a well adjusted description of the object. Other parameters include the range and sign of the orientations to consider and the normalization rule. These two were set to the values suggested by the authors to perform the best. Implementation values for the parameters are discussed in section 3. Figure 2 shows an example of the HOG descriptor visualization. Once the descriptors for all the training samples per point are computed, a GentleBoost classifier [4] is trained.

2.2 Prior Distribution Model

In order to take advantage of the strength of the descriptor and to reduce the computational cost of the system, the search for the feature points locations needs to be bounded. This is achieved using prior information about the locations of the points from the data available in the training set. Furthermore, this model takes into account the global distribution consistency to locate more reliably the facial feature points. To create a prior distribution model of the points, the mean and the covariance of the points locations are extracted from the training data per feature point. Subsequently the model is referenced to an arbitrary origin, allowing it to "float" to any position in the image and it is rescaled to an arbitrary predefined size, which provides robustness against changes in scale. We fixed the origin to the center pixel of each eye in this work. Figure 2 shows a pseudo-image of the distribution model of the points.

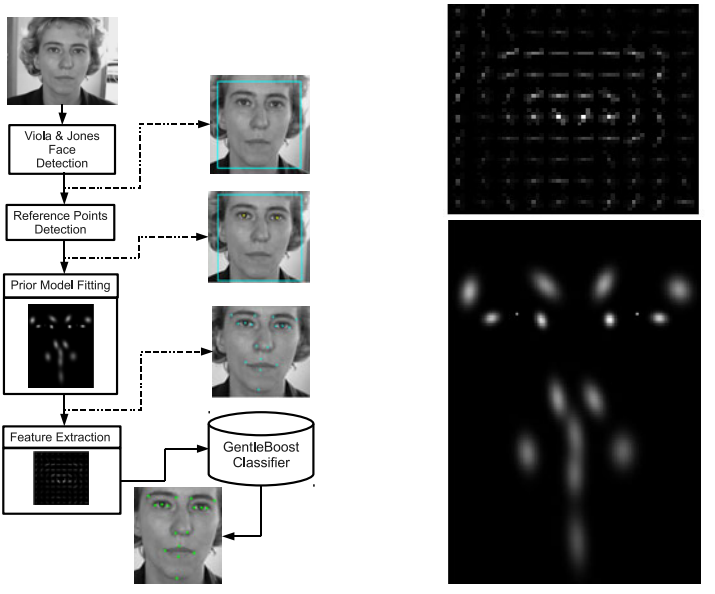
2.3 Detection Algorithm

The algorithm presented follows a standard classification framework. Once the training stage learns the descriptor for each feature point and the prior point distribution model is built, the detection stage can be performed. The algorithm starts with the bounding box returned by the Viola & Jones face detector. With this reference frame two regions are extracted from the upper half of this bounding box in order to obtain the centers of the eyes that constitute the reference points for the rest of the procedure. Thereafter the angle between the line joining the eyes and the horizontal is computed and the scale of the current face is determined[1]. With these parameters the prior point distribution model is adjusted to fit the scale of the query image providing the centers and cues for the size of the search regions. Finally using these locations cues, the points are searched within variable size regions. The search regions size is adjusted by the covariance parameters learned in the prior model. Figure 2 illustrates the major blocks and intermediate results of the method.

3 Experiments and Results

HOG Implementation. As mentioned in section 2.1 the HOG method uses 6 basic parameters recalling: number of orientation bins, range of orientations to

[1] The distance between the eyes once they have been rotated to a zero degree difference with the horizontal, is considered as the current scale.



(a) (b)

Fig. 2. Left: Outline of the method. **Right** Top: HOG descriptor output for the feature point of the pupil. The intensity of the lines is proportional to the strength of the gradient in that direction. Bottom: Prior Distribution Model. The clouds represent two standard deviations of the training data.

be considered, cell size, block size, overlap and normalization rule. In this study, we implemented the L_2-norm suggested in [3], that is:

$$v_n = \frac{v_i}{\sqrt{\sum ||v_i||^2 + \varepsilon^2}} \tag{1}$$

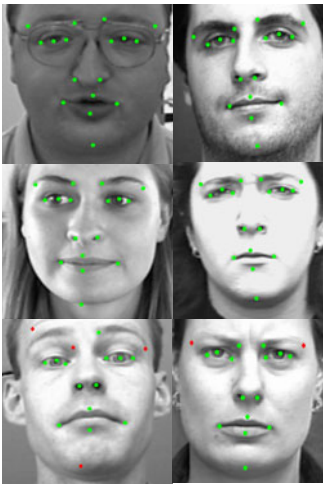
where v_i is the histogram of the ith cell and ε is a regularization parameter.

With respect to the range of orientations the current implementation uses an unsigned gradient, that is, the orientation bins are evenly spaced over 180 degrees. The number of bins to quantize the orientation histogram is 9, thus grouping angles in ranges of 20 degrees per bin. Each cell is 6×6 pixels and each block is 3×3 cells, with an overlapping factor of 75%. The final descriptor is built from an intensity patch of 30×30 pixels, by concatenating the block normalized cell histograms.

Datasets and Experimental Setup. The method proposed was evaluated with 570 images of the Cohn Kanade [6] and 350 images the BioID [5] databases. Given the rather small sample size of each dataset the error rate was estimated with a 3-Fold cross validation scheme. The results shown for the performance

Table 1. Performance results for the system on the Cohn Kanade and BioID data sets per Point. **Right.** Qualitative Results for both data sets. The first two rows present positive detections, the third detections outside of the threshold radius.

Point	BioID	Error	Cohn Kanade	Error
1	99.57% (0.84)	2.72%	100.00% (0.00)	1.84%
2	96.98% (2.53)	3.24%	99.82% (0.34)	2.08%
3	89.66% (6.76)	6.70%	98.07% (1.24)	3.12%
4	94.40% (9.29)	4.80%	98.77% (0.91)	3.08%
5	73.71% (9.29)	6.92%	83.68% (4.13)	5.92%
6	77.59% (3.38)	6.63%	88.60% (5.34)	4.77%
7	76.29% (5.91)	6.50%	92.63% (2.38)	4.31%
8	81.47% (14.36)	6.09%	88.07% (4.99)	5.11%
9	99.57% (0.84)	3.63%	99.65% (0.69)	3.28%
10	98.71% (0.84)	3.33%	100.00% (0.00)	2.55%
11	99.57% (0.84)	3.24%	99.82% (0.34)	2.91%
12	97.41% (0.00)	3.69%	100.00% (0.00)	2.80%
13	96.55% (0.00)	4.04%	99.82% (0.34)	2.66%
14	98.71% (0.84)	3.53%	100.00% (0.00)	2.57%
15	98.71% (2.53)	3.89%	99.12% (1.72)	2.88%
16	73.28% (13.52)	9.70%	93.16% (3.32)	4.63%
17	54.74% (2.53)	16.52%	68.95% (7.60)	8.26%



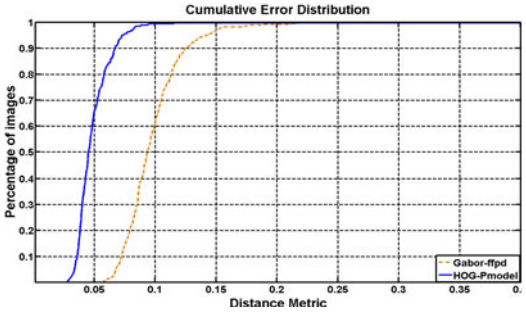


Fig. 3. Comparative of the Cumulative Error Distribution of the point to point error measured on the BioID Data set

are given with a confidence interval (shown in brackets in table 1) for a 95% confidence level. The classification accuracy for any point P_i given in table 1 was obtained from the normalized accumulated thresholded error:

$$Acc_i = \frac{\sum_{j=1}^{M} err_{ji} < T_h}{M} \tag{2}$$

where M is the number of images in the evaluation set and err_{ji} is the error defined as a function of the Inter-Ocular Distance –IOD (i.e. the magnitude of the vector joining the centre of the eyes) as:

$$err_i = \frac{||P_{oi} - \hat{P}_i||}{d_{IOD}} \tag{3}$$

where P_{oi} is the ground truth coordinate pair for interest point P_i, $\hat{P}_i$ is the detected coordinate pair for point P_i, $||\cdot||$ is the euclidean distance and d_{IOD} is the Inter-Ocular Distance. In this study the threshold T_h for the error assessment was set to 9%.

It can be seen in table 1 that the performance for all the points, but number 17, is rather high. In the case of the BioID dataset, the four points in the eyebrow area, exhibit a classification rate around 80% (see figure 1 to cross reference the point numbers). These points present a high variance in appearance and are highly sensitive to illumination changes. Moreover, the inter-observer agreement for the labeling of these points is low making them quite unstable for detection. In the case of point 16 in the BioID dataset, it suffers from one of the limitations of the method. Given the out-of-plane rotations of the face, specifically in yaw (nodding motion), the variation of the vertical face dimension is high, thus the fitting process of the Prior Model by a general scaling factor fails to properly allocate the lower points (lips and chin) correctly. In both data sets point 17 the chin, presents a rather low performance. This point has on top of the characteristics mentioned for the other 4 points, another that influences the instability of the point for the detection task. It is the high variability in appearance due to physical factors i.e., the sharpness of the jawline makes it a point rather dependent on the illumination conditions. We compared our method to the state of the art, represented by the implementation of [13] available from the authors web page. Figure 3 shows the cumulative error distribution of the mean error over all points. It can be seen that the method proposed has predictions with lower error: 20% of the images have an average error of less than 4% (2 pixels) and 90% of the images have an average error lower than 7% of d_{IOD}, as compared to the *Gabor-ffpd* that has no images with an error lower than 4% and 90% of the images lay between 5% and 13% of the d_{IOD}.

4 Discussion

We have presented a method for finding facial salient points in an input image of frontal faces, based on a prior Point Distribution Model and a robust local descriptor. The method exploits the local appearance information, the consistency of the global geometric information and its flexibility, for localizing the facial salient points. The use of this geometric Prior Distribution Model significantly improves the detection accuracy.

We evaluated the proposed method on two separate data sets, the Publicly available BioID and the Cohn-Kanade data base. The results show effectiveness in the accurate detection of facial salient points and robustness to illumination, scale and some degree of out-of-plane rotation. The limitation concerning the out-of-plane rotation, can be address by means of a more robust Prior Model that considers separately the variations suffered by the vertical and horizontal dimensions.

Our method shows better accuracy than current algorithms for facial point detection, with an average detection of 94.72% for all 17 points as compared to a 82.71% of the Gabor-ffpd, for the Cohn-Kanade dataset.

Acknowledgments. This work was partially supported by MEC grants TIN2009-14404-C02-01 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

1. Cohn, J.F., Sayette, M.A.: Spontaneous facial expression in a small group can be automatically measured: An initial demonstration. Behavior Research Methods (in press)
2. Cristinacce, D., Cootes, T.: Feature detection and tracking with constrained local models. In: Proceedings of British Machine Vision Conference, vol. 3, pp. 929–938 (2006)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886–893 (2005)
4. Hastie, J.F.T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. The Annals of Statistics 28(2), 337–374 (2000)
5. Jesorsky, O., Kirchberg, K., Frischholz, R.: Robust face detection using the hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 90–95. Springer, Heidelberg (2001)
6. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
7. Kozakaya, T., Shibata, T., Yuasa, M., Yamaguchi, O.: Facial feature localization using weighted vector concentration approach. Image and Vision Computing 28(5), 772–780 (2010)
8. Lepetit, V., Fua, P.: Keypoint Recognition Using Randomized Trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1465–1479 (2006)
9. Mayer, C., Wimmer, M., Radig, B.: Adjusted pixel features for robust facial component classification. Image and Vision Computing 28(5), 762–771 (2010)
10. Shinohara, Y., Otsuf, N.: Facial expression recognition using fisher weight maps. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 499–504 (2004)
11. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2729–2736 (2010)
12. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision 57(2), 137–154 (2002)
13. Vukadinovic, D., Pantic, M.: Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. In: IEEE International Conference on Systems, Man and Cybernetics (2005)

Rectifying Non-euclidean Similarity Data through Tangent Space Reprojection

Weiping Xu, Edwin R. Hancock, and Richard C. Wilson

Dept. of Computer Science,University of York, UK
{elizaxu,erh,wilson}@cs.york.ac.uk

Abstract. This paper concerns the analysis of shapes characterised in terms of dissimilarities rather than vectors of ordinal shape-attributes. Such characterisations are rarely metric, and as a result shape or pattern spaces can not be constructed via embeddings into a Euclidean space. The problem arises when the similarity matrix has negative eigenvalues. One way to characterise the departures from metricty is to use the relative mass of negative eigenvalues, or negative eigenfraction. In this paper, we commence by developing a new measure which gauges the extent to which individual data give rise to departures from metricity in a set of similarity data. This allows us to assess whether the non-Euclidean artifacts in a data-set can be attributed to individual objects or are distributed uniformly. Our second contribution is to develop a new means of rectifying non-Euclidean similarity data. To do this we represent the data using a graph on a curved manifold of constant curvature (i.e. hypersphere). Xu et. al. have shown how the rectification process can be effected by evolving the hyperspheres under the Ricci flow. However, this can have effect of violating the proximity constraints applying to the data. To overcome problem, here we show how to preserve the constraints using a tangent space representation that captures local structures. We demonstrate the utility of our method on the standard "chicken pieces" dataset.

Keywords: Dissimilarity, Embedding, Ricci flow, Spherical embedding, Tangent space.

1 Introduction

Geometric shape representation and recognition is an active area of research in computer vision and pattern recognition. Graph-based representations have found widespread use in shape analysis, for example, in the use of shock graphs to represent shape-skeletons [5]. When such a representation is adopted, measures such as graph-edit distance provide the natural way of capturing the similarity of different shapes. This provides a powerful and natural way of capturing the relationships between objects that are not characterised by ordinal measurements or feature vectors [6]. One way to construct a shape-space for such data is to represent the dissimilarity data using a weighted graph, and to embed the graph on a manifold. This produces a vectorial representation of the data by

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 379–386, 2011.

projecting dissimilarity data into a fixed-dimensional vector space. Examples of this approach include multidimensional scaling (MDS) and Isomap [10].

However, one of the problems with dissimilarity representations and their embeddings, is that the distance measures can not be used to construct a shape space if the underlying dissimilarity matrix contains negative eigenvalues. If this is the case the shapes can not be embedded into a real-valued Euclidean space [3], and must instead be embedded into a complex valued or Krein space.

In order to analyse non-Euclidean dissimilarity data using traditional geometric machine learning or pattern recognition techniques, we must first attempt to rectify the data so as to minimize the non-Euclidean artifacts. Before the analysis of such data is attempted, it is advisable to assess the degree and extent to which non-Euclidean artefacts affect the data-set. One measure that has proved useful in this respect is the negative eigenfraction [1] which is the total mass of negative eigenvalues as a fraction of the total mass of unsigned eigenvalues. However, in this this paper, we introduce a finer measure that assesses the contribution of each object to the mass of negative eigenvalues. In this way it is possible the determine whether the non-Euclidean artefacts are attributable to the outlying dissimilarities of a few objects or are uniformly distributed throughout the dataset.

However, our main contribution in the paper is to consider how to rectify the data to minimise the effects of non-Euclidean artefacts. Xu et. al. [8] have explored the idea of embedding the disimilarity data on a locally hyperspherical surface of constant curvature. They then flatten the manifold composed of local hyperspherical patches by reducing the curvatue according to a Ricci flow. As a result both local and global distances are modified by the flattening. One of the problems they have encountered in applying Ricci flow to a constant curvature Riemannian manifold to evolve the distance measures is that due to the piecewise nature of the manifold, the structure of the data is distorted. Moreover, they also encounter instabilities due to local fluctuations in edge curvature. Although this latter problem can to some extent be remedied by regularizing the Gaussian curvature [8], the problem of preserving structure persists.

To overcome this problem, we aim to reduce the reliance on the piecewise embedding and its effect on individual edges. We turn to the tangent space representation of data using the exponential and log maps [2], which provide a means of preserving the distance between the points on the manifold and the origin of the map. This allows us to to flatten the manifold while preserving the global structure of the data.

2 Characterising Non-euclidean Data

In this paper we are concerned with embedding data represented in terms of pairwise dissimilarities or distances, and in particular the case where the data is non-Euclidean. Our overall aim is to rectify a given set of non-Euclidean dissimilarity data so as to make them more Euclidean. One way to gauge the degree to which a pairwise distance matrix exhibits non-Euclidean artefacts is to analyse

the properties of its centralised Gram matrix. For an $N \times N$ symmetric pairwise dissimilarity matrix D with the pairwise distance as elements, the centralized Gram matrix $G = -\frac{1}{2}JD^2J$,where $J = I - \frac{1}{N}11^T$ is the centering matrix and 1 is the all-ones vector of length N. The degree to which the distance matrix departs from being Euclidean can be measured by using the relative mass of negative eigenvalues or "negative eigenfraction " $F_{eigS} = \sum_{\lambda_i<0} |\lambda_i| / \sum_{i=1}^{N} |\lambda_i|$ [1]. This measure is zero when the distances are Euclidean and increases as the distance becomes increasingly non-Euclidean.

If the non-Euclidean artefacts are contributed solely by the set of distances to a few "outlier" objects, it is possible to restore the data to a Euclidean state by editing (i.e. removing) these objects from the dataset. Based on this idea we introduce the notion of measuring the contribution of each object to the negative eigenfraction of a dissimilarity matrix. That is, the fraction given by the sum of the negative distances originating from an individual object to all the remaining objects, divided by the total.

The matrix of kernel embedding co-ordinates is given by $Y = \sqrt{\Lambda}\Phi^T = (y_1, ..., y_N)$, where $\Lambda = diag(\lambda_1, ..., \lambda_N)$ is the diagonal matrix with the ordered eigenvalues of centered Gram matrix as elements and $\Phi = (\phi_1|...|\phi_N)$ is the eigenvector matrix with the ordered eigenvectors ϕ_1, ..., ϕ_N as columns. When the centered Gram matrix has negative eigenvalues then those dimensions of the embedding associated with negative eigenvalues are represented by imaginary numbers, and those associated with positive eigenvalues by real numbers. In other words, the data are embedded into a pseudo Euclidean or Krein space [3].

Under the embedding, the coordinate vector of point j is $y_j = (\sqrt{\lambda_1}\Phi_{1j}, ..., \sqrt{\lambda_i}\Phi_{ij}, \sqrt{\lambda_N}\Phi_{Nj})^T$. The contribution to the negative squared distance between two points k and e is $d^2_{ke} = \sum_i \big(y_k(i) - y_e(i)\big)^2 = \sum_i \lambda_i (\phi_{ik} - \phi_{ie})^2$. The sum of negative squared distances from point k to all the remaining points is $d^2_{k-} = \sum_{\lambda_i<0} \lambda_i \sum_{e \neq k} (\phi_{ik} - \phi_{ie})^2$. On the other hand, the sum of positive distances from point k to the remaining points is $d^2_{k+} = \sum_{\lambda_i>0} \lambda_i \sum_{e \neq k} (\phi_{ik} - \phi_{ie})^2$. Thus the fraction of negative squared distances from point k is $C_{neig} = \frac{|d^2_{k-}|}{|d^2_{k-}|+|d^2_{k+}|}$.

3 Spherical Embedding

Spherical embedding [9] provides a means by which to embed objects represented in terms of dissimilarity data onto a hypersphere. The optimal radius of the hypersphere minimises the distortion of the geodesic distances between objects. It is desirable that the degree of the nodes of the embedded graph and the ranking of distances (dissimilarities) are preserved under the embedding. Given the hypershere of optimal raidus, the embedding coordinates are obtained through the eigendecomposition of the inner product matrix[9], and for the node indexed p the vector of embedding co-ordinates is $y_p = \sqrt{\Lambda}\Phi^T$, where Λ is the diagonal matrix with the ordered eigenvalues of the inner product matrix $Z = cos(\frac{D}{r})$

as elements and Φ is the matrix with the ordered eigenvectors of Z as columns, D is the distance matrix and $r \in R^+$ is the optimal radius of the embedding hypersphere.

4 The Exponential and Log Map

The exponential map $\mathrm{Exp_p}[.]$ is a mapping from points on the manifold to points in the tangent space at a reference point on a manifold. The log map $Log_p[.]$ the inverse mapping from points in the tangent space at the reference point to points on the manifold. The exponential map [2] preserves the distance between the points on the manifold and the reference point or origin of the tangent space to the manifold. Our aim is to flatten the global manifold by gradually

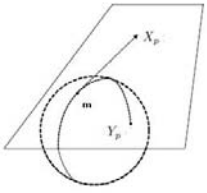


Fig. 1. The exponential map and log map

smoothing out the local patches. This is achieved by representing sub-graphs of objects on the local hyperspheres, mapping the points to the tangent space through the log-map function, reducing the curvatures (i.e.increasing the radii) of the individual hyperspherical patches, and then mapping the data back onto the inflated hyperspheres through the exponential-map function. The increase in radius of the hypersheres is determined by the Ricci flow [7] and satisfies the equation

$$\frac{dg_{ij}}{dt} = -2R_{ij}. \tag{1}$$

where g_{ij} is the metric tensor of the manifold and R_{ij} is the Ricci curvature. We model the embedding manifold as consisting of a set of local patches with individual constant Ricci curvatures. The solution of the differential equation is straightforward. Commencing with the initial conditions curvature $K = K_0$ at time $t = 0$, then at time t we have $K_t = \frac{K_0}{1 \pm 2K_0 t}$ with the positive sign for the elliptic space (hypersphere).

On the spherical manifold, the log and exp maps give the following co-ordinate trasformations [9]:

$$x_p = \frac{\theta}{\sin\theta}(y_p - y_m \cos\theta); y_p = y_m \cos\theta + \frac{\sin\theta}{\theta} x_p. \tag{2}$$

where y_p is the coordinate vector for point p on the manifold, x_p is the coordinate vector for point p in the tangent space, m is the reference point or origin of the

map with the length equals to the corresponding radius, and θ is the angle between radius vectors to the points m and p on the hypersphere. This set of transformations is illustrated in Figure 1.

Once, the inflation and reprojection onto the hypersphere are complete, we compute the new coordinate vector of a point on the inflated hypersphere based on the old coordinate vector on the original hypersphere by using Equation 2:

$$y_{p_{n+1}} = (1 + 2K_n t)^{\frac{1}{2}}(y_{m_n} \cos\theta_{n+1} + \frac{\sin\theta_{n+1}}{\sin\theta_n}(y_{p_n} - y_{m_n} \cos\theta_n)). \quad (3)$$

where the angle on the original sphere is $\theta_n = K_n acos < y_m, y_p >$.

As the geodesic distances to the origin are preserved, we can compute the angles on the inflated sphere $\theta_{n+1} = \frac{K_{n+1}^{\frac{1}{2}}}{K_n^{\frac{1}{2}}}\theta_n$, given the curvatures of the original and the inflated spheres, and updated radial angles on the original sphere.

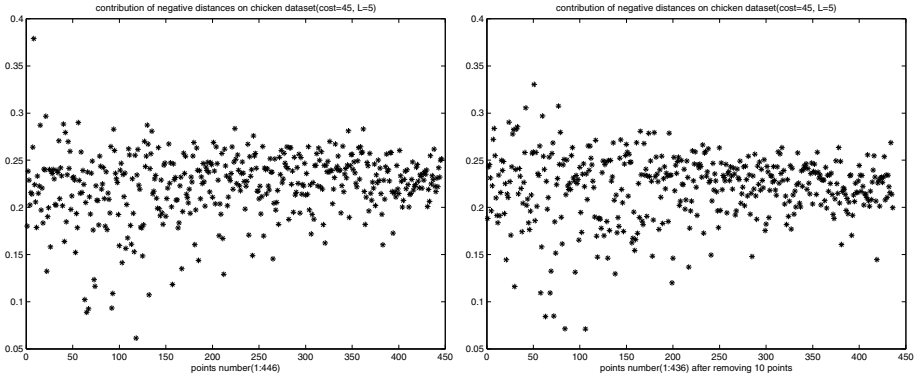
Then we compute new geodesic distances for the points on the inflated hypersphere. Reprojection under the log map preserves the geodesic distances to the origin of the tangent space. However, the geodesic distances between points are modified by the inflation and reprojection. The updated geodesic distances on the inflated hypersphere can be computed using the new co-ordinates on the inflated hypersphere. The update equation for the geodesic distance between point m and p on the inflated hypersphere is

$$d_{Gmp} = r_{n+1}\theta_{n+1} = \frac{acos(< y_{m_{n+1}}, y_{p_{n+1}} > K_{n+1})}{K_{n+1}^{\frac{1}{2}}}. \quad (4)$$

5 The Algorithm

Given a set $Y = \{y_1, \cdots, y_N\}$ of N objects and a dissimilarity measure d, a dissimilarity representation is an $N \times N$ matrix D_G with the elements $d_G(u, v)$ representing the pairwise geodesic distance between objects y_u and y_v. The following algorithmic steps can be used to perform Euclidean rectification of the distance matrix, and suppress its non-Euclidean artefacts:

1. Construct a local patch for a second order k-NN graph, consisting of the first and second neighbors of the reference object.
2. Perform hyperspherical embedding to obtain the initial curvature and the co-ordinates of the objects in the local patch.
3. Update the hyperspherical radius (i.e. curvature) with a small time step derived from Equation 1.
4. Obtain the new coordinates on the inflated hypersphere using Equation 3.
5. Obtain the new geodesic distance matrix $d_{G_{n+1}}$ for the local patch using Equation 4. The distances between pairs of objects external to the patch are approximated using the old dissimilarity matrix. The geodesic distance between pairs of objects external to the patch and objects internal to the patch are approximated by adding the geodesic distances over the of edge-connected path between the objects.



(a) The original distribution

(b) The distribution after removing 10 highly-contributed points

Fig. 2. The distribution of negative contribution of each point before and after removing 10 most highly-contributing objects

6. Obtain the updated global distance matrix $D_G^{(1)}$ containing rectified geodesic distances between objects, and repeat from step 1 until D_G stabilises, i.e. there are no further decreases in the negative eigenfraction. Ideally, the centralized Gram matrix should have no negative eigenvalues.

6 Experiments

We use the well known "Chicken pieces" shape dataset [4] for experimentation. The data-set poses the problem of classifying binary images of different types of chicken joint into shape-classes. It contains 446 binary images falling into five shape classes, namely a) breast (96 examples), b) back (76 examples), c) thigh and back (61 examples), d) wing (117 examples) and e) drumstick (96 examples). The data exists in the form of a set of non-Euclidean shape dissimilarity matrices, generated using different parameter settings. The parameters are the length of straight line segments of the chicken contours L and the insertion and deletion costs for computing edit distances between boundary segments C. Our experimental results are for the dissimilarity data with parameters $C = 45$ and $L = 5$, 10, 15, 20, 25 and 30. The originally asymmetric dissimilarities are made symmetric by averaging.

We commence by showing the distribution of the individual object contributions to the negative eigenfraction for the chickenpieces data. In Figure 2 we show the distribution for the data with L =5.0; C=45, both before and after removing the 10 most strongly ontributing points. Figure 3 shows the negative eigenfraction of the chickenpieces data with $C = 45$ and $L = 5$, 10, 15, 20, 25 and 30, again both before and after removing the 10 most strongly contributing objects. Removing the most strongly contributing objectss has little effect

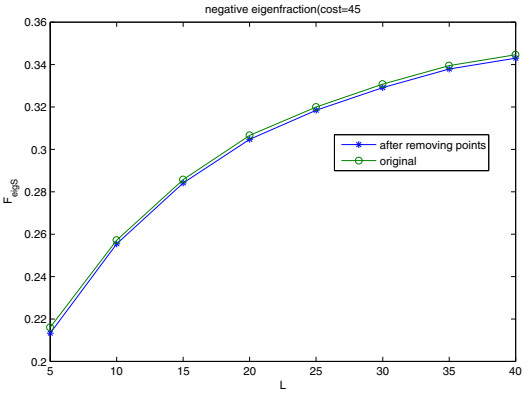
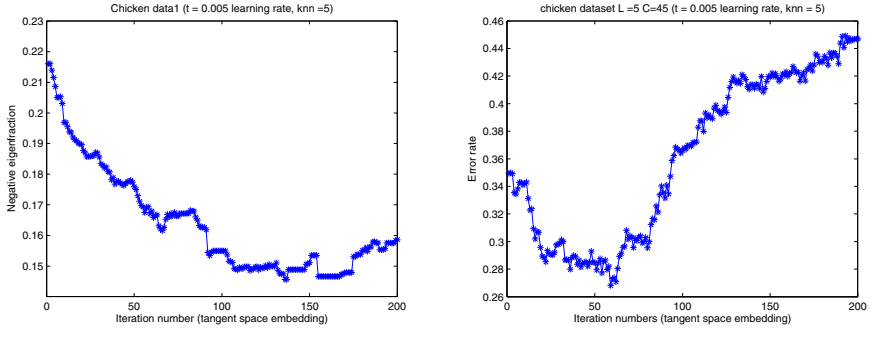


Fig. 3. The mass contribution of negative eigenvalues before and after removing the 10 most highly-contributing objects



(a) The negative eigenfraction (b) The 1NN error rate

Fig. 4. The negative eigenfraction and 1NN error rate as a function of iteration number

on he distribution, and this indicates that the non-Eucludean artefacts can not be attributed to outliers. Next, we explore the effectiveness of the rectification process. Figure 4 shows the negative eigenfraction and 1NN error rate for the shape-classes as the distance matrix is evolved. The negative eigenfraction drops from 22% to 15% and then increases again, indicating that the evolution has succeeded in flattening the manifold, but then deteriorates. This demonstrates that the distance measures can be corrected or flattened, but it is essential to have a halting criterion. The deterionation is caused by the path-based approximation of geodesic distances between internal and external objects on the local patches. These distances have been inflated more rapidly than the local distances on the hyperspherical surface. This exaggerates the overall curvature.

7 Conclusion

In this paper, we have made two contributions. First, we have presented a method to gauge the distribution of non-Euclidean artefacts in dataset Second, we have shown how to evolve a patchwise hyperspherical manifold so as to rectify such artefacts in a dataset. The method uses a tangent-space reprojection method to inflate the local hyperspherical patches, while maintaining the consistency of the pattern of geodesic distances. Applying our method to the chicken pieces shape dataset, we demonstrate that the method can be used to improve the classification of shape-data.

Acknowledgments. Work was supported by the project SIMBAD (213250). Edwin Hancock was supported by a Royal Society Wolfson Research Merit Award.

References

1. Pękalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-euclidean or non-metric measures can be informative. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR 2006 and SPR 2006. LNCS, vol. 4109, pp. 871–880. Springer, Heidelberg (2006)
2. Fletcher, P.T., Lu, C., Pizer, S.M., Joshi, S.: Barbara and Bunke, Horst: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Transactions on Medical Imaging, 995–1005 (2004)
3. Goldfarb, L.: A new approach to pattern recognition. Progress in Pattern Recognition, 241–402 (1985)
4. Andreu, G., Crespo, A., Valiente, J.M.: Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recognition. In: ICNN, pp. 1341–1346 (1997)
5. Torsello, A., Hancock, E.R.: Computing approximate tree edit distance using relaxation labeling. Structural, Pattern Recognition Letters, 1089–1097 (2003)
6. Sanfeliu, A., Fu, K.-S.: A Distance measure between attributed relational graphs for pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics, 353–362 (1983)
7. Chow, B., Luo, F.: Combinatorial Ricci flows on surfaces. J. Differential Geom., 97–129 (2003)
8. Xu, W., Hancock, E.R., Wilson, R.C.: Regularising the ricci flow embedding. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR&SPR 2010. LNCS, vol. 6218, pp. 579–588. Springer, Heidelberg (2010)
9. Wilson, R.C., Hancock, E.R.: Spherical embedding and classification. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR&SPR 2010. LNCS, vol. 6218, pp. 589–599. Springer, Heidelberg (2010)
10. Tenenbaum, J.B., Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science (2000)

Gait Identification Using a Novel Gait Representation: Radon Transform of Mean Gait Energy Image

Farhad Bagher Oskuie, Karim Faez,
Ali Cheraghian, and Hamidreza Dastmalchi

Dept. Electrical Engineering
Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
F.bageroskuee@aut.ac.ir

Abstract. Gait is one of the most practical biometric techniques which present the capability to recognize individuals from distance. In this study, we propose a novel gait template based on Radon Transform of Mean Gait Energy Image, as RTMGEI. Robustness against image noises and reducing data dimensionality can be achieved by using Radon Transform, as well as capturing variations of Mean Gait Energy Images (MGEIs) over their centers. Feature extraction is done by applying the Zernike moments to RTMGEIs. Orthogonal property of Zernike moment basis functions guarantees the statistically independence of coefficients in extracted feature vectors. The Euclidean minimum distance is used as the classifier. The our proposed method is evaluated on the CASIA database. Results show that our method outperforms recently presented works due to its high performance.

Keywords: Gait Recognition, Radon Transform, Zernike Moment.

1 Introduction

Gait means the manner of walking. Studies at [1] showed that people can recognize each other from their gaits. Therefore gait as a biometric can be used for recognizing people. Recognition based on gait has a unique property which requires no contact such as automatic 2D or 3D faces, fingerprints or iris recognition systems. All this methods need individual cooperation near to the systems. However, some variations such as clothing, types of shoes, the environment of walking and age affect the gait.

Various techniques and algorithms have been developed for human gait recognition recently. These techniques are generally categorized into two main groups: model-based and model-free approaches. Model-based method tends to explicitly model human body or motion, and they usually implement model matching in each frame of a gait sequence to measure the parameters such as trajectories according to the matched model [2, 3]. In this paper, we focus on the model-free algorithms.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 387–394, 2011.

In [4] a new spatio-temporal gait representation called Gait Energy Image (GEI) is proposed to address human walking properties for individual recognition. Also, a novel approaches for gait recognition is proposed by combining statistical gait features from real and synthetic templates. Similarly, other temporal templates called Gait History Image (GHI) is proposed in [5] which models gait static and dynamic characteristics more comprehensively. GHIs are used to learn discriminating features with statistical approaches. In another work, two types of gait features, GEI and Motion Silhouette Image (MSI) are extracted to identifying individuals by using the fused output of nearest neighbor classifiers [6]. In [7] a novel algorithm based on Fuzzy Principal Component Analysis (FPCA) is proposed to extract the eigenvectors from the GEIs. The eigenvectors are then projected to the subspace with lower dimensionality and the Nearest Neighbor (NN) classifier is used.

Radon Transform is used in [8], as new feature extractor, directly on binary silhouettes of each gait sequence. Transformed silhouettes are used for computation of a template which subsequently is subjected to LDA and subspace projection. Test feature vector is compared with feature vectors in the gallery to recognition and verification.

The structure of the paper is as follows. In section 2, we present our proposed gait recognition system which includes overview, Radon Transform and Zernike Moments sub-sections. Section 3, describes our experimental results, and, finally, conclusions are drawn in section 4.

2 Gait Identification System

2.1 Overview

Our proposed system is shown in Fig. 1. As many other researches, the silhouettes have been extracted from the original gait video sequences and the pre-processing procedure [9] is applied on them. The denoising process is needed before applying the gait recognition algorithm. First, the centers of silhouettes in the gait sequence have been calculated and then each silhouette is segmented into predetermined size over its center. Then the segmented silhouettes have been aligned using their calculated centers.

One of the recently developed spatio-temporal gait templates is the Gait Energy Image (GEI) which was first represented by Ju Han at [4]. GEI has no sensitivity to incidental silhouette errors in individual frames. As expected, GEI represents fundamental shapes of silhouettes as well as their changes over the gait cycle [4]. But in gait cycles which have incomplete silhouettes or occlusion with objects, the recognition based on the GEI leads to incorrect results. In order to avoid the mentioned problems, we prefer to use the Mean Gait Energy Image (MGEI) as a base representation for extracting features [10]. Definition for calculating MGEI of i^{th} sequence is as following:

$$MGEI_i(x, y) = \frac{1}{M_i} \sum_{j=1}^{M_i} GEI_i(x, y), \tag{1}$$

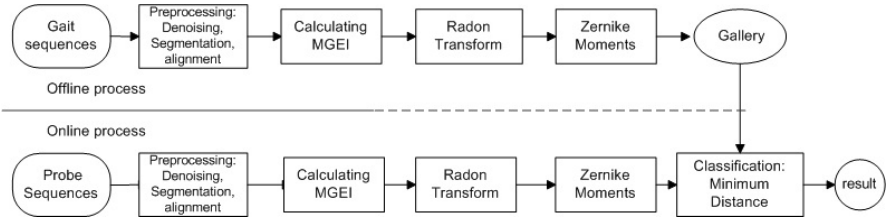


Fig. 1. Block diagram of the proposed system for gait recognition

where, M_i is the number of different gait cycles existing in the i^{th} sequence, x and y are the values of two dimensional image coordinates. $GEI_{i,j}$ is the Gate Energy Image for j^{th} cycle of i^{th} sequence and is calculated by the following equation:

$$GEI_i(x,y) = \frac{1}{N_j}\sum_{t=1}^{N_i} I_{i,j,t}(x,y). \tag{2}$$

According to the Fig. 1, after calculating MGEI, the Radon Transform is applied and the novel spatio-temporal template is produced. We call this template RTMGEI. RTMGEI is defined as the following:

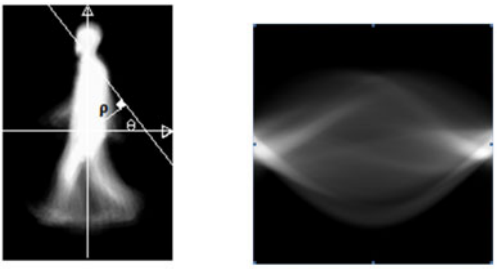
$$RTMGEI_i = RT_{MGEI_i}(\rho,\theta). \tag{3}$$

where, ρ is the radial distance from the center of image or data to be transformed, and θ is the angle in the polar coordination system. Definition of Radon Transform and its arguments are described in detail in the next sub-section.

Fig. 2 illustrates one calculated MGEI and its corresponding RTMGEI. Since MGEI and RTMGEI are normalized and transformed from the frames of the sequences, we can display them as a viewable image. As it is evident in the figure, the RTMGEI are smoother and has less noise than its corresponding MGEI. This is because of summation property of Radon Transforms which reduces the noise of MGEIs and yields a better temporal template in presence of noise. Actually the first level of denoising is done by summing single frames to construct MGEIs and the second level of denoising is done taking Radon Transform of MGEIs. Also using RTMGEIs will result in considerable reduction of data dimensions and will increase the separability of data in classification sub-space.

After calculating RTMGEIs, Zernike moments are employed for size-reduction and feature extraction on each sequence.The Zernike moments are described in detail at the sub-section (2.3). We use the Zernike moments of up to the 15^{th} order which result in feature vector with 136 coefficients.

In the offline process, as shown in Fig. 1, the features vectors of sequences, which are produced as mentioned, are saved in the Gallery set. In the online process, the feature vector of probe sequence is produced and compared with Gallery set. The Euclidean Minimum distance is used as the classifier.



(a) MGEI (b) RTMGEI

Fig. 2. (a) Calculated MGEI for a sequence, (b) related RTMGEI

2.2 Radon Transform

Hough and especially Radon Transforms have found various applications within the computer vision, image processing, pattern recognition and seismic. Mapping the two-dimensional images with lines into a sub-space is one of best abilities of these two transforms; where each line in the image will give a peak, positioned at the corresponding line parameters.

One can find several definitions of the Radon transform in mathematics, but the very popular form is as the following [8]:

$$RT_f(\rho,\theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\delta(\rho - x\cos\theta - y\sin\theta), \tag{4}$$

which, expresses the lines in the form of $\rho = x\cos\theta - y\sin\theta$, where θ is the angle and ρ is the smallest distance to the origin of the coordinates system. The Radon transform for a set of parameters (ρ,θ) is the line integral through the image $f(x,y)$, where the line is positioned corresponding to the value of (ρ,θ). The Dirac delta function $\delta(.)$ is defined as ∞ for argument zero and 0 for all other arguments (it integrates to one). In the digital domain, the Kronecker delta will be used instead, which is defined as 1 for argument zero and as 0 for other all others. Thus the Radon Transform will be simplified to summation of pixel intensities along the discrete lines (Fig. 2.(a)).

2.3 Zernike Moments

Zernike moment is some kind of orthogonal complex moments in which its interesting properties such as rotation invariance, translation invariance and scale invariance have been improved [9, 10, 11]. Zernike moments kernels consist of Zernike complete orthogonal polynomials. These polynomials are defined over the interior region of the unit disc in the polar coordinates space. Let $f(r,\theta)$ be the image intensity function, and the two-dimensional Zernike moments of order m with repetition n are defined as:

$$Z_{mn} = \frac{m+1}{\pi}\int_0^{2\pi}\int_0^1 f(r,\theta)V_{mn}^*(r,\theta)rdrd\theta, \mathrm{r} \le 1, \tag{5}$$

where $V^*_{mn}(r,\theta)$ is the complex conjugate of Zernike polynomial $V_{mn}(r,\theta)$; and m and n both are integer and the relation between m and n can be described as:

$$(\mathrm{m} - |\mathrm{n}|) \text{ is even and } |\mathrm{n}| \leq \mathrm{m}. \tag{6}$$

The Zernike polynomial $V_{mn}(r,\theta)$ is defined as:

$$V_{mn}(r,\theta) = R_{mn} e^{jn\theta}, \tag{7}$$

where $j = \sqrt{-1}$; and the orthogonal radial polynomial $R_{mn}(r)$ is given by:

$$R_{mn}(r) = \sum_{s=0}^{\frac{m-|n|}{2}} (-1)^s \frac{(m-s)!}{s!(\frac{m+|n|}{2}-s)!(\frac{m-|n|}{2}-s)!} r^{m-2s}. \tag{8}$$

For the discrete image, let $P(r,\theta)$ to be the intensity of the image pixels, and (5) can be represented as:

$$Z_{mn} = \frac{m+1}{\pi} \sum_r \sum_\theta P(r,\theta) V^*_{mn}(r,\theta). \tag{9}$$

The orthogonal radial polynomials result in Zernike moments which have less redundancy [12]. Structural and static information of individuals in related RMGEI can be represented by the low-order Zernike moments, and the dynamic information of RMGEI can be represented by high-order Zernike moments.

3 Experimental Results

Our Method is carried out on the CASIA database. The CASIA includes three sub-databases named DatasetA, DatasetB and DatasetC. We use the gait sequences of DatasetB which are captured from 124 subjects. Each subject includes 10 different gait sequences from 6 different camera views in the same scene. The original Image size of the database is 320x240 pixels. For each person there is: 6 normal gait sequences (set A), two bag carrying sequences (set B) and two coat wearing sequences (set C). For each set, we take first sequences of 90° view as the training subset and named them as set A1, set B1 and set C1 respectively. The rest sequences of 90° view are taken as the test subsets and named them as set A2, set B2 and set C2 respectively.

First, each frame of the sequence is segmented, aligned and resized to an image with 201x181 pixels. Then the RTMGEI's of each sequence is calculated as described in section (2). The RTMGEIs can be illustrated as a 180x180 pixel image. Final step is calculating Zernike moments of up to the order of 15^{15} order which result in the feature vector of length 136 coefficients. In the Zernike moments computing step, the pixel coordinates are transformed into the range of the unit circle, i.e. $x^2 + y^2 \leq 1$ and the center of the image is considered as the origin. Our method is carried out on a PC with 2.2GHz Core2 Duo CPU and 2.00GByte DDR2 RAM. Algorithm is implemented in MATLAB software. For

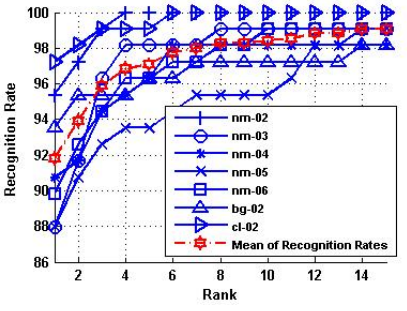


Fig. 3. Recognition rates for different probe sets

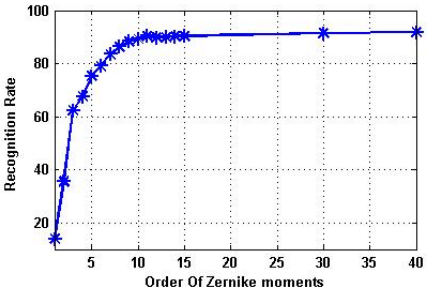


Fig. 4. Recognition rates for different order of Zernike moments

each person, Zernike moments extraction from RTMGEI by processing all steps, on the mentioned PC, takes 844.4 *ms* to complete in average. Also, minimum distance classifier takes 1 ms to complete. Therefore, it will take 845.4 ms to recognize a person. Thus, the proposed algorithm is capable of implementing on real-time gait recognition systems.

Fig. 3 demonstrates the recognition rates in different ranks. Each normal sequence existing in set A2 is indicated by 'nm'suffix. 'bg-02' indicates the test set B2 and 'cl-02' indicates the test set C2. As it is shown, the mean of recognition rate in Rank 1 and Rank 5 is 91.79% and 97.08% respectively.

For determining the optimum order of the Zernike moments used in our algorithm, we modify our algorithm to prepare a test using different order of Zernike moments. Thus, we applied the set A2 to our proposed system with predetermined values for Zernike moment's order ranging from 1 to 40. The mean recognition rates for various the Zernike moments are showed in Fig. 4. Therefore increasing the order of Zernike moments from the value of 10 has slight effect on recognition rate. But for marginal consideration, the final order of the Zernike moments is supposed to be 15 in all other experiments.

The mean of recognition rates on the normal, carrying bag and wearing coat for different works are shown in the table I. We compare our proposed algorithm with the baseline algorithm [13]. Also we compare our work with MSCT&SST algorithm [14] which is the human gait recognition using the fusion of motion and static spatio-temporal templates. GEI [4] and GEnI [15] are the state-of-the-art methods in recent years. KPCA algorithm [16] is based on the mean gait energy image (MGEI) which utilizes kernel principle component analysis (KPCA) for capturing high-order statistics which are particularly important for MGEI structure. Also in recent work, PCA algorithm was carried out to comparison between KPCA and single PCA. Another PCA based method which is called Fuzzy PCA (FPCA), as described in section 1, is compared with our proposed method. In the table, Rank 1 is the correctness of subjects which is placed at top in the rank list and, in the similar manner, Rank 5 indicates

Table 1. Recognition rates of our proposed algorithm and some of the other related works

Algorithms	Performance	
	Rank1	Rank5
Baseline	73%	88%
GEnI [15]	70.6%	-
TM	60.7%	-
GEI [4]	90%	94%
PCA [16]	80.6%	-
KPCA [16]	87%	-
MSCT&SST [14]	80%	92%
FPCA [7]	89.7%	-
RTMGEI-Zer (Our Method)	**91.79%**	**97.08%**

the percentage of the correct subjects appearing in any of the first five places of the output rank list. Results show that our method outperforms the other algorithms.

4 Conclusions

In this paper, we proposed a new gait representation called RTMGEI, which extract the dynamic and static characteristics of gait sequences. RTMGEI can be extracted from incomplete sequences and has better noise characteristics over the other gait representations. It is because of summation property of Radon Transform. Also we use Zernike moments to extract feature vectors. Due to orthogonal properties of Zernike basis functions, individual coefficients in feature vector have minimum redundancy. Finally, Euclidean Minimum distance is used to compare the probe feature vector with the stored feature vectors in the Gallery. Measuring the time needed to accomplish the recognition process, determines our proposed algorithm has the capability of implementing in the real-time identification systems.

The algorithm is evaluated on the CASIA gait database. Our results show significantly better performance compared to the other mentioned methods and our algorithm outperforms recent works.

References

1. Sarah, V.S., Mark, S.N., Kate, V.: Visual analysis of gait as a cue to identity. Applied Cognitive Psychology 13(6), 513–526 (1999)
2. Johnson, A.Y., Bobick, A.F.: A multi-view method for gait recognition using static body parameters. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 301–311. Springer, Heidelberg (2001)
3. Lee, L., Grimson, W.E.L.: Gait analysis for recognition and classification. In: Proce. IEEE Int. Confr. Automatic Face and Gesture Recognition, Washington DC, pp. 155–162 (2002)

4. Ju, H., Bir, B.: Individual Recognition Using Gait Energy Image. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(2) (February 2006)
5. Jianyi, L., Nanning Z.: Gait History Image:A Novel Temporal Template For Gait Recognition. In: Mulimedia and Expo, pp. 663–666 (July 2007)
6. Hong, S., Lee, H., Kim, E.: Fusion of Multiple Gait Cycles for Human Identification. In: IEEE Int. Joint. Confr., Japan, pp. 3171–3175 (2009)
7. Su-li, X., Qin-jin, Z.H.: Gait Recognition using Fuzzy Principal Component Analysis. In: e-Business and Information System Security, Wuhan, pp. 1–4 (2010)
8. Nikolas, V.B., Zheiwei, X.C.: Gait Recognition Using Radon Transform and Linear Discriminant Analysis. IEEE Trans. On Image Processing 16(3), 731–740 (2007)
9. Ye, B., Peng, J.: Invariance analysis of improved Zernike moments. Journal of Optics A: Pure and Applied Optics 4(6), 606–614 (2002)
10. Ye, B., Peng, J.: Improvement and invariance analysis of Zernike moments using as a region-based shape descriptor. Journal of Pattern Recognition and Image Analysis 12(4), 419–428 (2002)
11. Chong, C.W., Raveendran, P., Mukundan, R.: Translation invariants of Zernike moments. Pattern Recognition 36(8), 765–773 (2003)
12. Maofu, L., Yanxiang, H., Bin, Y.: Image Zernike Moments Shape Feature Evaluation Based on Image Reconstruction. Geo-spatial Information Science 10(3), 191–195 (2007)
13. Sarker, S., Jonathon Phillips, P., Liu, Z., Vega, I.R., Grother, P., Bouyer, K.W.: The Human ID Gait Challenge problem: data sets, performance and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(2) (February 2005)
14. Lam, T.W., Lee, R.S.T., Zhang, D.: Human gait recognition by the fusion of motion and static spatio-temporal templates. IEEE. Jorn. Pattern Recognition 40, 2563–2573 (2007)
15. Bashir, K., Xiang, T., Gong, S.: Gait Recognition Using Gait Entropy Image. In: IEEE Int. Confr. Crime Detection and Prevention, London, pp. 1–6 (2009)
16. Xiang-tao, C., Zhi-hui, F., Hui, W., Zhe-qing, L.: Automatic Gait Recognition Using Kernel Principal component Analysis. In: IEEE Int. Confr. Biomedical Engineering and Computer Science, Wuhan, pp. 1–4 (2010)

New Algorithm for Segmentation of Images Represented as Hypergraph Hexagonal-Grid

Dumitru Burdescu, Marius Brezovan, Eugen Ganea, and Liana Stanescu

University of Craiova, Craiova,
Bd. Decebal 107, Romania
{burdescu_dumitru,brezovan_marius,ganea_eugen,
stanescu_liana}@software.ucv.ro

Abstract. This paper presents a new method for segmentation of images into regions and for boundary extraction that reflect objects present in the image scene. The unified framework for image processing uses a grid structure defined on the set of pixels from an image. We propose a segmentation algorithm based on hypergraph structure which produces a maximum spanning tree of a visual hypergraph constructed on the grid structure, and we consider the HCL (Hue-Chroma-Luminance) color space representation. Our technique has a time complexity lower than the methods from the specialized literature, and the experimental results on the $Berkeley$ color image database show that the performance of the method is robust.

1 Introduction

The problem of image segmentation remains a great challenge for computer vision. Image segmentation techniques can be distinguished into two groups: region-based, and contour-based approaches. These two approaches need not to be different one from other, because boundary of regions can be defined to be contours. If one enforces closure in a contour-based framework [1], then can obtain regions having as boundary the detected closed contours. Conversely if one can obtain regions from an image, then the closed contours of the extracted regions can be determined. The method proposed in the article uses new techniques that leads to simpler and more efficient segmentation algorithms; we propose a low-level method for color image segmentation. The segmentation task implements a region based segmentation method that captures both certain perceptually important local and non-local image features. The detected visual objects are then analyzed by the boundary extraction task, which implements a faster algorithm for object boundaries detection. The proposed feature-based segmentation method uses a hypergraph constructed on a hexagonal structure containing half of the image pixels in order to determine the maximum spanning tree for each connected component representing a visual object. In [2] was presented an overview of a hypergraph-based image representation that considered Image Adaptive Neighborhood Hypergraph ($IANH$) model. The proposed segmentation method is original and uses a virtual graph structure constructed on the image pixels in order to determine the regions from the image and the syntactic features which can give the signature of each region. Our segmentation

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 395–402, 2011.

method, unlike other methods that rely on detecting boundaries between regions, returns a set of closed contours that are accurate polygonal approximation of simple or compound objects. Thus the image segmentation is treated as a hypergraph partitioning problem. The predicate for determining the set of nodes of connected components is based on two important features: the color distance and syntactic features [3], that are geometric properties of regions and their spatial configurations.The rest of the paper is organized as follows. Section 2 presents the color characteristics and hexagonal structure of an image. Section 3 describes our proposed method for hypergraph-based image segmentation and the proposed method for extracting visual object boundaries. Section 4 gives our experimental results and Section 5 concludes the paper.

1.1 Related Work

In this section we consider some of the related work that is most relevant to our approach. The segmentation process of a 2D image can be seen as three major phases[4]: preprocessing phase, feature extraction phase and decision phase. In the preprocessing phase it is removed from the image the information that is undesired for the given domain. The feature extraction phase provides the pixel features such as color and texture, extracted from the preprocessed image. In the final phase, the decision phase, the image is segmented into regions by partitioning the feature space. Segmentation algorithms for $2D$ images may be divided first into homogeneity-based and boundary-based methods [5]. Most graph-based segmentation methods attempt to search a certain structures in the associated edge weighted graph constructed on the image pixels, such as minimum spanning tree [6], or minimum cut [7]. The major concept used in graph-based clustering algorithms is the concept of homogeneity of regions. For color segmentation algorithms the homogeneity of regions is color-based, and thus the edge weights are based on color distance. Early graph-based methods use fixed thresholds and local measures in finding a segmentation. The method [6] uses an adaptive criterion that depends on local properties rather than global ones. The methods based on minimum cuts in a graph are designed to minimize the similarity between pixels that are being split [7]. In [3] a source of additional information denoted by the term of syntactic features is presented, which represent geometric properties of regions and their spatial configurations. Examples of such features include homogeneity, compactness, regularity, inclusion or symmetry. The $IANH$ model is extended in [8] with weight hyperedges and is applied a multilevel hypergraph partitioning technique for segmentation image. Another method using hypergraph structure for image segmentation is presented in [9], where it is proposed a machine learning algorithm for determination of hypergraphs based on seeds-pixels.

2 Color Features and Hexagonal Structure of an Image

The methods which analyze the distribution of the pixel colors in a color space consider that each pixel is represented by a color point with three values. We propose a color image segmentation approach by pixel clustering in a HCL color space. In the first

step we determine the color components corresponding to the HCL color space starting from the RGB color space. There are different algorithms to compute the HCL components; the algorithm used is [10]. In the second step, which corresponds to the pre-processing phase, we add the filters for removing the image noise and consuming less computational time in the segmentation image phase.

2.1 The Hexagonal Grid-Graph Structure

Our technique is based on a new representation of pixels that are integrated into a virtual graph. We use a flexible hexagonal structure as a grid-graph on the image pixels, as presented in Figure 1. The vertices of the hexagonal structure cover half of the pixels from the image. For each hexagon h in this structure there are 6-hexagons, neighbors in a 6-connected sense, and the determination of the indexes for hexagons neighbors, having as input the index of current hexagon is simple. We defined on this structure the two colors list of pixels: $L1$ and $L2$ corresponding to the color of pixels which belong to structure and to the color of complementary pixels which belong to the image, but not belong to the hexagonal grid [11]. The main goal of the structure with hexagons instead

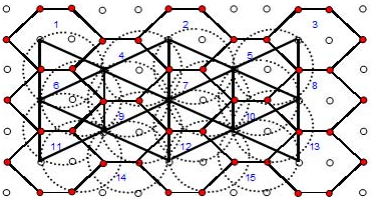


Fig. 1. The hexagonal structure of virtual graph on the image pixels

of pixels as the primitive element is the reduction of the running time for the segmentation algorithms. The grid of hexagons is stored such as a linear vector of numbers $[1 \ldots N]$, where N is the total number of hexagons. The $6 - HCL$ colors associated with an hexagon are used for determining the structure of the hypergraph which represents an image as in 3.1. The mapping of the pixels network on the hexagons network is immediately and it is not time consuming. For an image we have defined the column number of the hexagon grid , $columnNb$, as ($imageWitdh$ - $imageWitdh$ mod 4) / 2. In order to determine the distance between two neighboring hexagons, we use the following color distance formula [10]:

$$D_{HCL} = \sqrt{2 \times (L_1 - L_2)^2 + A_{CH} \times (C_1^2 + C_2^2 - 2C_1C_2cos(H_1 - H_2)} \quad (1)$$

where the value for A_{CH} is $(H_1 - H_2) + 0.16$. If the distance is less than a threshold then the current neighbor hexagon is added to the current region of hexagons. The threshold value is determined for each image in the pre-processing task as average between $sMedDist$, the mean value of the distances and $devDist$, the standard deviation of the distances.

3 HyperGraph-Based Image Segmentation

The proposed segmentation method produces a proper segmentation of pixels which are mapped on the hexagon network according to the definition of distance in HCL color space. In the subsection 3.2 we describe the algorithm for determination of the color regions. The dimension of the color regions list is given by the number of distinct colors which are founded in the image in accordance with the distance computed with formula 1 and with the threshold. Elements from the color list are members of one or more elements corresponding to the regions with the same color but distinct in the image space. For each sub elements we stored the attachment list of hexagons and the last hexagon of this list, which is a hexagon from the contour of the region because we cross the hexagon network from the center of the region to the border of the region.

3.1 Hypergraph Construction Algorithm

A hypergraph is defined as pair $HG = (V;E)$, where $V = v_1, v_2 \ldots v_n$ is the set of vertices and $E = E_1, E_2 \ldots E_m$, with E_i included in V for $i = 1 \ldots m$, is the set of hyperedges. If $G(V; E)$ is a graph then the hypergraph having the vertices of G as vertices and the neighborhood of these vertices as hyperedges is called the neighborhood hypergraph of graph G [8]. Initially, we worked on modeling of image as an hypergraph, taking into account the color distance on the hexagonal-grid. We use two strategies: the local information represented as color distance between two neighboring hexagons and the global information represented as color distance between hexagon which is the center of the current hyperedge and the current hexagon. In the initial phase for each hexagon we have an hyperedge which can connect it with $0 \ldots 6$ neighboring hexagons. The decision to add an edge formed by two neighboring hexagons to an hyperedge is made using the algorithm 1. In Figure 2 there is presented an example of hypergraph representation for an area of an image, obtained by applying the algorithm 1. By using as processing unit element a hexagon, any two neighboring elements have in common two vertexes (an edge). Thus, the color distance between two hexagons is given by the color weight of the common edge.

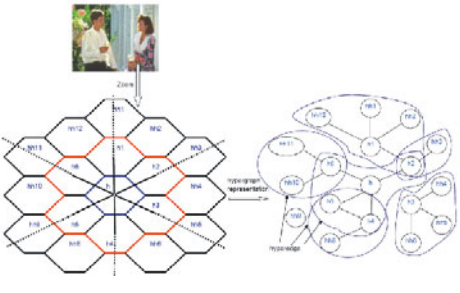


Fig. 2. An example of hypergraph representation

Algorithm 1. Procedure createHyperedge

Input: The index of an hexagon $index_h$
The column number of the hexagon grid-graph $columnNb$
The two list of color pixels: $L_1 = \{p_1, \ldots, p_{6n}\}, L_2 = \{p_1^c, \ldots, p_{6n}^c\}$
Output: The hyperedge corresponding to the hexagon $index_h$

```
1  * init hyperEdge and determine the neighbors hexagon for hexagon index_h;
2  for k ← 1 to 6 do
3      * compute the color distance between first pixel of current hexagon and
       neighborHexagon[k];
4      if * color Distance <= threshold distance then
5          * add edge [index_h,neighborHexagon[k]] to hyperEdge;
6      end
7  end
```

The algorithm 1 create an hyperedge for each hexagon from the initial grid. For this goal are determined the neighbors hexagon and are computed the color distances using the formula 1. If the condition is satisfied the edge create by the hexagon and the neighbor hexagon are added to the hyperedge. This procedure transform the initial hexagonal-grid in the hypergraph-hexagonal grid as a representation of an image.

3.2 Segmentation Algorithm

In this subsection there is presented the algorithm 2 for determining the regions of the image. The output of this algorithm is a composed list corresponding to the region colors from the image.

Algorithm 2. Procedure HyperSegmentation

Input: The total number of hexagons from hexagonal grid n
The hypergraph representation for an image (HG), obtained with algorithm 1
Output: A compose list with the hexagons of regions
$C = \{\{c_1\}, \ldots, \{c_k\}\}, c_i = \{color, \{r_1, \ldots, r_p\}\}$

```
1  * initialize the HGStack ;
2  * get indexH as index of the first unmarked hexagon
3  while * exist unmarked hexagon do
4      * mark as visited the hexagon indexH and push in HGStack the hyperedge
       corresponding to hexagon indexH;
5      while !empty(HGStack) do
6          * pop an hyperedge from HGStack
7          * the current hexagon is the first hexagon of the current hyperedge
8          * add the current hexagon to current region and mark hexagon
9          * push in HGStack the next unmarked hexagon of the current hyperedge
10     end
11     * add region to the list and get indexH as index of unmarked hexagon
12 end
```

The procedure $HyperSegmentation$ returns a composed list; the elements of composed list are determined for each distinct color from the input image as a list of regions that contains hexagons which have the same dominant color. The current region is an instance of the class $ColorRegion$ and represents the data structure corresponding to an item from the output list. The attributes of the class $ColorRegion$ are: the color of the region and the list of hexagons which have the same color. The current hyperedge is represented as a data structure which stores the index of the current hyperedge (managed by the stack), the attribute visit used to process once an hyperedge of the hypergraph.

Proposition 1: The running time of the procedure $HyperSegmentation$ is $O(n^2)$, where n is the number of the grid-graph of hexagons.

Proof: In the exterior WHILE loop it is considered each hexagon from the hexagonal grid structure represented by the list with length n. In the interior WHILE loop it is considered each hyperedge from the hypergraph hexagonal grid which are n elements. The running time of the entire algorithm is therefore $O(n^2)$.

3.3 Boundary Extraction of Visual Objects

In this section we present the method for extraction of the contours of regions. For each object from the image we detect the closed contour which is the accurate polygonal approximation. We choose the last added hexagon as the current hexagon and we determine his neighbors with at least one neighbor from the exterior of the current region. In this way we decide if a hexagon is a contour hexagon and add it to the contour list regions. When the contour is closed, the algorithm ends and there will be returned the list of hexagons which formed the contour of the current region. The procedure which detect the contours of the regions returns the list of the hexagons from the contour of the current region. From the construction of the list with hexagons corresponding to the same region (the procedure $HyperSegmentation$), the last item from the input list L is a hexagon from the contour (the current hexagon is initially set to h_r). We determined all the neighbors hexagons of the current hexagon which belongs to the list L and determined for each neighbor hexagon if is a hexagon from contour.

4 Experiments

We tested our method for image segmentation on a Berkeley Segmentation Dataset color image database ($BSDB$). For evaluating the performance of the our segmentation method we used two quality measures: precision and recall which give us the fraction of detections that are true positives rather than false positives and respectively the fraction of true positives that are detected rather than missed. We include for comparison the segmentation results obtained with other two alternative segmentation algorithms: Mean-Shift [12] and Local Variation [6]. In Figure 3 there are shown examples of segmentation images from $BSDB$. The evaluation of the proposed segmentation method is performed with precision-recall metric [13]. We implemented our precision-recall framework based on the contours of the visual objects. For combining the quantities of

Fig. 3. Examples of images segmented. From left to right: Human segmentation, Our segmentation, Mean-Shift segmentation and Local Variation segmentation.

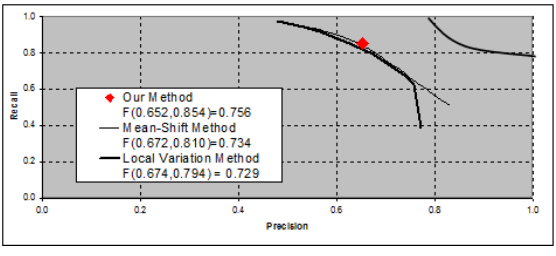


Fig. 4. Precision-Recall Curve and F-Value Maximal

Precision (P) and Recall (R) in a single quality measure, we used the harmonic mean function F. The maximal value of function F gives us the performance of segmentation on a set of images. We use for evaluation the test set of 100 images and we determine the average values for P and R for the test set of images. Because the other evaluated methods are parameterized we use 15 different values for parameters as follows: for Local Variation we tested the input parameter k within [100; 1500] and Mean-Shift algorithm has two main parameters: the spatial bandwidth [1; 15], and the range bandwidth [10; 25]. We obtained a representation in the $Precision - Recall$ diagram for each method and we retained one point (P;R) which represents the best $F - measure$. In Figure 4 it is shown the $Precision - Recall$ curves of the segmentation method and maximal $F - value$ 0.756 (diamond point) obtained using flexible hexagonal grid-graph structure. The values obtained for the human segmentations which are provided with the $BSDB$ image dataset are represented by the top right curve.

5 Conclusions

In this paper we presented a method for image segmentation and extraction of the contours for regions. The novelty of our contribution concerns three aspects: (a) In order to minimize the running time we construct a hexagonal structure based on the image

pixels and we use as data structure, an hypergraph structure; (b) We proposed an efficient method for segmentation of color images based on color distance and hypergraph partitioning algorithm; (c) We developed a fast method for extracting the contour of the detected regions, which will be used in the shape recognition phase. The experiments showed that the segmentation can be yielded with good results regardless of the area of the images that come.

Acknowledgment

The support of the The National University Research Council under Grant CNCSIS IDEI 535 is gratefully acknowledged.

References

1. Jacobs, D.: Robust and efficient detection of salient convex groups. IEEE Trans. on Pattern Analysis and Machine Intelligence 18, 23–37 (1996)
2. Bretto, A., Gillibert, L.: Hypergraph Based Image Representation. In: Graph Based Representations in Pattern Recognition, pp. 1–11 (2005)
3. Bennstrom, C.F., Casas, J.R.: Binary-partition-tree creation using a quasi-inclusion criterion. In: Proc. of the Eighth International Conference on Information Visualization, pp. 259–294 (2004)
4. Gonzalez, W.: Digital image processing, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
5. Salembier, P., Marques, F.: Region-Based Representations of Image and Video: Segmentation Tools for Multimedia Services. IEEE Trans. on Circuits and Systems for Video Technology 9 (1999)
6. Felzenszwalb, P.F., Huttenlocher, W.D.: Efficient Graph-Based Image Segmentation. Intl. Journal of Computer Vision, 167–181 (2004)
7. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 731–737 (1997)
8. Rital, S., Cherifi, H., Miguet, S.: Weighted Adaptive Neighborhood Hypergraph Partitioning for Image Segmentation. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3687, pp. 522–531. Springer, Heidelberg (2005)
9. Ding, L., Yilmaz, A.: Image Segmentation as Learning on Hypergraphs. In: Proc. International Conference on Machine Learning and Applications (2008)
10. Sarifuddin, M., Missaoui, R.: HCL: a new Color Space for a more Effective Content-based Image Retrieval. Departement d'informatique et d'ingenierie, Universite du Quebec en Outaouais (2005)
11. Burdescu, D.D., Brezovan, M., Ganea, E., Stanescu, L.: A New Method for Segmentation of Images Represented in a HSV Color Space. Advanced Concepts for Intelligent Vision Systems (2009)
12. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: Color image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 750–755 (1997)
13. Fowlkes, C., Martin, D., Malik, J.: Learning affinity functions for image segmentation: combining patch-based and gradient-based approaches. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, pp. 54–61 (2003)

Statistical and Wavelet Based Texture Features for Fish Oocytes Classification

Encarnación González-Rufino[1], Pilar Carrión[1], Arno Formella[1], Manuel Fernández-Delgado[2], and Eva Cernadas[2]

[1] Computer Science Department, Univ. de Vigo, Campus As Lagoas 32004 Ourense, Spain
nrufino@uvigo.es

[2] Dept. of Electronics and Computer Science, Univ. de Santiago de Compostela Campus Vida, 15872, Santiago de Compostela, Spain

Abstract. The study of biology and population dynamics of fish species requires the estimation of fecundity parameters in individual fish in many fisheries laboratories. The traditional procedure used in fisheries research is to classify and count the oocytes manually on a subsample of known weight of the ovary, and to measure few oocytes under a binocular microscope. With an adequate interactive tool, this process might be done on a computer. However, in both cases the task is very time consuming, with the obvious consequence that fecundity studies are not conducted routinely. In this work we develop a computer vision system for the classification of oocytes using texture features in histological images. The system is structured in three stages: 1) extraction of the oocyte from the original image; 2) calculation of a texture feature vector for each oocyte; and 3) classification of the oocytes using this feature vector. A statistical evaluation of the proposed system is presented and discussed.

Keywords: Image analysis, Texture classification, Fish oocyte, Statistical classifiers, Classification trees, Neural networks.

1 Introduction

The description of the reproductive strategies and the assessment of fecundity are fundamental topics in the study of biology and population dynamics of fish species. Studies about reproduction, including aspects such as fecundity, assessment of size at maturity, duration of the reproductive season, daily spawning behavior and spawning fraction, allow a quantification of the reproductive capacity of an individual fish [1]. This information increases the knowledge that fisheries researchers need in order to improve the standard assessments of many commercially valuable fish species. To estimate fish female fecundity, researchers need to count the number of oocytes (i.e., ovarian cells precursors of the ovules) that are developing during the breeding season.

Researchers in this field use stereological techniques [2] on histological sections to estimate fecundity and other reproductive parameters. Figure 1 (upper panels) shows histological sections of fish ovary including several oocytes. The

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 403–410, 2011.

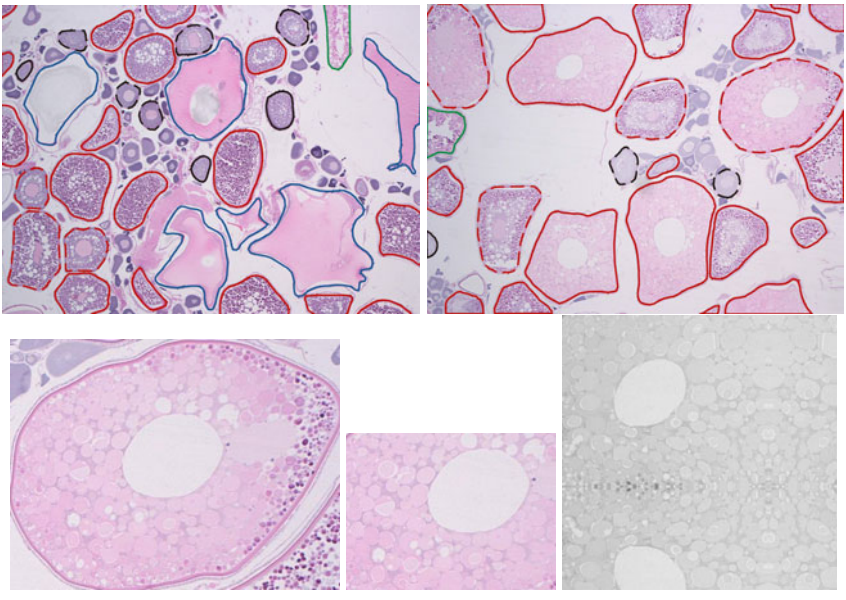


Fig. 1. Upper panels: typical histological images of a fish ovary. Dotted line: with nucleus (WN); continuous line: without nucleus (WON). Black: alveoli corticales (AC); blue: hydrated (HID); red: vitelline (V); green: atresic (AT). **Lower panels**: Digital images of oocytes. Left: oocyte. Center: maximum rectangle of oocyte (389×292 pixels). Right: result of applying the Mirror Expand Method (size 512 × 512 pixels).

most usual method to count the oocytes is Weibel stereometry [3], which requires a classification of the oocytes depending on the presence/absence of the nucleus, and their state of development ("alveoli corticales", "hydrated" and "vitelline/atresic").

We developed a computer vision system that automatically classifies oocytes in histological images and runs on-line. The paper is organized as follows: section 2 describes the image acquisition process and the proposed methods for the different stages of the system, and Section 3 discusses the statistical results achieved so far. Finally, Section 4 draws the conclusions.

2 Methods

The image acquisition process uses standard histological procedures to section the ovaries. The sections are stained with *Haematoxylin–Eosin* and captured with a *LEICA*® *DRE* research microscope with a total magnification of 6.3 and a calibration of 0.547619 microns per pixel. A *LEICA*® *DFC320* digital camera, directly connected to the microscope, combined with the *LEICA IM50*© is used to digitalize the images. The camera resolution is 3.3 megapixels (2088 × 1550 pixels), using square pixels of 3.45 μm. The exposure time and color balance are set automatically.

Our computer system is composed of three stages: 1) extraction of the oocyte from the image; 2) calculation of a texture feature vector for each oocyte; and 3) classification of the oocytes using this feature vector. We studied and evaluated the automatic detection of oocytes in previous works [4]. In this paper, our objective is to evaluate the capability of several texture features combined with well-known classifiers to determine the *class* of each oocyte ("with nucleus" and "without nucleus") and its *state* of development ("alveoli corticales", "hydrated" and "vitelline/atresic"). Since the oocyte contour detected by the computer are still not exact, we used the true contours annotated by the human experts in order to compute texture features without precision loss or error accumulation. We computed all the texture features over square regions (whose side is a power of 2) in the image, in order to allow the comparison with some of them, which required such a square region. So, our region of interest (ROI) extractor selects a square region in the true oocyte contour. In the following subsections we describe the ROI extractor, the texture features and the classifiers used.

2.1 Square ROI Extraction

First, we extract the maximum rectangle contained in the oocyte using the Vandevoorde's maximum rectangle algorithm [5], whose computational complexity is linear in the number of pixels of the image. Once we have a rectangular image, there are several methods available to extract a square region: *crop*, *mirror expand* and *black*. The latter two expand the image to be the nearest power of two square above the actual size, filling the square with black pixels (*black*) or with a mirror copy of their neighbors (*mirror expand*). The image to be the nearest power of two square below the actual size is kept in the *crop* method. We did not consider the *black* method because it introduces many artefacts in the image. The *crop* was also discarded because it can not be applied to the smallest oocytes. The lower panels of figure 1 shows an example of the extraction process of an oocyte with the mirror expand method.

2.2 Texture Feature Generation

We use some of the most popular texture features proposed in the literature ([6,7]). **First Order Statistics (FOS)** provide information of the grey level distribution in the image. Let x be the random variable representing the grey levels in the region of interest, $P(x)$ be its histogram and N_g be the possible number of grey levels. We define the following first-order statistics: a) *moments* ($m_i = \sum_{x=0}^{N_g-1} x^i P(x),\ i = 1, 2$), b) *central moments* ($\mu_i = \sum_{x=0}^{N_g-1} (x - m_1)^i P(x),\ i = 2, 3, 4$), where m_1 is the mean grey level of the region, c) *absolute central moments* ($\hat{\mu}_i = E[|(x - E[x])|^i],\ i = 2, 3, 4$) and d) *entropy* ($H$).

Second order statistics provide information about the relative positions of the grey levels within the image: 1) **Haralick Coefficients (HC)**: the Grey Level Co–occurrence Matrix (GLCM) of the image encode the repeated occurrence of some grey-level configuration separated by a distance d and direction α. This matrix could be used as a texture–feature vector. However, it is common to use

the following derived features from the matrix: energy, entropy, correlation, inverse difference moment, inertia, cluster shade and cluster prominence. 2) **Grey Level Run Length Statistics (GLRLS)**: a set of consecutive pixels in the image having the same grey level value is called grey level run. The *length of the run* is the number of pixels in the run. So, a run length matrix results from which the following features are derived: short run emphasis, long run emphasis, grey level non-uniformity, run length non-uniformity and run percentage. 3) The **Neighboring Grey Level Dependence Statistics (NGLDS)** consider the relationship between an element and all its neighboring elements at one time. It is based on the calculation of a grey level spatial dependence matrix of an image. The usual numerical measures calculated from this matrix are: small number emphasis, large number emphasis, number non-uniformity, second moment and entropy. The common value for distance to neighbors is 1 and for the difference of grey levels is 0.

Wavelet transform is used for multiscale filtering providing information about the image contained in the time-frequency domain. The application of the discrete wavelet transform and variants thereof for texture identification have received considerable attention in the literature. Wavelet packets are a generalization of orthonormal and compactly supported wavelets. For two-dimensional images, the basic function can be expressed as the tensor product of two one-dimensional basis functions in the horizontal and vertical directions, with corresponding 2-D filter coefficients that represent low-pass and high-pass filtering effects in the x and y direction, respectively. Daubechies wavelet are orthonormal, regular wavelets with compact support and they are therefore suitable for analysis of signals with finite support, particularly for image analysis. The orthogonality condition ensures that the representations of the signals at different levels of decomposition are uncorrelated. The regularity condition provides a sufficient decay of the mother wavelet in the frequency domain which makes it continuous in the spatial domain. We computed the energy (E), entropy (H), variance (μ_2), 3^{rd} (μ_3), and 4^{th} (μ_4) statistical moments for the decomposed wavelet packets, using the libwavelet [1] (version 1.3.1) with Daubechies-8 filter.

2.3 Classification

When the texture features are computed, the classification stage assigns each query case to a pre-established class and state. We consider two classes—oocytes "with nucleus" (WN) and "without nucleus" (WON)—and three states of development—"alveoli corticales" (AC), "hydrated" (HID) and "vitelline /atresic" (V/AT). Both, classes and states, are labelled by expert biologists. For both classifications we tried classifiers belonging to three Machine Learning paradigms: 1) Statistical classifiers: **K-Nearest Neighbors (KNN)** [8], which assigns a test pattern to the most voted class among its nearest neighbors in the training set. We used the R (version 2.7.1) implementation of KNN (package *class*) with the Euclidean distance, tuning the number of neighbors with values in the

[1] `http://sourceforge.net/projects/wavelet/`

range 1:20. 2) Tree-based classifiers: **Adaboost** [9] of classification trees. Adaboost is a meta-algorithm which creates an ensemble of base classifiers which focuses on the most difficult patterns. Initially, the whole training set is used for the first classifier, and for each step a new classifier is added, trained with a set where the patterns missclassified at the previous step are more probable. We used the R implementation (function `adaboost.M1`, package *adabag*). 3) Connectionist classifiers: **Support Vector Machine (SVM)** and **Multi-Layer Perceptron (MLP)**. The SVM [10] is based on the Statistical Learning Theory, and it trains efficiently a linear classifier in the high-dimensional hidden (or feature) space, using a kernel mapping to achieve linear separability in the hidden space. We used the LibSVM[2] (version 2.84) library with Gaussian kernels, tuning the regularization parameter (C) and the kernel spread in the ranges $\{2^n, n = -5, \ldots, 14\}$ and $\{2^n, n = -15, \ldots, 0\}$ respectively. The MLP is a classical feed-forward neural network trained with the Back-propagation algorithm [8]. We used the MLP implementation provided by R (package *nnet*), tuning the number of hidden neurons with values in the set $\{3, 5, 7, 9, 12, 15\}$.

3 Results and Discussion

Our system was tested to classify oocytes in two classes (with and without nucleus) and three states of development ("alveoli corticales", "hydrated" and "vitelline/atresic") using 20 histological images with a total of 533 oocytes. We extracted the square ROI (Section 2.1) considering the *mirror expand* method, and we calculated the texture features (Section 2.2) for all the oocytes. Features were grouped in two vectors: 1) statistical features (FOS, HC, GLRLS and NGLDS); and 2) wavelet features, including the five statistics (E, H, μ_2, μ_3 and μ_4) calculated over the three decomposed wavelet packets.

We randomly generated 10 sets of patterns (given by permutations of the patterns) for training, validation and test including 50%, 25%, and 25% of the available patterns (267, 133 and 133 patterns, respectively). Each classifier (Section 2.3) was trained on the 10 training sets, and its tuning parameters were selected to maximize the average classification accuracy over the 10 validation sets. Finally, the average performance of each classifier with its best parameter values was evaluated on the 10 test sets (two upper rows in table 1).

In both classifications (classes and states) the statistical features show higher discrimination capability than wavelet features, providing the best results for all classifiers, with an average difference of 6%±2.4 for classes and 2%±0.7 for states. In the discrimination between classes (WN and WON) SVM achieves 75.6% of accuracy, followed by KNN (74.6%) and MLP (74.5%). We must take into account the class WON has 63.97% of the patterns, so the base accuracy is 63.97% when all the patterns are assigned to WON. Therefore, although the four approaches seem to achieve similar accuracies, relatively small differences in the average accuracy may lead to high differences in the class sensitivities. For instance, SVM (75.6% accuracy) achieves high sensitivities (80.3% and 72.6%

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Table 1. Two upper rows: average test accuracy (in %) achieved by KNN, Adaboost of classification trees, SVM and MLP using statistical and wavelet features. Lower rows: results using feature subsets. The results better than with the whole feature set are in bold. N is the number of features.

			Classes				States			
		N	KNN	Adaboost	SVM	MLP	KNN	Adaboost	SVM	MLP
Statistical		26	74.6	73.4	**75.6**	74.5	85.3	**88.5**	85.9	83.9
Wavelet		60	68.1	70.3	67.1	66.6	82.8	85.9	84.1	82.8
Statistical	FOS	9	67.8	66.6	65.1	63.0	80.5	84.2	79.4	83.3
	HC	7	73.5	68.9	62.9	74.1	81.7	78.1	59.6	82.0
	GLRLS	5	**77.3**	71.6	62.9	**78.3**	82.6	75.6	50.0	81.1
	NGLDS	5	74.3	72.9	60.1	74.9	81.8	81.7	64.8	83.5
Wavelet	E	12	64.7	67.9	60.1	60.5	79.2	81.3	66.2	78.3
	H	12	60.8	64.7	60.1	58.6	75.3	82.2	64.1	68.9
	μ_2	12	67.6	68.8	57.7	62.2	77.8	81.4	64.9	77.9
	μ_3	12	64.4	66.1	59.5	59.9	77.8	79.9	67.1	73.5
	μ_4	12	63.2	68.4	56.5	60.5	78.5	80.5	65.9	77.0
	Level 1	20	62.5	68.7	62.6	59.8	75.9	84.6	78.7	75.9
	Level 2	20	65.9	65.9	58.1	62.6	77.2	83.7	74.6	74.8
	Level 3	20	66.1	70.8	62.3	63.3	82.9	86.4	76.2	82.1

for WN and WON respectively), so that the discrimination is quite reliable. On the contrary, KNN (74.6% accuracy) achieves much lower sensitivities (61.2% for WN and 82.3% for WON), being less reliable than SVM. The same happens with the other classifiers.

In the discrimination among states of development (AC, HID and V/AT), Adaboost achieves 88.5% of accuracy, followed by SVM (85.9%) and KNN (85.3%). The state V/AT has 71.1% of the patterns, and the state HID has only 8.44%, so that relatively high accuracies may give low sensitivities. Adaboost achieves sensitivities above 80% for the three states (83.8% for AC, 81.9% for HID and 90.3% for V/AT), while SVM achieves low sensitivity (66.8%) for class AC, with better values for HID and V/AT (80.4% and 92.5% respectively). Thus, the relatively low difference between the accuracies of Adaboost and SVM has a big impact in the state sensitivities.

Table 2 reports the average confusion matrices (in %) over the 10 permutations achieved by SVM for classes and Adaboost for states using the statistical features. The items in this matrices represent the percentage of correct and incorrect classification of the expected class (true class, in rows) in relation to the observed class (class provided by the classifier, in columns). The classification is more reliable for WN (sensitivity 80.3%) than for WON (72.6%). With states, the sensitivities are higher: 83.8% for AC, 81.9% for HID and 90.3% for V/AT. Only 16% of AC patterns are misclassified as V/AT, and 7% of V/AT patterns are assigned to AC. The overlapping between HID and V/AT is lower (2% and 18%), and states AC and HID are discriminated very well (0.6% and 0%).

Table 2. Average (in %) confusion matrices of SVM for classes—with nucleus (WN) and without nucleus (WON)—and Adaboost for states of development—vitelline/atresic (V/AT), hydrated (HID) and alveoli corticales (AC)—both using statistical features

SVM	WN	WON	Adaboost	AC	HID	V/AT
WN	29.8	7.3	AC	14.5	0.1	2.7
WON	17.1	45.5	HID	0.0	5.9	1.3
			V/AT	5.4	1.9	68.2

For a deeper analysis, and in order to reduce the number of features required, we applied the classifiers to texture features subsets (lower rows in table 1 above). We divided the statistical textures feature in four subsets: First–Order Statistics (FOS), Haralick coefficients (HC), Grey Level Run Length Statistics (GLRLS) and Neighboring Grey Level Dependence Statistics (NGLDS). With wavelet texture features, we developed an analysis considering each feature (E, H, μ_2, μ_3 and μ_4) over all the decomposed wavelet packets and all the statistics over each decomposed wavelet packet (Levels 1, 2 and 3). Usually, the results achieved with subsets are slightly worse than the results corresponding to whole feature sets. However, there are two exceptions in the class discrimination: MLP achieves 78.3%, and KNN 77.3%, both using the GLRLS subset (SVM achieved 75.6% using the whole feature set). Thus, GLRLS (with only 5 features) provides better results than the whole set (26 features), with sensitivities (74.6% and 80.4% for MLP and classes WN and WON respectively) similar or better than SVM (80.3% and 72.6%) using the whole set.

The good results of statistical features compared to wavelet features, both for class and state discrimination, is relevant. In fact, in the previous experiments we computed the statistical features over square regions in order to allow this comparison. Therefore, we developed additional experiments with the same classifiers using some statistical features (specifically, the Haralick coefficients) over the true oocyte, which is an irregular region. With classes, we achieved 72.3% accuracy using SVM (worse than SVM using statistical features in a square region, 75.6%), but with low sensitivities (43.7% and 86.9% for WN and WON, respectively). In the state classification, all the classifiers achieved better results using irregular regions. Specifically, SVM achieved 90.1%, with high sensitivities (85.1%, 84.5% and 92.4% for AC, HID and V/AT, respectively), clearly better than Adaboost using square regions (88.5%, sensitivities 83.5%, 82.3% and 90.4%). These results suggest that the computation of statistical features over square regions reduces the information available to classify oocytes, at least for the discrimination among states.

4 Conclusion

In this work we test the performance of a wide set of texture features (wavelets and statistics of first and second order) and several classifiers (KNN, Adaboost

of classification trees, SVM and MLP) to determine the presence/absence of nucleus and the state of development of fish oocytes in histological images. The statistical textures features (26 inputs) provided better results in both classifications compared to wavelet features (60 features). The best performance was achieved by SVM to discriminate between classes (75.6% accuracy, with sensitivities between 70-80%) and by Adaboost of classification trees to discriminate among states (88.5%, with sensitivities above 80%). We tested feature subsets, and for the detection of nucleus in the oocyte we achieved slightly better results (78.3%) using MLP and the GLRLS subset (only 5 inputs). The use of the true oocyte to compute the Haralick coefficients revealed better results (90% accuracy, sensitivities above 85%) for state classification with respect to square regions. The future work will include other ensemble classifiers for both problems and other color and/or texture features.

Acknowledgment

This investigation was partly supported by the Xunta de Galicia (regional government) project PGIDIT08MMA010402PR.

References

1. Murua, H., Saborido-Rey, F.: Female reproductive strategies of marine fish species of the North Atlantic. J. of Northwest Atlantic Fishery Science 33, 23–31 (2003)
2. Saborido-Rey, F., Witthames, P., Thorsen, A., Murua, H., Kraus, G., Junquera, S.: Procedures to estimate fecundity of marine fish species in relation to their reproductive strategy. J. of Northwest Atlantic Fishery Science 33, 33–54 (2003)
3. Weibel, E.R., Gómez, D.M.: A principle for counting tissue structures on random sections. J. of Appl. Physiology 17, 343 (1962)
4. Alén, S., Cernadas, E., Formella, A., Domínguez, R., Saborido-Rey, F.: Comparison of region and edge segmentation approaches to recognize fish oocytes in histological images. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4142, pp. 853–864. Springer, Heidelberg (2006)
5. Vandevoorde, D.: The maximal rectangle problem. Dr. Dobb's Journal (April 1998)
6. Mirmehdi, M., Xie, X., Suri, J.: Handbook of texture analysis. Imperial College Press, London (2008)
7. González, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Pearson Prentice Hall, London (2008)
8. Duda, R.O., Hart, P.E., Storck, D.G.: Pattern classification. John Wiley & Sons, Inc., Chichester (2001)
9. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of on-line Learning and an Application to Boosting. In: Proc. of the 2nd European Conf. on Computational Learning Theory, pp. 23–37 (1995)
10. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)

Using Mathematical Morphology for Similarity Search of 3D Objects

Roberto Lam[1] and J.M. Hans du Buf[2]

[1] University of Algarve – Instituto Superior Engenharia, Faro, Portugal
[2] University of Algarve – Vision Laboratory – FCT, Faro, Portugal

Abstract. In this paper we use the erosion and dilation operators for characterizing 3D polygonal objects. The goal is to perform a similarity search in a set of distinct objects. The method applies successive dilations and erosions of the meshes in order to compute the difference volume as a function of the size of the structuring element. Because of appropriate pre-processing, the resulting function is invariant to translation, rotation and mesh resolution. On a set of 32 complex objects with different mesh resolutions, the method achieved an average ranking rate of 1.47, with 23 objects ranked first and 6 objects ranked second.

Keywords: 3D Shape similarity - volume models - manifold meshes - mathematical morphology.

1 Introduction

Because of the increasing number of databases of 3D mesh models there is a strong interest in methods for 3D similarity analysis [13]. Similarity analysis is a fast way to discard many irrelevant objects from a database, i.e., before precise object recognition must be applied by very time-consuming matching methods because of all variations that may occur: different position (object origin), rotation, size and also mesh resolution. Similarity analysis does not require precise shape comparisions, global nor local. Instead, this approach is based on computing a feature vector (FV) of a query object and comparing its FV with all FVs of the known objects in a database. The FVs can be obtained by a variety of methods, from very simple ones (bounding box, area-volume ratio, eccentricity) to very complex ones like the curvature distribution of the sliced volume, spherical harmonics or 3D Fourier coefficients [8]. Mesh smoothing serves to reduce noise, for example for decreasing the mesh size by re-triangulation of planar areas. A characteristic function based on mesh smoothing, which also eliminates structural mesh details, has also been applied to similarity analysis [6].

The theory of mathematical morphology (MM) arose in the middle of the 1960s [7,10]. Envolving geometric analyses of shapes and textures, it became increasingly important in 2D image processing and computer vision. Despite all theoretical developments and generalization to 3D, most MM work is still being applied to 2D image processing [10]. The work done in 3D is rather scarse and mostly limited to three-dimensional surfaces. Jackway [5] developed an approach

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 411–419, 2011.

for the recognition of 3D objects in range data through the matching of local surfaces. Lee et al. [4] analyzed the composition of 3D particle aggregates by processing one hemisphere of the particles.

In this paper we apply MM to similarity search of 3D polygonal objects, with the goal of computing object signatures which could complement other multi-scale signatures, for example those based on mesh smoothing [6]. The rest of this paper is organized as follows: Section 2 briefly reviews the basic concepts of mathematical morphology. Section 3 presents the proposed method and Section 4 the experimental results. We conclude with a discussion in Section 5.

2 Brief Review of Mathematical Morphology

Mathematical morphology is based on set-theoretic concepts and its main operators, erosion and dilation, can be seen as operations between two sets. In image processing the sets are defined in R^2 [10], but in general one can work in R^n [11]. If A is a set of points $p = (x_1, x_2, \ldots, x_N)$ in Euclidean space R^3, then the *binary* set A defines the surface and interior of a 3D object. Furthermore, if B is also a set in R^3, $(B)_x$ denotes the translation of B by x and $\hat{B}$ is the reflection of B. The dilation of A by B is denoted by $A \oplus B$ and defined by

$$A \oplus B = \{x \in R^3 \mid (\hat{B})_x \cap A \neq \emptyset\}. \tag{1}$$

Likewise, the erosion of A by B is denoted by $A \ominus B$ and defined by

$$A \ominus B = \{x \in R^3 \mid (B)_x \in A\}. \tag{2}$$

The set B is called structuring element. In this paper we will only use 3D spheres as structuring elements. If applied to an object A, dilation will lead to a bigger object and erosion to a smaller one, but both new objects will lack detail smaller than the size of the sphere. Such details are also lost when the opening and closing operators are applied, which combine dilation and erosion operators [7,9]. Finally, boundary extraction in 2D, see [9], can be defined by

$$\beta(A) = (A \oplus B) - (A \ominus B). \tag{3}$$

In 3D, the boundary β is calculated on the dilated and eroded versions of object A where B is a sphere with radius r. In the limit case $r \downarrow 0$, β is the surface of A and this is related to the fractal dimension [12]. As for mesh smoothing [6], the basic idea in this paper is to characterize 3D objects by controlled elimination of detail. This is illustrated in 2D in Fig. 1. The left shows a triangle and a square with the structuring element, a circle with radius r, on the corners of the original objects. The dilated objects are bigger (only the contours are shown) and the eroded objects (in grey) are smaller. The surface β between both as a function of radius r is shown to the right: the two curves are linear but have different slopes. This effect will be exploited below in the 3D case.

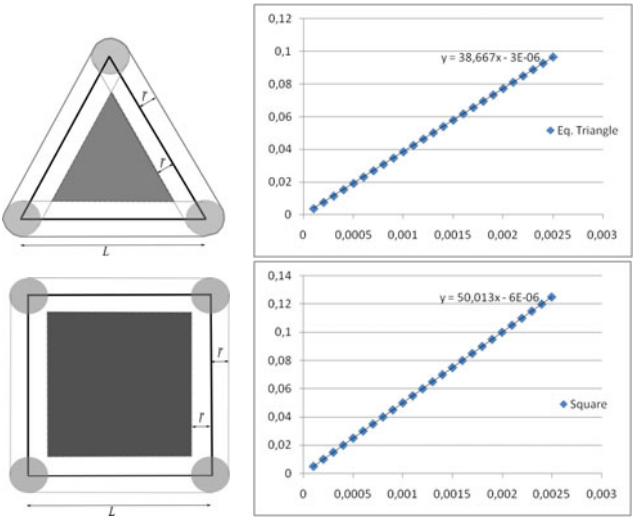


Fig. 1. Left: 2D erosion and dilation of an equilateral triangle and a square. Right: area β as a function of radius r (x axis) of the structuring element.

3 Method

There are a few important issues when applying mathematical morphology to 3D objects. One is associated with the type of representation: voxel or mesh [2,11]. The voxel representation involves 3D arrays with, depending on the object's resolution, very big dimensions, although the voxels themselves are binary: object vs. background. An advantage is that many algorithms from mathematical morphology have been developed for 2D image processing, and these can easily be adapted to 3D. Polygonal meshes, on the other hand, have a more complex data structure. After applying the erosion and dilation operators, the new meshes must be determined, very close vertices can be collapsed, and self-intersecting facets must be detected and removed. In our method we extend boundary extraction (3) from 2D to 3D. Due to the fact that we use polygonal meshes we can apply a similar solution. If $A^c = 1 \backslash A$ is the set outside A, then

$$\beta(A) = A^c \cap (A \oplus B) + A \cap (A \ominus B)^c \tag{4}$$

is the sum of the expanded and shrunken volumes, i.e., the difference volume.

In order to limit distortions in the transformations, we iteratively apply a sphere of which the radius r is a function of edge lenght. To avoid inconsistencies between different mesh resolutions, we select $r = \hat{L}/20$, where $\hat{L}$ is an object's edge length with the maximum occurrence. This can be easily determined by filling a lenght histogram with 50 equal bins from $L_{\min}$ to $L_{\max}$ of each object.

We use a set of 32 models, each one represented by four different mesh resolutions. The models were selected from the AIM@SHAPE database [1].

Fig. 2. Examples of models. Left to right: Elk, Mouse, DancingChildren, Dragon and Eros, with increasing model resolutions.

This database contains high-definition objects which can be converted to other mesh resolutions by means of one parameter between 9.9 (max mesh size) and 5.5 (min mesh size). The models were downloaded in PLY format and only "watertight" ones—closed, without gaps and 2-manifold meshes—were selected. Figure 2 shows a few examples, and Table 1 lists all objects with their mesh resolutions: the first three resolutions are used for extracting feature vectors, the fourth one is used as test object for similarity search. In order to obtain invariance to translation and scale (mesh size), each model was normalized to the unitary sphere (radius of 1.0) after the origin of the model was translated to the center of the sphere. Rotation invariance is achieved by the fact that our characteristic function is global to the model as proven in [14].

Because of the mesh representation, after applying the erosion and dilation operators, (i) vertices within the neighborhood defined by the structuring element should be merged, and (ii) self-intersecting facets must be detected and removed. All this is done without introducing distortions and by keeping the mesh closed, 2-manifold and without degenerated facets.

Table 1. All 32 models with their mesh resolutions, the last resolution was used in similarity search

N	Model	Resolutions	N	Model	Resolutions
1	Bimba	6.0; 8.5; 9.5; 8.0	17	Fish	6.0; 7.5; 8.0; 8.0
2	Blade	6.5; 7.5; 9.9; 8.0	18	Grayloc	6.0; 7.5; 9.9; 7.8
3	Block	5.0; 6.5; 8.0; 8.5	19	GreekSculpture	6.5; 7.0; 7.7; 8.5
4	Bunny	6.5; 7.5; 9.9; 8.0	20	Horse	6.0; 7.5; 9.9; 8.0
5	CamelA	6.0; 7.5; 9.9; 7.8	21	IsidoreHorse	6.0; 7.5; 9.9; 7.0
6	Carter	6.0; 8.5; 9.5; 7.3	22	Kitten	6.0; 7.5; 9.9; 7.3
7	Chair	6.5; 7.5; 9.9; 6.9	23	Liondog	6.0; 7.5; 9.9; 8.0
8	Cow	6.0; 6.4; 9.9; 7.1	24	Mouse	6.0; 7.5; 9.9; 7.8
9	Cow2	6.0; 7.5; 9.9; 8.9	25	Neptune	6.0; 8.0; 9.5; 7.6
10	Dancer	6.0; 7.5; 99; 7.7	26	Pulley	6.0; 7.5; 9.9; 7.0
11	DancingChildren	6.0; 7.5; 9.9; 6.8	27	Ramesses	6.0; 7.5; 9.9; 8.0
12	Dente	6.0; 7.5; 9.9; 7.0	28	Rocker	6.0; 7.5; 9.9; 7.1
13	Dragon	6.0; 8.0; 9.5; 7.7	29	Screwdriver	6.0; 7.5; 9.9; 7.0
14	Duck	6.0; 7.5; 9.9; 6.7	30	Squirrel	6.0; 7.5; 9.9; 7.2
15	Elk	6.0; 7.5; 9.9; 7.9	31	Torso	6.0; 7.5; 9.9; 7.7
16	Eros	6.0; 7.5; 9.9; 6.5	32	Vaselion	6.0; 7.5; 9.9; 8.0

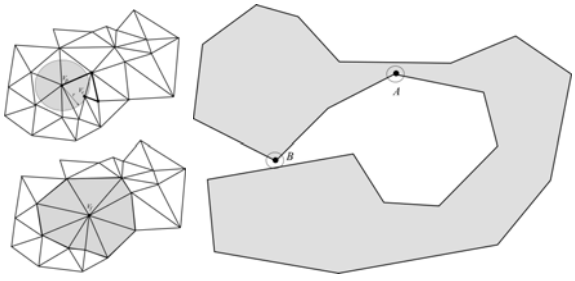


Fig. 3. Left: merging neighboring vertices, before (top) and after (bottom). Right: triangles at vertex A will self-intersect during erosions; those at B during dilations.

Dilations are obtained by displacing all vertices a distance r (the radius) in the direction of the normal vector. Since normal vectors always point outside, this is $-r$ in the case of erosions. Both operators are applied in two distinct steps. The first one is intented to acquire the volumes of the objects after the erosion/dilation process. Each operator is repeatedly applied until the first self-intersection occurs. In this step we do not remove any element of the mesh, vertex nor facet. In the second step we use the dilated (biggest) and the eroded (smallest) objects, generated in the first step, as a new starting point. The operators are repeatedly applied to the corresponding object: erosion to the smallest and dilation to the biggest object. After each erosion/dilation, we search the mesh for vertices that have a neighbor vertex in their vicinity, i.e., in the sphere with radius r centered at the vertex being processed, V_p. If there is a candidate vertex, V_c, it must be connected to V_p by at most 3 edges but it may not possess a direct edge to V_p. These restrictions must be satisfied in order to keep the mesh 2-manifold. The search for the vertices with the shortest path from V_p to V_c is done by using Dijkstra's algorithm. Vertices V_p and V_c are merged by removing all edges and vertices, which causes a gap in the mesh, and then by inserting a new vertex, V_f, with coordinates equal to the average of the removed vertices. In the last step V_f is connected to the vertices forming the gap; see Fig. 3 (left). The elimination of self-intersecting facets is also necessary in situations where the nearest vertex is outside the vicinity sphere, the structuring element. Figure 3 (right) shows two situations which both lead to a self-intersection. Elimination is done using the TransforMesh Library [15], i.e., the method `ensureEdgeSizes` with the flags `fixDegeneracy=Yes` and `smoothing=No`.

4 Results

Applying a sphere as structuring element to all vertices leads to a smaller object in case of erosion and a bigger object in case of dilation. In the case of the Horse model, repeated erosion will cause discontinuity of the legs. The processed Horse models as shown in Fig. 4 (left) were obtained using $r = 0.0170$ in erosion and $r = 0.0259$ in dilation at model resolution 7.5. The small stumps, shown in red

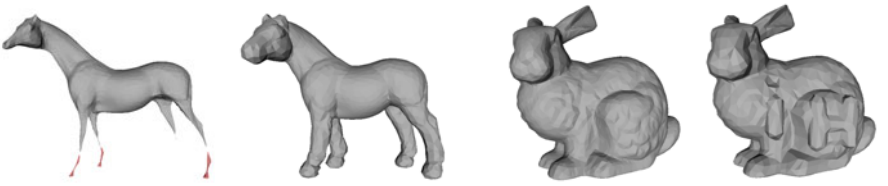


Fig. 4. Left to right: Horse (res. 7.5) after erosion and dilation, Bunny and Bunny iH

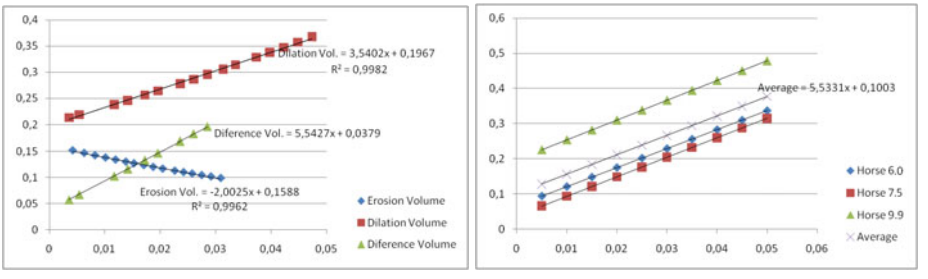


Fig. 5. Difference volume of Horse model as a function of radius r. Mesh resolution 7.5 (left) and average difference function of resolutions 6.0, 7.5 and 9.9 (right).

and created by erosion, were deleted and their volumes were excluded from the computation of the Horse's volume. The same procedure was applied to the other models.

For each model resolution, the difference volume defined as dilated volume minus eroded volume according to Eq. (4), yields an approximately linear function of the radius of the structuring element, see Fig. 5. After least-squares line fitting by $Ar + B$, the slope coefficient A reflects the complexity of the surface of the object. The coefficient B also reflects the complexity, but with emphasis on the capacity of the object to be eroded and dilated without self-intersections, i.e., the first step (i) of the two-step process as described in Section 3. The entire procedure can be summarized as follows. First, in feature extraction, a model is eroded and dilated until the first self-intersection occurs, in the eroded model and the dilated one separately. From these two modified models we compute the difference volume V_i at each mesh resolution i as listed in Table 1, from which the average difference volume V is computed.

Then, for each mesh resolution, both models are further eroded and dilated and the difference volumes as a function of total sphere radius r are computed. From these functions the line parameters A and B are determined by linear regression. Again, these parameters are averaged over the first three mesh resolutions as given in Table 1.

The same parameters are computed in the case of the fourth mesh resolution given in Table 1. Finally, the parameters of the fourth mesh resolution of each model are compared with the averaged parameters of the first three mesh

resolutions of all models, and the models are ranked using the Euclidean distance of the vectors (A, B, V). Table 2 lists the first six ranking positions with increasing Euclidean distance. The six ranking positions are due to object number 13 (Dragon), which gave the worst result (see below). An additional test involved a modified object named "Bunny iH", see Fig. 4 (right), which was not part of the dataset. Bunny iH was correctly ranked first as Bunny, object number 4.

If R_i is the correct ranking position of object i, which is shown boldfaced in Table 2, then the average ranking rate is $\bar{R} = \frac{1}{32} \sum_{i=1}^{32} R_i$. Our results are summarized by $\bar{R} = 1.47$, with a root-mean-square (RMS) error of 0.973. Although the distribution is asymmetric, this means that most objects were ranked first or second. Indeed, in Table 2 we can see that 23 objects of all 32 were ranked first, 6 objects second and only 2 third. Object number 13 (Dragon) was an exception: Fig. 2 shows the ranked models in the Dragon query. The dataset tested is too small to compute advanced performance measures as used in the SHREC contest, but our correct recognition rate of $23/32 = 0.72$ is at the top of the range between 0.45 and 0.70 achieved in the 2010 contest [3].

Table 2. Results

N	Test model	Ranking	N	Test model	Ranking
1	Blade	**1**-4-12-22-28-13	17	Fish	27-**17**-2-18-9-31
2	Bimba	**2**-27-18-17-9-21	18	Grayloc	26-6-**18**-3-14-23
3	Block	**3**-14-23-32-30-11	19	GreekSculpture	**19**-8-20-7-31-9
4	Bunny	1-**4**-22-12-28-13	20	Horse	**20**-7-19-25-5-29
5	CamelA	**5**-25-29-7-20-19	21	IsidoreHorse	**21**-2-27-18-17-9
6	Carter	26-**6**-18-3-14-23	22	Kitten	**22**-4-1-12-28-13
7	Chair	**7**-20-25-5-29-19	23	Liondog	14-3-**23**-32-30-16
8	Cow	31-**8**-9-17-18-27	24	Mouse	16-**24**-15-13-11-30
9	Cow2	**9**-17-31-18-27-8	25	Neptune	**25**-5-29-7-20-19
10	Dancer	**10**-29-5-25-7-20	26	Pulley	**26**-6-18-3-14-23
11	DancingChildren	**11**-15-16-30-24-13	27	Ramesses	**27**-18-17-2-9-31
12	Dente	**12**-1-4-22-13-24	28	Rocker	**28**-21-22-4-1-12
13	Dragon	16-15-24-11-30-**13**	29	Screwdriver	5-**29**-25-7-20-19
14	Duck	**14**-23-3-32-30-16	30	Squirrel	**30**-16-11-15-24-13
15	Elk	**15**-11-16-24-13-30	31	Torso	**31**-9-8-17-18-27
16	Eros	**16**-15-24-11-30-13	32	Vaselion	**32**-23-30-14-3-16
	Bunny iH	**4**-22-1-12-28-13			

5 Discussion

Taking into account the complexity of the mesh models, the results obtained are quite good. The average ranking rate of 1.47, with 23 objects ranked first and six objects ranked second, was achieved by using only three parameters: the difference volume V of the eroded and dilated objects until the first self-intersections occurred, plus the linear regression parameters A and B of the difference volumes after additional erosions and dilations. This means that these parameters,

which reflect systematic object deformations with decreasing detail as a function of increasing structuring element, are capable of characterizing different shapes. However, this does not mean that similar or other signatures based on mathematical morphology cannot perform even better; it makes sense to experiment with other signatures in the future. In addition, two of our parameters provide multi-scale signatures because of the increasing structuring element, and it makes sense to combine these parameters with other multi-scale signatures, for example those based on mesh smoothing [6]. The latter method achieved on a similar set of objects an average ranking rate of 1.29, but seven instead of three parameters were used. Therefore, a general conclusion might be that the combination of perhaps two multi-scale parameter sets with perhaps as much as six parameters will boost performance.

Acknowledgements

Research supported by the Portuguese Foundation for Science and Technology (FCT), through the pluriannual funding of the Inst. for Systems and Robotics through the POS_Conhecimento Program (includes FEDER funds), and by the FCT project SmartVision: active vision for the blind (PTDC/EIA /73633/2006).

References

1. AIM@SHAPE (2008), `http://www.aimatshape.net`
2. Campbell, R., Flynn, P.: A survey of free-form object representation and recognition techniques. Computer Vision and Image Understanding 81(2), 166–210 (2001)
3. Veltkamp, R.C., et al.: SHREC 2010 track: Large scale retrieval. In: Proc. of the Eurographics/ACM SIGGRAPH Symp. on 3D Object Retrieval, pp. 63–69 (2010)
4. Smith, L., Lee, J., Smith, M., Midha, P.: A mathematical morphology approach to image based 3D particle shape analysis. Machine Vision and Applications 16(5), 282–288 (2005)
5. Jackway, P.T.: Morphological Scale-Space with Application to Three-Dimensional Object Recognition. PhD thesis, Queensland University of Technology, Australia, Supervisor-Boles, W. W (1995)
6. Lam, R., Hans du Buf, J.M.: Invariant categorisation of polygonal objects using multi-resolution signatures. In: Proc. KDIR, pp. 168–173 (2009)
7. Matheron, G.: Random sets and integral geometry. John Wiley & Sons, New York (1975)
8. Pang, M.-Y., Dai, W., Wu, G., Zhang, F.: On Volume Distribution Features Based 3D Model Retrieval. In: Pan, Z., Cheok, D.A.D., Haller, M., Lau, R., Saito, H., Liang, R. (eds.) ICAT 2006. LNCS, vol. 4282, pp. 928–937. Springer, Heidelberg (2006)
9. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, New York (1982)
10. Serra, J.: Introduction to mathematical morphology. Comput. Vision, Graphics and Image Processing 35(3), 283–305 (1986)
11. Shih, F.: Object representation and recognition using mathematical morphology model. Journal of Systems Integration 1, 235–256 (1991)

12. Soille, P., Rivest, J.-F.: On the validity of fractal dimension measurements in image analysis. Journal of Visual Communication and Image Representation 7(3), 217–229 (1996)
13. Tangelder, J.W., Veltkamp, R.C.: A survey of content based 3D shape retrieval methods. Multimedia Tools Appl. 39(3), 441–471 (2008)
14. Vranic, D.V.: 3D Model Retrieval. PhD thesis, University of Leipzig (2004)
15. Zaharescu, A., Boyer, E., Horaud, R.: TransforMesh: A topology-adaptive mesh-based approach to surface evolution. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 166–175. Springer, Heidelberg (2007)

Trajectory Analysis Using Switched Motion Fields: A Parametric Approach*

Jorge S. Marques[1,2], João M. Lemos[1,3], Mário A.T. Figueiredo[1,4], Jacinto C. Nascimento[1,2], and Miguel Barão[3,5]

[1] Instituto Superior Técnico, Portugal
[2] Instituto de Sistemas e Robótica, Portugal
[3] INESC-ID, Portugal
[4] Instituto de Telecomunicações, Portugal
[5] Universidade de Évora, Portugal

Abstract. This paper presents a new model for trajectories in video sequences using mixtures of motion fields. Each field is described by a simple parametric model with only a few parameters. We show that, despite the simplicity of the motion fields, the overall model is able to generate complex trajectories occuring in video analysis.

1 Introduction

The analysis of trajectories plays an important role in computer vision [1]-[8]. Consider, *e.g.*, a video surveillance system, tracking moving people or vehicles in a parking lot or in a street. The trajectory of each object is a rich source of information about its behavior. We should therefore be able to learn what are the typical trajectories and how can they be characterized so that we can distinguish typical behaviors from abnormal ones and discriminate different types of common behaviors.

A trajectory model must be rich enough to allow different types of behaviors occurring at the same place. For example, several types of trajectories may occur in a hotel lobby. The same happens if we wish to characterize the traffic in a city or in part of it. A single motion field is not enough to characterize people of vehicle motion in a scene.

A generative model for trajectory analysis based on *switched motion fields* was recently proposed in [1]. The model is equipped with estimation methods that are able to learn a set of motion fields describing typical behaviors of objects in a scene. It is assumed that each trajectory is driven by one of the motion fields at each instant of time, the so-called *active field*. Switchings between active fields are allowed and may occur at any position and any instant of time, according to suitable probabilities. This model is rich enough to describe a variety of behaviors and simple enough to be efficiently learned from experimental data.

* This work was supported by FCT (plurianual funding) through the PIDDAC Program funds and by project PTDC/EEACRO/098550/2008.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 420–427, 2011.

The work presented in [1] adopts a non-parametric model for the motion fields based on first order splines. In this paper, we extend that work to simple parametric models that depend on a small number of parameters and discuss if it is still possible to obtain flexible overall behaviors and efficient estimation from observed data. The main contribution consists in a parametric approach to the switched motion field model.

2 Switched Motion Field Model

Let $x = (\mathbf{x}_1, \ldots, \mathbf{x}_L)$, $\mathbf{x}_t \in \mathbb{R}^2$, denote the trajectory of an object in the image. We assume that x is generated by a bank of K vector fields $\mathbf{T}_i : \mathbb{R}^2 \to \mathbb{R}^2$, $i \in \{1, \ldots, K\}$, according to

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{T}_{k_t}(\mathbf{x}_{t-1}) + \mathbf{w}_t \ , \tag{1}$$

where k_t is the label of the active field at instant of time t and $\mathbf{w}_1, \ldots, \mathbf{w}_L$ is a white random sequence with normal distribution $\mathbf{w}_t \sim \mathcal{N}(0, \sigma^2_{k_t}\mathbf{I})$.

Furthermore, we assume that label sequence $k = (k_1, \ldots, k_L)$ is a Markov chain with space varying transition probabilities, *i.e.*, the next active field depends on the current active field as well as on the position of the object in the scene. This is an important issue. For example, consider a cross between two streets. Many pedestrians and vehicles change their direction and velocity at the cross. Therefore, the transition probabilities should be higher at the cross than elsewhere.

To be specific, model switching is characterized by the transition probabilities,

$$P(k_t = j | k_{t-1} = i, \mathbf{x}_{t-1}) = B_{ij}(\mathbf{x}_{t-1}) \ , \tag{2}$$

where $B_{ij}(\mathbf{x})$ is the probability of switching from the field i to the field j at position $\mathbf{x}$. Therefore, $\mathbf{B}(\mathbf{x}) = \{B_{ij}(\mathbf{x})\}$ is a field of stochastic matrices which verify the following properties at each position $\mathbf{x}$:

$$B_{ij}(\mathbf{x}) \geq 0, \quad \sum_{p=1}^{K} B_{ip}(\mathbf{x}) = 1, \quad \forall i, j \ . \tag{3}$$

Figure 1 shows a set of of trajectories which can be easily generated by this model using three vector fields. This figure suggests the variety of behaviors that can be generated with the proposed approach.

The pair $(\mathbf{x}_t, k_t)$ can be considered as a hybrid state since it summarizes all the past information needed to generate the future samples of the process. The joint probability function associated to the pair of sequences $\{x, k\}$ is given by

$$p(x, k) = p(\mathbf{x}_1, k_1) \prod_{t=2}^{L} p(\mathbf{x}_t | k_t, \mathbf{x}_{t-1}) p(k_t | \mathbf{x}_{t-1}, k_{t-1}), \tag{4}$$

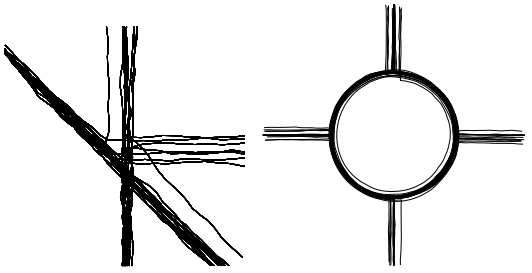


Fig. 1. Synthetic trajectories for the cross and roundabout problems

where

$$p(\mathbf{x}_t|k_t, \mathbf{x}_{t-1}) = \frac{1}{2\pi\sigma_{k_t}^2} e^{-\frac{1}{2\sigma_{k_t}^2}\|\mathbf{x}_t - \mathbf{x}_{t-1} - T_{k_t}(\mathbf{x}_{t-1})\|^2}, \tag{5}$$

and $p(k_t|\mathbf{x}_{t-1}, k_{t-1}) = B_{k_{t-1},k_t}(\mathbf{x}_{t-1})$.

The parameters to be learned from the video data are: i) the number of models K; ii) the motion fields $\mathbf{T}_1, \ldots, \mathbf{T}_K$; iii) the field of transition matrices $\mathbf{B}$; and iv) the noise variances $\sigma_1^2, \ldots, \sigma_K^2$.

In [1] the fields $\mathbf{T}_1, ..., \mathbf{T}_K$ and $\mathbf{B}$ are modeled in a non-parametric way. They are specified at the nodes of a regular grid and interpolated using first order splines

$$\mathbf{T}_k(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{t}_k^i \phi_i(\mathbf{x}), \qquad \mathbf{B}(\mathbf{x}) = \sum_{i=1}^{N} \mathbf{b}^i \phi_i(\mathbf{x}) \,, \tag{6}$$

where $\mathbf{t}_k^i, \mathbf{b}^i$ are the velocity vector and the transition matrix associated to the $i-th$ node of the grid and $\phi_i(\mathbf{x})$ is the corresponding spline function, centered at the i-th node. As a consequence, that approach can be classified as non-parametric since we are not imposing any kind of structure, and each field depends on a large number of parameters (typically a few hundreds) which have to be estimated from the data. Some kind of regularization (Gaussian field priors, in [1]) is required to obtain meaningful estimates for these parameters.

In this paper we follow a different approach by adopting parametric models for the motion fields. This results in a much smaller number of parameters to be estimated. Although we are making strong assumptions about each motion field, a flexible trajectory model is expected at the end because trajectories are decomposed into segments, each of them generated by a different motion field. Parametric field may be tuned to a specific space region, if necessary.

3 Parametric Motion Fields

We consider several parametric motion models, which are often used in image alignment and registration, namely [9]: translation (T), Euclidean (E), similarity (S) and affine (A) transforms. All these models are expressed by

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{t} \,, \tag{7}$$

Table 1. Parametric motion models

Name	Transformation	Motion field
T	$\mathbf{z} = \mathbf{x} + \mathbf{t}$	$\mathbf{T}(\mathbf{x}) = \mathbf{t}$
E	$\mathbf{z} = \mathbf{R}\mathbf{x} + \mathbf{t}$	$\mathbf{T}(\mathbf{x}) = (\mathbf{R} - \mathbf{I})\mathbf{x} + \mathbf{t}$
S	$\mathbf{z} = s\mathbf{R}\mathbf{x} + \mathbf{t}$	$\mathbf{T}(\mathbf{x}) = (s\mathbf{R} - \mathbf{I})\mathbf{x} + \mathbf{t}$
A	$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{t}$	$\mathbf{T}(\mathbf{x}) = (\mathbf{A} - \mathbf{I})\mathbf{x} + \mathbf{t}$

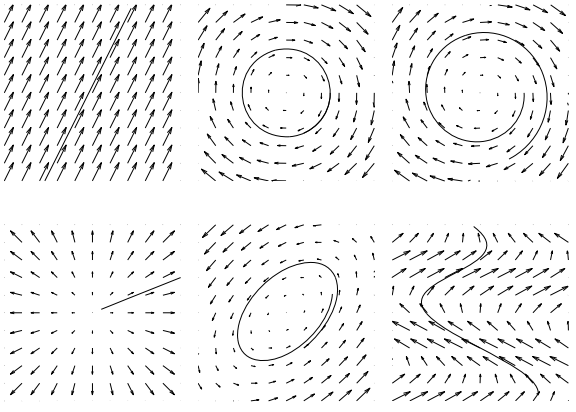


Fig. 2. Examples of motion fields and trajectories generated by the T, E, S models (top row) and A, A, non parametric models (bottom row)

where $\mathbf{z}$ is the transformed position of the point $\mathbf{x}$, $\mathbf{A}$ is a 2×2 matrix and $\mathbf{t}$ is a 2×1 translation vector. The only difference between these models lies in the structure of matrix $\mathbf{A}$ as shown in Table 1. In this table, $\mathbf{R}$ (a rotation matrix) and $s\mathbf{R}, s \in \mathbb{R}$, have the following structure

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad s\mathbf{R} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}, \tag{8}$$

and $\mathbf{A}$ is an arbitrary 2×2 matrix.

Figure 2 shows examples of the motion fields and trajectories generated by these models and by a non-parametric one. Only the translation and the Euclidean transform generate trajectories in which the object moves with constant speed since the eigenvalues of the matrix $\mathbf{A}$ lie on the unit circle. In the other cases, velocity has an exponential growth or decay. This is not a major problem since each model is only used for a short period of time. We stress that, at this point, no switching was allowed in the generation of these trajectories. More complex trajectories can be generated if we allow model switching.

4 Model Estimation

In an ideal setting, we would like to turn on the camera and ask the system to learn the behavior of all the pedestrians and vehicles in the scene. Assuming

that the tracking task is solved (although this is by no means a trivial task) we would like to estimate the number of fields and the field parameters from a set of observed trajectories $\mathcal{X} = \{x^{(1)}, \ldots, x^{(S)}\}$ where $x^{(s)} = (\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{L_s}^{(s)})$ is the s-th trajectory.

The maximum likelihood (ML) estimates of all the model parameters, collectively denoted as θ, can be obtained by solving the following optimization problem

$$\hat{\theta} = \arg\max_{\theta} \ \log p(\mathcal{X}|\theta). \tag{9}$$

However, the likelihood function cannot be directly computed. Since we do not know the sequence of active models underlying each trajectory, we should marginalize the complete likelihood function $p(\mathcal{X}, \mathcal{K}|\theta)$, *i.e.*,

$$p(\mathcal{X}|\theta) = \sum_{\mathcal{K}} p(\mathcal{X}, \mathcal{K}|\theta) = \sum_{\mathcal{K}} \prod_{s=1}^{S} p(x^{(s)}, k^{(s)}|\theta), \tag{10}$$

where $\mathcal{K} = \{k^{(1)}, \ldots, k^{(1)}\}$ are the (unobserved) label sequences (active models) and $p(x^{(s)}, k^{(s)}|\theta)$ is the joint density defined in (4). The marginalization involves a sum for all sequences of labels $\mathcal{K}$ which is unfeasible since it involves a huge number of operations.

This difficulty can be circumvented by using the Expectation-Maximization (EM) method. The EM method is based on the optimization of an auxiliary function: the conditional expectation of the complete log-likelihood

$$U(\theta, \hat{\theta}) = \mathbb{E}\left\{\log p(\mathcal{X}, \mathcal{K})|\mathcal{X}, \hat{\theta}\right\}, \tag{11}$$

where $\hat{\theta}$ is the currently available estimate of the model parameters. The EM method generates a sequence of estimates by iteratively optimizing $U(\theta, \hat{\theta})$ with respect to θ, to update the parameter estimates:

$$\hat{\theta}(t+1) = \arg\max_{\theta} \ U(\theta, \hat{\theta}(t)). \tag{12}$$

The expected value of the complete log-likelihood with respect to these variables can be written as

$$U(\theta, \hat{\theta}) = \bar{\mathcal{A}}(\mathcal{X}, \mathcal{K}) + \bar{\mathcal{B}}(\mathcal{X}, \mathcal{K}), \tag{13}$$

with

$$\begin{aligned} \bar{\mathcal{A}}(\mathcal{X}, \mathcal{K}) = C - \textstyle\sum_{s=1}^{S} \sum_{t=2}^{L_s} \sum_{i=1}^{K} w_i^{(s)} \Big[\log(2\pi\sigma_i^2) \\ + \tfrac{1}{2\sigma_i^2} \|\mathbf{x}_t^{(s)} - \mathbf{x}_{t-1}^{(s)} - \mathbf{T}_i(\mathbf{x}_{t-1}^{(s)})\|^2 \Big], \end{aligned} \tag{14}$$

$$\bar{\mathcal{B}}(\mathcal{X}, \mathcal{K}) = \sum_{s=1}^{S} \sum_{t=2}^{L_s} \sum_{i,j=1}^{K} w_{i,j}^{(s)} \log B_{ij}(\mathbf{x}_{t-1}^{(s)}), \tag{15}$$

where $w_i^{(s)}(t) = P(k_t^{(s)} = i|\mathbf{x}^{(s)}, \hat{\theta})$ is the probability of label i at time t and $w_{i,j}^{(s)}(t) = P(k_{t-1}^{(s)} = i, k_t^{(s)} = j|\mathbf{x}^{(s)}, \hat{\theta})$ the probability of the pair of labels i, j at

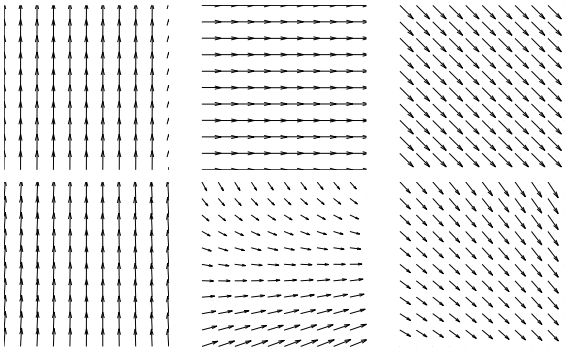


Fig. 3. Estimated fields for the cross problem with translation (1st row) and affine (2nd row) models ($K = 3$)

consecutive instants of time. The probabilities $w_i^{(s)}(t), w_{i,j}^{(s)}(t)$ are computed in the E-step of the EM method using the forward-backward algorithm [10].

The M-step maximizes U with respect to the model parameters. The maximization with respect to the noise variances and switching matrix field is done as in [11]. The optimization with respect to the motion parameters depends on the motion model adopted but this is straight forward.

5 Results

The proposed model was applied to synthetic trajectories. Figure 1 shows two sets of 30 trajectories which are denoted as cross and roundabout. The first case simulates a cross between three streets with two entries. The second case simulates a roundabout with four entries and combines linear and circular segments. The field of transition matrices was modeled in a non-parametric way, using a 11×11 regular grid of points and first order interpolation splines as in [1].

Figure 3 shows the estimates obtained by the EM method assuming affine motion fields and translation fields. These results were obtained after 5 iterations. Both models are able to correctly extract the correct motion fields. The second affine motion field is not uniform but it is approximately uniform in the region of interest. We stress that we do not have any *a priori* knowledge about the active model at each instant of time. This information must be guessed by the EM algorithm using the soft assignment variables $(w_i^{(s)}(t), w_{i,j}^{(s)}(t))$.

The second problem is more complex since it involves a larger number of fields (five) and a mixture of circular and uniform fields. Figure 4 (1st row) shows the estimates obtained using Euclidean motion fields showing that the EM method is able to correctly estimate the field parameters. It should be mentioned that the output of the EM method depends on the initialization as shown in the Figure 4 (bottom): two motion fields are associated to the circular motion and the third motion field in the Figure tries to represent two different motion directions.

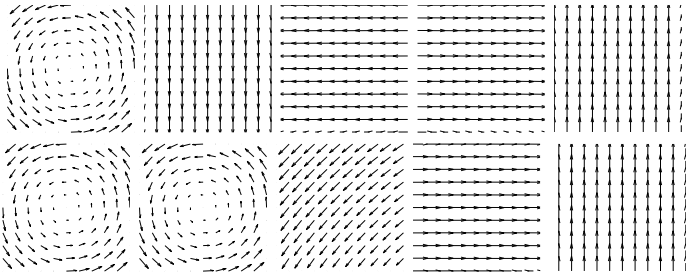


Fig. 4. Estimated fields for the roundabout problem with Euclidean transform model ($K = 5$) and two different initializations

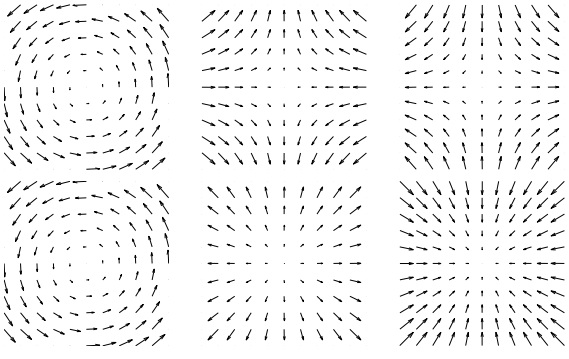


Fig. 5. Estimated fields for the roundabout problem with the affine model ($K = 3$) and two different initializations

This is shown in Figure 5 which displays the output of the EM method, assuming an affine model with a smaller number of fields (three) and two different initializations.

Although excellent results were achieved for both problems, the algorithm may converge to local maxima of the likelihood function, leading to poor estimates of the motion fields. This is a consequence of the EM estimation method which does not guarantee the convergence towards the global maxima [12]. However, this effect is stronger in the estimation of parametric vector fields depending on a small number of parameters (< 10) then in the case of non-parametric models which depend on hundreds of parameters. The presence of global restrictions increase the attraction towards local maxima.

6 Conclusions

This paper presents an extension of the trajectory model with multiple motion fields presented in [1]. It shows that complex trajectories can be generated using a set of simple motion fields (parametric fields) depending on a small number of

parameters. The key point is the ability to switch between motion fields at any position in space. In addition we use space dependent switching probabilities which allow different switching behaviors in different regions of space.

Despite the good results obtained there are several open question to be addressed in the future: how to initialize the EM method in an efficient way? how to determine the best number of motion fields for a given problem? how to select the most appropriate model for each field? application of this model to real data and the comparison with the non-parametric model [1]. These questions will be addressed in a forthcoming paper which will include extensive results and a comparison with non-parametric techniques.

References

1. Nascimento, J.C., Figueiredo, M.A.T., Marques, J.S.: Trajectory Analysis in Natural Images Using Mixtures of Vector Fields. In: IEEE International Conference of Image Processing (2009)
2. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviours. IEEE Trans. on Systems and Cybernetics, Part C 34, 334–352 (2004)
3. Duong, T., Bui, H., Phung, D., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
4. Fu, Z., Hu, W., Tan, T.: Similarity based vehicle trajectory clustering and anomaly detection. In: IEEE International Conference on Image Processing, Genoa, Italy (2005)
5. Wang, X., Ma, K., Ng, G., Grimson, E.: Trajectory analysis and semantic region modeling using a nonparametric Bayesian model. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
6. Oliver, N., Rosario, B., Pentland, A.: A Bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 831–843 (2000)
7. Junejo, I., Javed, O., Shah, M.: Multi feature path modeling for video surveillance. In: International Conference on Pattern Recognition (2004)
8. to be included
9. Szeliski, R.: Image Alignment and Stiching: a Tutorial. Foundations and Trends in Computer Graphics and Computer Vision 2, 1–104 (2006)
10. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
11. Barão, M., Marques, J.S., Lemos, J.M.: An Improved EM-method for the Estimation of Transition Probabilities in Multiple Model Switching Systems. In: IFAC Symposium on Nonlinear Control Systems, Bolonha (2010)
12. Wu, C.: On the convergence properties of the EM algorithm. The Annals of Statistics 11, 95–103 (1983)

Semi-supervised Probabilistic Relaxation for Image Segmentation*

Adolfo Martínez-Usó, Filiberto Pla, José M. Sotoca, and Henry Anaya-Sánchez

Institute of New Imaging Technologies - Dept. of Computer Languages and Systems
Universitat Jaume I, 12071 Castellón, Spain

Abstract. In this paper, a semi-supervised approach based on probabilistic relaxation theory is presented. Focused on image segmentation, the presented technique combines two desirable properties; a very small number of labelled samples is needed and the assignment of labels is consistently performed according to our contextual information constraints. Our proposal has been tested on medical images from a dermatology application with quite promising preliminary results. Not only the unsupervised accuracies have been improved as expected but similar accuracies to other semi-supervised approach have been obtained using a considerably reduced number of labelled samples. Results have been also compared with other powerful and well-known unsupervised image segmentation techniques, improving significantly their results.

Keywords: Semi-supervised, Image segmentation, Probabilistic Relaxation.

1 Introduction

Relaxation methods find numerical solutions for a wide range of problems in physics and engineering and, more concretely, probabilistic relaxation has demonstrated to be very useful for pattern recognition [13]. A general framework for the theoretical foundations of relaxation processes can be found in [9,6]. This general relaxation structure has attracted important interest, being often refined in a number of ways by means of *ad hoc* or heuristic choices [12].

Relaxation approaches are *iterative* processes that are used for reducing ambiguities in assigning symbolic labels to a set of nodes (clusters, objects, etc.) which is often known as equilibrating or *relaxing* a system. Relaxation methods involves *contextual information* that describes relations between single components [6], defining a neighbourhood in accordance with the properties of the system. The contextual information is generally introduced into the process from our *a priori* knowledge of the problem. Therefore, these approaches present two interesting features; the use of the context of the problem and the expected good performance to obtain a robust solution [8].

Medical imaging problems are generally too specialised to use unsupervised techniques but, at the same time, too time-consuming for an expert if every single case

* This work was supported by the Spanish Ministry of Science and Innovation under the projects Consolider Ingenio 2010 CSD2007-00018, AYA2008-05965-C04-04/ESP, TIN2009-14103-C03-01 and by Caixa-Castelló foundation under the projects P1 1B2009-45.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 428–435, 2011.

should be solved manually, spending a large amount of time isolating the most interesting parts of the images. Therefore, a process that is able to do this work in an accurate way and, at the same time, does not need too much participation of an expert, has become a very demanding task on this field.

Semi-supervised learning has received an increasing attention for the last years and has been widely extended to many fields [2], being specially useful for medical imaging applications. Semi-supervised approaches arise from the idea of using together a large amount of unlabelled data, which is often cheap and easy, and few labelled data, which is hard to obtain since it requires human experts or special devices. The important point here is to manage a better classifier (or clustering result) than from the unlabelled data alone [14].

A semi-supervised approach for colour image segmentation based on probabilistic relaxation is presented in this paper. The main contributions of this work are:

- We propose a probabilistic relaxation method that using few labelled samples introduced by an expert (contextual information) is able to propagate this information to the whole system.
- We also offer preliminary experimental evidences that our method improves the results obtained by other unsupervised techniques of the literature. The results of a semi-supervised approach based on the Expectation-Maximisation algorithm have been also improved in terms of the number of labelled samples needed.
- The usefulness of this technique will be shown for a particular application in Dermatology.

2 Probabilistic Relaxation Methodology

A probabilistic relaxation method is an iterative process that assigns consistent labels to a initial set of nodes on the basis of the contextual information; which is also introduced into the model. A node is a point in a graph that represents objects, clusters, regions, etc. whereas the contextual information is generally related to the relationship among those nodes, that is, arcs among the nodes in the graph.

Let us suppose a set of nodes $\mathcal{N} = \{n_1, n_2, \ldots, n_N\}$ and a set of class labels $\mathcal{L} = \{\omega_1, \omega_2, \ldots, \omega_L\}$. Suppose a *support* function $Q^s(n_i \leftarrow \omega_k)$ representing that the node n_i would be labelled with ω_k. This support function results from each binary relation with the set $\overline{\mathcal{N}}_i$ neighbouring nodes of n_i at the step s^{th} of the iterative process,

$$Q^{s+1}(n_i \leftarrow \omega_k) = Q^s(n_i \leftarrow \omega_k) + \frac{1}{\left|\overline{\mathcal{N}}_i\right|} \sum_{j \in \overline{\mathcal{N}}_i} \mathcal{C}_{ij} \;\; P^s(n_j \leftarrow \omega_k) \tag{1}$$

where $\mathcal{C}_{ij}$ are the coefficients representing the *strength of interaction* between nodes n_i and n_j. These coefficients are independent of the estimated posterior probabilities (P) and can be computed ahead of time, remaining constant during the entire process. Coefficients $\mathcal{C}_{ij}$ satisfy that $\sum_{j=1}^{N} \mathcal{C}_{ij} = 1$. Our approach defines these coefficients as:

$$\mathcal{C}_{ij} = \frac{\mathcal{D}(i,j)}{\sum_l \mathcal{D}(i,l)} \tag{2}$$

being

$$\mathcal{D}(i,j) = exp\left(\frac{c_{ij}}{gmin}\right) \qquad gmin = min\left(c_{mn}\right) \;\; \forall m,n \in \mathcal{N} \tag{3}$$

Coefficients $c_{ij} = \frac{\alpha_j}{d(i,j)}$ represent the relationship between the density of the modes and the distance among them. Thus, α_j is the prior probability for node n_j whereas $d(i,j)$ is the distance between the nodes n_i and n_j. Note likewise that $\mathcal{D}(i,j)$ is a potential term that acts as a relative measure of potential similarity function. It will be high for the neighbouring node with the best rate for coefficient c_{ij} and very low for the rest of the nodes.

The updating formula for calculating the posterior probability $P^s(n_i \leftarrow \omega_k)$ that node n_i would be labelled with ω_k is:

$$P^{s+1}(n_i \leftarrow \omega_k) = \frac{P^s(n_i \leftarrow \omega_k)Q^s(n_i \leftarrow \omega_k)}{K} \tag{4}$$

where $K = \sum_{l=1}^{|\mathcal{L}|} P^s(n_i \leftarrow \omega_l)Q^s(n_i \leftarrow \omega_l)$ is a normalising constant.

The system (P^0) is initialised on the basis of the *a priori* information of the problem statement, also setting $Q^0 = P^0$ in this initialisation. These initial probabilities act as the *compatibility coefficients* among the nodes.

2.1 The Algorithm

The algorithm here presented for a semi-supervised probabilistic relaxation (**semi-PR**) has three input sources: the number of classes, the labelled and unlabelled data set, dividing this last set into two subsets for training and test. In addition, the algorithm will consider that each class can have multiple initial probability distributions (modes). The number of modes of each class is estimated in the initialisation stage, being each initial mode a node in the **semi-PR**.

Initialisation stage: The same initialisation stage presented in [10] is used in our approach for estimating the number of Gaussian modes per class. Thus, the training data is divided into N modes where the mean and covariance of each mode is estimated. Figure 1 shows several examples of this initialisation stage in its central row.

Semi-supervised Probabilistic Relaxation: The Gaussian modes found in the initialisation stage are used as initial nodes for our proposed semi-supervised algorithm. The initial probabilities (P^0) for each node are based on the *a priori* information that we have from our problem, that is, from initial modes found in the initialisation stage and from the labelled samples. The labelled samples allow us to label some of the initial nodes/modes to certain classes. The probabilities of the modes where no *a priori* information is available are equally distributed for each class before starting the relaxation process.

The objective of the **semi-PR** algorithm is to iteratively optimise Equation (4) until all the initial nodes would be consistently labelled for each class. For this optimisation, Equation (1) must be also calculated in each iteration, providing the *level of agreement* for each node and their possible labels. Distance $d(i,j)$ in coefficients c_{ij} is worked out using the Mahalanobis distance between the nodes n_i and n_j.

The process stops when, for each node of the system, one of the class probabilities exceeds $1 - \epsilon_1$, being $\epsilon_1 \ll 1$. As the literature suggest [3], if the whole changes in the system do not exceed certain threshold ϵ_2, the system is supposed relaxed enough and the process also stops in this case. Note that the process needs to keep iterating until the previous stopping conditions are reached, even so, from the first iterations the propagation of information is channelled through the probabilistic relaxation model. Therefore, parameters ϵ_1 and ϵ_2 are not critical, although they can probably be considered application-dependent.

3 Experimental Results

3.1 Results on $2D$-Datasets and Two Classes

Figure 1 shows the results obtained on synthetic $2D$-datasets with two classes. These toy datasets are frequently used in the literature as illustrative examples to show the robustness of a technique. Datasets 1, 2 and 4 (columns 1, 2 and 4 of the figure) show very consistent results with a classification accuracy higher than 98% in all the cases. The approximate area position of the labelled samples used in each case have been also drawn in the resulting labelled dataset (light green in third row). From left to right, each dataset has needed 1, 3, 2 and 1 labelled samples per class respectivelly.

It is worth noticing how the relaxation process finds a slightly wrong solution for the dataset shown in the third column. The initialisation stage sometimes generates wrong modes that should probably have been divided into several ones. When this happens (p.e. the mode in green colour) the **semi-PR** method still assigns that mode to the most likely class, although the final results will certainly be partially mistaking. On the contrary, the second dataset shows how the algorithm finds the correct solution despite that there exist two wrong initialised modes (the ones in pink colour). In this last case, the algorithm has several alternatives for propagating the labels through the system, obtaining eventually the most suitable solution.

3.2 Results on Colour Image Segmentation

The main part of the experimental work is focused on segmenting colour images. From each image, the number of classes is known and a small number of labelled pixels is available. Each sample is a $2D$-vector representing the chroma of a pixel in the $L^*a^*b^*$ colour space, that is, the a^*b^* dimensions. The total amount of samples is equally divided into two sets, one for training and another for test. The supervised data take samples proportionally from each class, representing the 0.1% of training set.

Preliminary results are provided for images from a skin diagnosis application, where several skin pathologies have to be analysed. The proposed technique has been used to automatically isolate skin regions, in order to segment unhealthy areas out of healthy skin in a two-class problem. Currently, this unhealthy skin is drawn manually, which is time-consuming as well as prone to errors due to manual mistakes and tiredness. Hence, adding very few labelled data to the clustering process by means of an expert, for instance, just clicking on few image points, our semi-supervised proposal can improve this task in a quick and practical way, introducing some interactivity with the user.

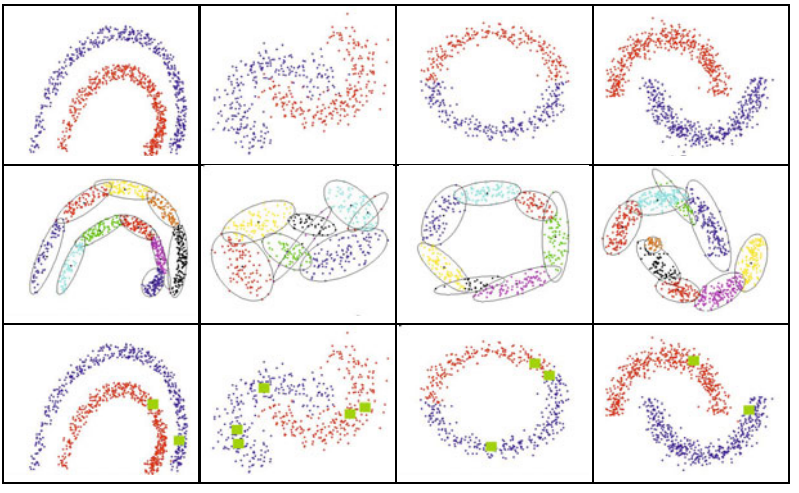


Fig. 1. Semi-supervised results of two-class datasets. From top to bottom in rows, ground-truth, initial modes and the proposed probabilistic relaxation result, providing initially only between 1 and 3 labelled samples per class (light green spots in third row).

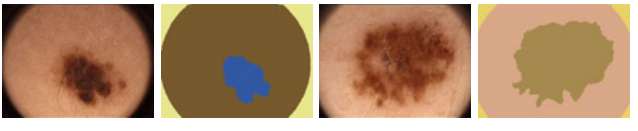


Fig. 2. Application in Dermatology. From left to right, *p02* image and its ground truth, *p34* image and its ground truth.

The classification accuracy of the proposed algorithm has been compared with the standard unsupervised EM algorithm for Gaussian mixtures [1] and with the semi-supervised approach based on the EM algorithm (**semi-EM**) presented in [10]. The ground-truth references (manually labelled) were available for the images shown in Figure 2 and, therefore, the classification rates for the example images *p02* and *p34* can be worked out. Table 1 shows the percentages with regard to the number of well-classified pixels when the clustering results are compared with the ground-truth images. Notice that the classification rates obtained on these images are quite similar for the proposed algorithm and the **semi-EM**. Nevertheless, the important point on this comparison is the number of labels used in each case. As authors specify in [10], the **semi-EM** uses the 1% of the total image samples for each class whereas the **semi-PR** only needs the 0.1% of the training set.

Figure 3 shows the image segmentation results for the example images of the Figure 2. The proposed technique is compared with other three well-known unsupervised image segmentation algorithms and the **semi-EM** algorithm. Firstly, **MS** is an effective algorithm proposed by Comaniciu and Meer [4] based on the *mean shift* algorithm. Secondly, **SRM** algorithm, proposed by Nock and Nielsen [11], is based on the idea of using perceptual grouping and region merging for image segmentation. Thirdly, **FH**

Table 1. Classification rates for the images of the dermatology application

	semi-PR	**semi-EM**	Unsup-EM
p02	**95.48%**	95.45%	78.76%
p34	90.24%	**91.87%**	90.94%

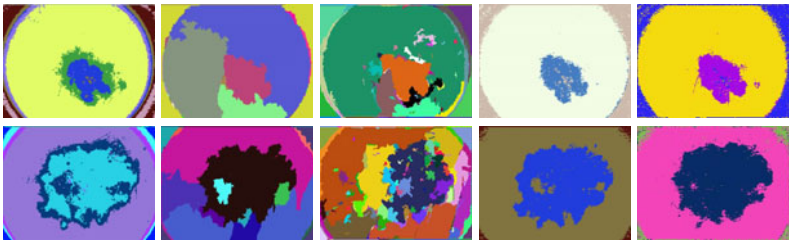


Fig. 3. Application in Dermatology. Top row show the results on *p02* image whereas the bottom row shows the results for *p34* image. In columns and from left to right, five image segmentation results corresponding to **MS**, **SRM**, **FH**, **semi-EM** and the proposed **semi-PR**.

algorithm proposed by Felzenszwalb and Huttenlocher [7] adaptively adjusts the segmentation criterion based on the degree of variability in neighbouring regions. Finally, the **semi-EM** algorithm presented in [10] has been used for image segmentation in the same application in dermatology.

Taking into account the resulting images from the segmentation, the results presented in Figure 3 show that,

1. The proposed semi-supervised method is significantly better than the **MS**, **SRM** and **FH** algorithms.
2. The results of the **semi-EM** algorithm are comparable to the proposed technique. Healthy and unhealthy skin have been well separated, allowing an easy and precise identification of each part. However, as it has been said, the **semi-PR** result has used a significant smaller amount of labelled samples for solving the same problem.

Note that there are some isolated pixels that are badly labelled. This is due to the fact that there is no spatial constraint in the process, but some spatial regularisation could be applied as in [5].

Summarising this experimental section, we have focused our research on an application in dermatology. From a practical point of view, the unsupervised approaches are not robust enough for this kind of application and their results have been shown as a matter of comparing the different methodologies. Regarding the semi-supervised approaches, the **semi-EM** algorithm would need an expert to select sets of pixels from each class since such amount of points cannot be provided with a few clicks. In the proposed **semi-PR** however, the expert could only select a few interesting points from the image, as it would be desirable.

It is also important to compare the performance of the two semi-supervised approaches in terms of the classification accuracy for different number of labelled samples. Thus, Figure 4 shows the classification rate (y-axe) for each semi-supervised

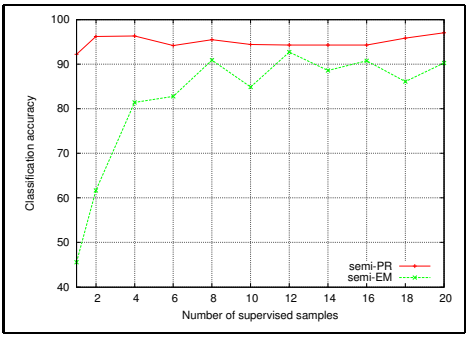


Fig. 4. Performance classification accuracy for **semi-EM** algorithm and proposed **semi-PR**

method related to the number of labelled samples provided (x-axe). This graph has been obtained using the second dataset (2nd column) shown in Figure 1, which is probably the most difficult one of the $2D$-datasets. The labelled samples have been randomly selected and each method has been carried out 10 times in order to reduce the influence of the stochasticity in the experiment. Graph in Figure 4 shows the average classification for each method. As it can be seen, **semi-PR** (red line) reaches the maximum classification rates quickly from the beginning, using a minimum number of labelled samples. The **semi-EM** algorithm (green line) is not robust enough when very few labelled samples are provided and its performance is always worse than our proposal.

4 Conclusions

The main objective of this work is to develop a robust semi-supervised algorithm using probabilistic relaxation for colour image segmentation. Our methodology satisfies the use of few labelled samples and the assignment of labels according to our contextual constraints. The proposed approach has improved the image segmentation results of other semi-supervised and unsupervised approaches.

This work supports the increasing attention that the semi-supervised learning is receiving during the last years. As an image segmentation problem, the methodology has demonstrated an improvement with respect to other well-known unsupervised segmentation algorithms. As a classification problem, the accuracy results obtained have been comparable to a recent semi-supervised approach based on the EM algorithm, being capable to reach the highest classification rates with very few labelled samples.

From a practical point of view, our proposal is more suitable for partial annotation in interactive image segmentation in applications like dermatology, since it is much better in terms of the amount of supervised information needed to finish the task.

In order to extend these promising results, further work includes testing the algorithm in other problems *i)* with multiple classes and higher dimensionality, *ii)* in presence of noise and *iii)* with a higher level of overlapping among the classes. Future improvements are also directed to introduce a procedure to give more or less relevance to the supervised data, expressing somehow the level of confidence in the contextual information.

References

1. Bouman, C.A.: Cluster: An unsupervised algorithm for modeling Gaussian mixtures (April 1997), http://www.ece.purdue.edu/~bouman
2. Chapelle, O., Scholkopf, B., Zien, A. (eds.): Semi-supervised Learning. MIT Press, Cambridge (2006)
3. Christmas, W.J.: Structural matching in computer vision using probabilistic reasoning. PhD thesis, CVSSP, University of Surrey (1995)
4. Comaniciu, D., Meer, P.: Robust analysis of feature spaces: Color image segmentation. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 750–755 (1997)
5. Diplaros, A., Vlassis, N., Gevers, T.: A spatially constrained generative model and an EM algorithm for image segmentation. IEEE Trans. on Neural Networks 18(3), 798–808 (2007)
6. Faber, P.: A theoretical framework for relaxation processes in pattern recognition: Application to robust nonparametric contour generalization. IEEE Trans. on PAMI 25(8), 1021–1027 (2003)
7. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
8. Haralick, R.M.: Decision making in context. IEEE Trans. on PAMI 5(4), 417–428 (1983)
9. Hummel, R.A., Zucker, S.W.: On the foundations of relaxation labeling processes. IEEE Trans. on PAMI 5(3), 267–287 (1983)
10. Martinez-Uso, A., Pla, F., Sotoca, J.M.: A semi-supervised gaussian mixture model for image segmentation. In: ICPR, pp. 2941–2944 (2010)
11. Nock, R., Nielsen, F.: Statistical region merging. IEEE Trans. on PAMI 26(11), 1452–1458 (2004)
12. Wang, H.F., Hancock, E.R.: Probabilistic relaxation using the heat equation. In: ICPR, vol. 2, pp. 666–669 (2006)
13. Kittler, J., Christmas, W.J., Petrou, M.: Structural matching in computer vision using probabilistic relaxation. IEEE Trans. on PAMI 17(8), 749–764 (1995)
14. Zhu, X.: Semi-supervised learning literature survey. Technical Survey 1530, Computer Sciences, University of Wisconsin-Madison (2005)

Poker Vision: Playing Cards and Chips Identification Based on Image Processing

Paulo Martins[1], Luís Paulo Reis[2], and Luís Teófilo[2]

[1] DEEC Electrical Engineering Department
[2] LIACC Artificial Intelligence and Computer Science Lab.
Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, 4200 Porto, Portugal
{ee04150,lpreis}@fe.up.pt, luisfgteofilo@gmail.com

Abstract. This paper presents an approach to the identification of playing cards and counting of chips in a poker game environment, using an entry-level webcam and computer vision methodologies. Most of the previous works on playing cards identification rely on optimal camera position and controlled environment. The presented approach is intended to suit a real and uncontrolled environment along with its constraints. The recognition of playing cards lies on template matching, while the counting of chips is based on colour segmentation combined with the Hough Circles Transform. With the proposed approach it is possible to identify the cards and chips in the table correctly. The overall accuracy of the rank identification achieved is around 94%.

Keywords: Poker, playing cards identification, image processing, template matching, colour segmentation, chips counting.

1 Introduction

To build an autonomous agent of Poker it is necessary to consider two very distinct modules: the intelligence of the agent and the agent's interaction with the real world.

There is a lot of research work in Poker agent intelligence. The most renowned work is Darse Billings PhD thesis [1] where were distinguished and discussed several possible poker agent architectures. Building Poker agents requires the use of opponent modelling techniques, because the agent is not aware of the full game state. Several articles were wrote about this matter, such as [2] or [3]. Most opponent modelling techniques are based on David Slansky publications, such as [4].

This work focused on the interaction of the agent with the real world, more specifically collecting information about the state of the poker game. There has been relatively little work on playing cards or chips recognition based on vision [5–8]. The perception layer designed and implemented relies completely on image acquisition and processing. Besides the tasks of recognition of cards and chips, it also features the detection of players in the game at the beginning of each hand and the perception of the dealers position.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 436–443, 2011.

2 Playing Cards Identification

2.1 Find Cards

Finding the cards present on the table relies on the great contrast between the poker table and the cards lying on it, where the former features dark colours, e.g. dark green, and the latter is always white coloured. The greater the contrast between the cards and the poker table, the stronger the edges between both will be, i.e. the border between the card and the table

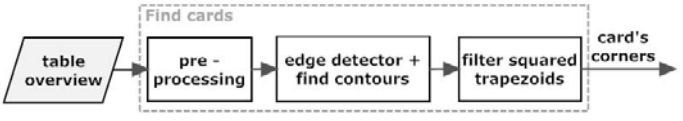


Fig. 1. Detect cards contours algorithm overview

The image captured of the table overview is first submitted to the pre-processing block. This block consists of low level image processing functions, where the image is first converted to gray scale, followed by smoothing with a Gaussian filter and finally the contrast between the card and the poker table is enhanced by a linear stretch to the histogram.

The image is subject of edge detection, more specifically Canny edge detector. The resulting output from the Canny detector is scanned for external contours, thus all the first level contours found are kept while the remaining are discarded. The first level contours correspond to the border between the table and the cards.

Finally, the algorithm approximates all the contours remaining by polygons and selects those which feature four vertices, correspondent to the four vertices of a card, thus remaining only trapezoids. Since playing cards have the shape of a rectangle, i.e. four right angles, the algorithm filters the trapezoids by their inner angles. All the trapezoids which inner angles are close to 90° are selected as cards, while the remaining are discarded. Accepting inner angles within a range around 90° makes the system robust enough to be in any angle relatively to the cards. If all the inner angles had to be exactly 90°, it would be required to place the camera perpendicular to the cards on the table.

2.2 Cards Extraction

The algorithm, shown in Fig. 2, is used to isolate each one of the trapezoid (cards) found so it is possible to process them individually. When the tripod is setup in one of the seats around the table, the position relative to the cards is not calibrated. Therefore the PokerVision does not know in advance in which angle it is relative to the community cards. In order to make this system robust enough to be positioned in any place around the poker table, the position of each card is computed using the coordinates information of its corners. This computation is based on the position of the corners relative to each other and to

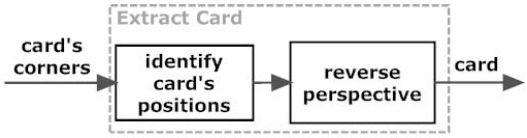


Fig. 2. Cards' extraction algorithm

the global coordinate system. The reverse perspective block is intended to reverse the perspective view of the card, outputting the image of the card with the correct size proportions, as a regular playing card, and without the perspective view effect, . It computes the matrix of perspective transform, based on the starting points, i.e. corners of the card on the original image, and destination points, i.e. the corners of the new rectangle shaped. The former are based on the corners coordinates of each card combined with the cards position/rotation determined previously. Since enlargement occurs, the final image will have some loss of quality.

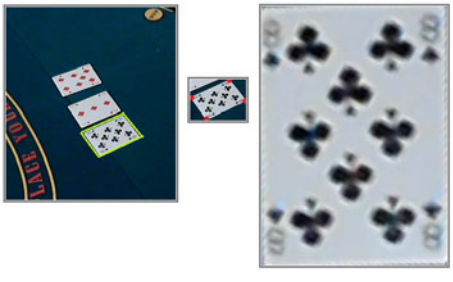


Fig. 3. Extraction of playing card reversing the perspective view

2.3 Rank Extraction

This block aims to extract the rank of the card and prepare it for the template matching. By performing this, the algorithm restricts the region to be subject of recognition, benefiting the rank identification reliability. Not only prevents possible misidentifications with something else drew on the card than the rank itself, but also enables the identification to compute much faster since the image area to be analyzed is reduced significantly.

The position and size of the rank are both the same across cards of the same deck, therefore the size and position of the crop was pre defined for the cards used along the work. After the region of interest is extracted, the algorithm removes unwanted noise by smoothing it with a Gaussian filter. Thereafter binarizes the resultant image with an Adaptive Threshold. The resulting image after binarization is eroded in order to remove some remaining thin noise.

In order to normalize the position of the rank, along with the filtering of remaining noise, the image is submitted to the isolate rank block. This block

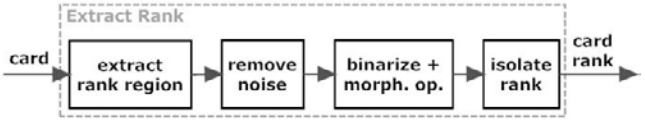


Fig. 4. Rank extraction algorithm overview

was implemented in order to tightly isolate the rank of the card from the rest of the crop. It seeks for all the elements, i.e. contours, present on the binarized image and encloses them in a box. Afterwards, the algorithm analyses the sizes of the enclosed contours as well as its centroids and discards all the contours which feature a non reasonable size, followed by selecting from the remaining, the one which centroid is the closest to the centre of the image. The resulting image is one featuring only the rank tightly isolated, Fig. 5 d).

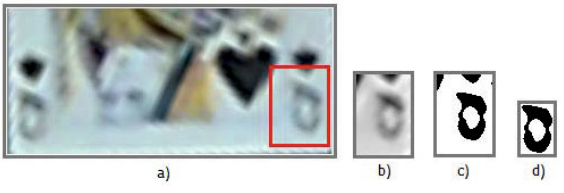


Fig. 5. Part of the rank extraction process

The rank 10 is constituted by two characters instead of one, which means the algorithm when iterating through the conditions referred above, discards one of the characters. The work around for this issue consists in drawing one horizontal line 1pixel thick, wide enough to superimpose the characters 1 and 0. This misleads the algorithm, which searches for contours, to interpret both characters as the same one.

2.4 Rank Identification

The approach used here, template matching [2], is a pattern recognition technique that allows detecting the presence of a specific object in an image. It can produce reliable results but needs to know exactly what to look for since it is based on the comparison of the image against a pre-defined template.

The method is performed by sliding a template, Fig. 6, over the input image (card rank) while it calculates the normalized square differences between both. The best match corresponds to the one closest to 0. A total of thirteen templates were prepared previously and based on the cards used, corresponding each one of them to a rank of a card, Fig. 7. To prevent false matches, it was defined empirically that the scores above 0.45 are not considered as a valid match.

It is worth mentioning that, preceding the template matching, both ranks depicted in the card are extracted and subject of recognition. This provides the Template Matching with more samples to identify, i.e. ranks, extracted from

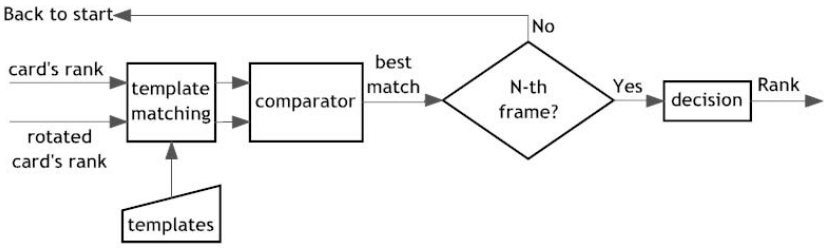


Fig. 6. Card Identification algorithm

Fig. 7. The thirteen templates used

the same card. Furthermore, since the samples come from different corners of the card, overcomes situations where illumination reflectance affects one of the corners, thus blinding the camera in that region, but not the other. From both resulting scores, prevails the lowest score, i.e. the best match.

In order to make the matching more consistent, it was implemented a loop of successive capturing and extraction succeeded by matching along N-frames, during which the best matches found are stored. This loop is followed by a decision function which takes the stored matches and outputs the result of the identification. The decision function used is the weighted mean, which considers the identified ranks and the number of times each one occurred. The card rank which match is the closest to the weighted mean result, is the best match found for the correspondent card.

2.5 Suit Extraction and Identification

Both the suit extraction and the suit identification algorithms are similar to the rank extraction and the rank identification, respectively. It is worth mentioning that the playing cards, on which this algorithm is based, have one suit under the rank. Therefore the size and position of the region to crop are compatible with the size and position of the suit.

3 Chips Identification

In order to build a complete humanoid poker player, the information about the chips on the table is essential since it will have a role on the decision during game play.

The remove cards block performs the removal of the community cards present on the captured poker table overview. This procedure aims to prevent any misidentification between chips and playing cards, since the latter features

Fig. 8. Chips counting - algorithm overview

elements the same colour as the chips used, such as the suit. The removal of the cards is achieved with a mask which distinguishes between the playing cards region and the rest.

The colour segmentation aims to separate chips of different values on the game, which are only distinguishable by their colours. Within the colour segmentation block, the captured image undergoes a smoothing filter with the purpose of neutralizing high variations on the pixel values. The smoothing is followed by colour segmentation based on the RGB vector of each colour of chips. The threshold of each RGB component is within a fixed range in order to cover more points that can belong to the same chip. If one thinks of the RGB color space, this defined range can be seen as a cube centered in the RGB vector which represents the colour of the chip. This range overcomes the different illumination inherent to the environment. The resulting binarized image allows defining the regions of interest.

The previous procedure spots the chips as well as other kinds of objects of the same colour. In case the latter are present on the captured image, these must be discarded. Moreover, it is intended to design the algorithm as robust as possible by making the detection of partially superimposed chips, since it is common to happen during the game play. All these requirements are achieved using the Hough Circles Transform.

The Hough Circles Transform outputs better results, the better the contrast between the chips and the background. Therefore, preceding it, it is computed the absolute difference between the gray image of the chips with itself. Only difference is that the first has the brightness inverted, while the second does not. This procedure results in a great contrast between the chips and the table.

4 Experimental Results

In order to test the reliability of the algorithm, the webcam was placed on a tripod 81 cm high relative to the poker table. The tripod was placed in one of the seats of an octagonal poker table. In each round of the tests, three cards were dealt and placed along the table, such that the cards were 88cm, 103cm and 119cm away from the webcam. The resolution used was 800x600.

To determine the rank identification accuracy, it was registered how many times, for one specific rank, the identification was correct, incorrect or if the rank didnt meet the pre-requisites for the decision function to output it, which happens when the template match score is above 0.45. For the suit the same procedure applies. A total of 265 cards were analyzed with an equally distributed number of ranks and suits.

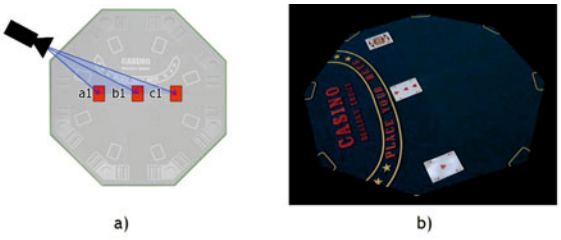


Fig. 9. a) Webcam setup relative to the three cards b) perspective view of the webcam

Table 1. Percentage of accuracy for rank identification

Card rank	King	Queen	Jack	10	9	8	7
Acc	96,0%	93,3%	94,5%	89,4%	94,2%	93,0%	96,3%
Card rank	6	5	4	3	2	A	
Acc	95,2 %	96,4%	93,6%	94,5%	95,4%	95,7%	

The results for the card rank identification are shown in the table above. Easily the 10 can be identified as being the one with the worst results. This relies on the fact that it is constituted by 2 digits, 1 and 0, and when the card is in the furthest position relative to the camera, it happens that in the pre-processing both of the digits get merged into one, making difficult the template matching. Just for reference, most of the times the 10 was identified as a 2. Concerning the suits, Spades features the worst accuracy, where most of the times was misidentified as Diamonds.

Table 2. Percentage of accuracy for suit identification

Card rank	Clubs	Hearts	Diamonds	Spades
Acc	94,4%	95,2%	99,0%	92,98%

5 Conclusion and Future Work

Regarding the rank identification, the method used here should not be discarded, since it achieves an overall accuracy of 94,4%. Concerning the chips identification, the presented algorithm enables to count chips accurately if these are not occluded. Otherwise, when more than 30% occluded, it presents errors on the counting. These errors are higher, the further the chips are from the capturing device. Some improvements should be performed in order to make the algorithm more reliable. The first concerns the distinction between a red card and a black card. This would immediately prevent the misidentification between red suits and black suits, as it happens frequently with the spades being misidentified as diamonds. Moreover, the upgrade of the video source to one with a higher

resolution would also improve the reliability of the system. The study and implementation of a stereo vision system would be of interest, in order to improve the counting of chips.

References

1. Billings, D.: Algorithms and Assessment in Computer Poker (2006)
2. Felix, D., Reis, L.P.: An Experimental Approach to Online Opponent Modeling in Texas Hold'em Poker. In: Zaverucha, G., da Costa, A.L. (eds.) SBIA 2008. LNCS (LNAI), vol. 5249, pp. 83–92. Springer, Heidelberg (2008)
3. Van, G., Kurt, B., Ramon, D.J.: Monte-Carlo Tree Search in Poker using Expected Reward Distributions. In: 1st Asian Conference on Machine Learning: Advances in Machine Learning, Nanjing, China, pp. 367–381 (2009)
4. Sklansky, D.: The Theory of Poker: A Professional Poker Player Teaches You How to Think Like One. Two Plus Two (2002)
5. Zheng, C., Green, R.: Playing Card Recognition Using Rotational Invariant Template Matching. University of Canterbury, Christchurch (2007)
6. Zutis, K., Hoey, J.: Who's Counting?: Real-Time Blackjack Monitoring for Card Counting Detection. University of Dundee (2009)
7. Hollinger, G., Ward, N.: Introducing Computers to Blackjack: Implementation of a Card Recognition System using Computer Vision Techniques. Colby College, Waterville (2003)
8. Chen, W.-Y., Chung, C.-H.: Robust poker image recognition scheme in playing card machine using Hotelling transform, DCT and run-length techniques. In: Digital Signal Processing, 3rd edn., vol. 20, pp. 769–779 (May 2010)

Occlusion Management in Sequential Mean Field Monte Carlo Methods

Carlos Medrano[1], Raúl Igual[2], Carlos Orrite[1], and Inmaculada Plaza[2]

[1] CVLab, Aragon Institute for Engineering Research,
c/ Mariano Esquillor s/n, 50018 Zaragoza, Spain
[2] EduQTech, E.U. Politécnica
c/ Ciudad Escolar s/n, 44003 Teruel, Spain
{ctmedra,rigual,corrite,iplaza}@unizar.es

Abstract. In this paper we analyse the problem of occlusions under a Mean Field Monte Carlo approach. This kind of approach is suitable to approximate inference in problems such as multitarget tracking, in which this paper is focused. It leads to a set of fixed point equations, one for each target, that can be solved iteratively. While previous works considered independent likelihoods and pairwise interactions between objects, in this work we assume a more realistic joint likelihood that helps to cope with occlusions. Since the joint likelihood can truly depend on several objects, a high dimensional integral appears in the raw approach. We consider an approximation to make it computationally feasible. We have tested the proposed approach on football and indoor surveillance sequences, showing that a low number of failures can be achieved.

Keywords: Multitarget tracking, occlusions, mean field.

1 Introduction

Object detection and tracking are basic tasks in computer vision with many potential applications. However, if a target is occluded its visual appearance changes a lot and conventional single-tracker approaches are very likely to fail. Therefore, tracking in crowded situations remains challenging.

Even though the approach we are going to present in this paper is general, our interest and experimental work focuses on human tracking mainly. The most promising approaches consider a Bayesian perspective. In [1] a multi-blob likelihood function is presented. It is based on the response of several filters at a grid of locations the image is overlaid with. A person is modelled as a generalised cylinder and the tracking problem is solved with a particle filter in the joint state space, which is also able to cope with objects entering or leaving the scene. In [2] the Hybrid Joint Separable filter, is presented. It considers both interactions and occlusions. Belief Propagation is applied to solve interactions, which are expressed as Markov Random Fields (MRF), whereas occlusions are tackled by an appearance likelihood that implements a physically-based model of the rendering process. The observation model is considered to be separable

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 444–451, 2011.

in image space and each term is approximated using a first order expansion. In [3] humans are modelled as ellipsoids in the 3D world, which allows a principled reasoning about image formation. The joint likelihood is based on the partition of the image into several regions corresponding to different objects and the background. A Maximum a Posteriori is found by running an algorithm derived from data driven Markov Chain Monte Carlo (MCMC). In [4–6] a variational approach is taken to track structured deformable shapes, articulated objects or multiple targets. Constraints for joint motion are modelled as pairwise MRFs, but every object is supposed to have its own independent observation likelihood. This approach leads to a set of collective Mean Field equations which allows finding a posterior estimation for each object.

In the top-down view images of ants analysed in [6], Mean Field was shown to be more effective than MCMC to avoid tracking failures. Thus, we think it is worth extending previous work using Mean Field [4–6] to deal with joint likelihoods and occlusions, which is a more realistic case. We point out the differences with previous formulations that only took into account MRFs and we perform some approximations to compute the multi-object functions that appear.

The paper is organized as follows. In section 2 we review the background of Mean Field and the equations that arise when pairwise potentials are included. Then we discuss how they change when a joint likelihood is considered. In section 3 we present the application of our approach to a football match sequence and to selected parts of indoor surveillance sequences, showing its weaknesses and strengths. Conclusions are drawn in section 4.

2 Theoretical Background

Let X_t be the joint state at time t, $X_t = \{x_{i,t}, i = 1 \dots M\}$ with M being the number of objects, and Z_t the measurement. The posterior $P(X_t|Z_t)$ is:

$$P(X_t|Z_t) \propto \phi(Z_t|X_t)P(X_t) \tag{1}$$

where $\phi(Z_t|X_t)$ is the likelihood function and $P(X_t)$ is the prior distribution. Applying a first order Markov assumption, the prior can be obtained by propagating the posterior at time t-1 using a dynamical model $P(X_t|X_{t-1})$. Given the practical difficulty in dealing with the posterior for a high dimensional state space, some kind of approximation has to be devised. In the conventional Mean Field approach we are assuming [4, 5], the posterior is approximated by the product of independent probability distributions, $Q_{i,t}(x_{i,t})$, which only depend on one target each:

$$P(X_t|Z_t) \approx \prod_i Q_{i,t}(x_{i,t}) \tag{2}$$

Distributions $Q_{i,t}(x_{i,t})$ are found by minimizing the Kullback-Leibler (KL) divergence between the true posterior and the product of independent distributions. As shown in [4, 5], this leads to:

$$Q_{i,t}(x_{i,t}) \propto exp(E_Q\left[logP(X_t, Z_t)|x_{i,t}\right]) \tag{3}$$

where

$$E_Q\left[logP(X_t, Z_t)|x_{i,t}\right] = \int_{X_t \backslash i} logP(X_t, Z_t) \prod_{j \neq i} Q_{j,t}(x_{j,t}) \tag{4}$$

and $X_t \backslash i$ means that the integration is taken over all the variables except $x_{i,t}$.

2.1 Previous Works

In previous works that used Mean Field for tracking [4–6], each target (or part of an object) was assumed to have its own measurement, $z_{i,t}$, so the likelihood was the product of individual likelihoods, $\phi(z_{i,t}|x_{i,t})$. In addition, pairwise MRFs $\psi(x_{i,t}, x_{j,t})$ were used to obtain a more realistic motion model. Under these assumptions, equation 3 can be rewritten to give [4–6]:

$$Q_{i,t}(x_{i,t}) \propto \phi(z_{i,t}|x_{i,t}) p_{i,t|t-1}(x_{i,t}) exp\left(\sum_{j \neq i} \int_{x_{j,t}} Q_{j,t}(x_{j,t}) log\psi(x_{i,t}, x_{j,t})\right) \tag{5}$$

where $p_{i,t|t-1}(x_{i,t})$ is the prior of object i assuming independent motion. Note that pairwise MRFs cause the argument of the exponential in equation 5 to be a sum over all the objects except the one under consideration, each addend being an integral over the state space of a single object.

Equations 5 ($i = 1 \ldots M$) are a set of fixed point equations that can be solved iteratively. In [4], a particle filter solution is adopted and the corresponding method is called Sequential Mean Field Monte Carlo (SMFMC). If $\{x_{i,t}^n, \omega_{i,t}^n\}_{n=1}^N$ is the particle set of object i, the difference with a traditional particle filter is that the weight $\omega_{i,t}^n$ also includes a factor regarding the discrete approximation of the exponential term in equation 5 [4–6].

2.2 Occlusions and Mean Field

In contrast to previous works, we assume that targets move independently but that the likelihood can not be decomposed into the product of independent likelihoods. In this case, equation 3 can be worked out to give:

$$Q_{i,t}(x_{i,t}) \propto p_{i,t|t-1}(x_{i,t})\, exp\left(\int_{X_t \backslash i} log\phi(Z_t|X_t) \prod_{j \neq i} Q_{j,t}(x_{j,t})\right) \tag{6}$$

The integral in the exponential is over all the objects except the one being considered, and it can not be split into the sum of single object integrals since $\phi(Z_t|X_t)$ depends jointly on all the targets. Even after discretization in a Monte Carlo implementation, the computation of the integral is a formidable task because it leads to many sums. In order to achieve a workable solution,

some approximation has to be taken. We perform a first order Taylor expansion of the log-likelihood around the mean state of objects other than object i:

$$log\phi(Z_t|X_t) \approx log\phi(Z_t|\hat{X}_{i,t}) + \sum_{j\neq i} \left[\frac{\partial log\phi(Z_t|X_t)}{\partial x_{j,t}}\right]^T_{\hat{X}_{i,t}} (x_{j,t} - \bar{x}_{j,t}) \qquad (7)$$

where $\bar{x}_{j,t}$ is the average state of object j, $\bar{x}_{j,t} = \int_{x_{j,t}} x_{j,t} Q_{j,t}(x_{j,t})$, and we have used the short notation $\hat{X}_{i,t} = \{\bar{x}_{1,t}, \ldots, x_{i,t}, \ldots, \bar{x}_{M,t}\}$, that is, the joint state in which all the objects are in their average state except the object under consideration, whose state is a free variable.

Since the expansion is around the mean term, the linear term gives no contribution at the end, so equation 6 now becomes simpler:

$$Q_{i,t}(x_{i,t}) \propto p_{i,t|t-1}(x_{i,t})\, \phi(Z_t|\hat{X}_{i,t}) \qquad (8)$$

This equation can be easily interpreted: regarding occlusions, object i interacts with the other targets as if they were in their mean states. The set of equations 8 with $i = 1 \ldots M$ is solved iteratively. From a given set of mean states, equation 8 allows updating $Q_{i,t}(x_{i,t}), i = 1 \ldots M$, which in turn allow computing a new set of mean states. The user has to provide initial distributions, but they can be based on simple reasoning: for instance, considering an incorrect but faster to compute individual likelihood, or just the prior $p_{i,t|t-1}(x_{i,t})$. An algorithmic implementation is shown in Fig. 1, which corresponds to a Sequential Importance Resampling filter solution to the approximation of the SMFMC method we are considering. We have found that a single improvement iteration gives good tracking results in all the tests performed.

2.3 The Joint Likelihood

Not every likelihood is suitable for a joint tracker. The likelihood should be able to compare fairly different hypotheses with different degrees of predicted occlusion or even with a different number of targets. In [1, 3, 7] several examples of joint likelihoods can be found, all of them sharing common features. We present our likelihood shortly, see [3] for a similar case. Given a joint state X_t, the image is partitioned into the expected background, $\bar{R}$, and the expected foreground, R. The expected foreground can be partitioned into the visible regions of each object, $R = \cup_i R_i$. Note that R_i are disjoint regions.The likelihood consist of two terms, foreground and background. The background term is based on the classification of each pixel u in $\bar{R}$ using a background substraction method and a threshold. The likelihood is increased when the pixel is classified as background, and decreased otherwise:

$$log\phi(Z_t^{\bar{R}}|X_t) = \lambda_{\bar{R}} \sum_{u\in\bar{R}} \phi_u,\ \phi_u = \pm 1 \qquad (9)$$

The object region likelihood is obtained from the comparison of colours histograms in R_i with reference histograms, using the Battacharyya coefficient

Input: resampled particle sets at $t-1$, $\{x^n_{i,t-1}, \frac{1}{N}\}^N_{n=1}$, $i = 1 \dots M$
Output: resampled particle sets at time t, $\{x^n_{i,t}, \frac{1}{N}\}^N_{n=1}$, $i = 1 \dots M$
(1) Initialization: $k \leftarrow 1$
1.a. Prediction: For each particle $\{x^n_{i,t-1}\}^N_{n=1}$ sample from $p(x_{i,t}|x^n_{i,t-1})$ to get $\{x^n_{i,t}\}^N_{n=1}$.
1.b. Assign initial weights $\omega^n_{i,t,k}, n = 1 \dots N$ and normalize. For instance, they can be based on simple models of individual likelihoods or just set to $1/N$.
1.c. Repeat 1.a and 1.b for $i = 1 \dots M$
(2) Iteration:
2.a Obtain the mean state of each object, $i = 1 \dots M$
$\bar{x}_{i,t,k} = \sum_n \omega^n_{i,t,k} x^n_{i,t}$
2.b. Re-weight: For each sample $\{x^n_{i,t}\}^N_{n=1}$, set its weight to:
$\omega^n_{i,t,k+1} \propto \phi(Z_t|\hat{X}^n_{i,t,k})$
where $\hat{X}^n_{i,t,k} = \{\bar{x}_{1,t,k}, \dots, x^n_{i,t}, \dots, \bar{x}_{M,t,k}\}$
2.c. Normalization: Normalize the weights $\omega^n_{i,t,k+1}$
2.d. Repeat 2.b and 2.c for $i = 1 \dots M$
2.e. Iteration: $k \leftarrow k+1$, iterate until convergence.
(3) Result:
$\{x^n_{i,t}, \omega^n_{i,t}\}^N_{n=1} \leftarrow \{x^n_{i,t}, \omega^n_{i,t,k}\}^N_{n=1}$ for each object $i = 1 \dots M$
(4) Resample:
resample $\{x^n_{i,t}, \omega^n_{i,t}\}^N_{n=1}$ to get $\{x^n_{i,t}, \frac{1}{N}\}^N_{n=1}$ for each object $i = 1 \dots M$

Fig. 1. Sequential Mean Field Monte Carlo algorithm with joint likelihood

between two histograms, $D(q, q^*)$. The first term in the exponential favours the difference between the hypothesis and the background, while the second favours its similarity with a reference object.

$$log\phi(Z^R_t|X_t) = \lambda_R \sum_i |R_i|(-\lambda_b D(q_i, q^*_{bg}) + \lambda_f D(q_i, q^*_i)) \tag{10}$$

Model parameters $\lambda_{\bar{R}}$, λ_R, λ_b, λ_f weight the relative importance of each term and give the likelihood a suitable bandwidth. $|R_i|$ is the surface of region i.

3 Experiments

3.1 Football Sequence

We have tested our approach on the sequence of football players from VS-PETS2003[1] data base, using the 2500 images in the TESTING/CAMERA3 directory. In this sequence there are many occlusions, some of them involving more than two people. Players are modelled as 2D ellipses. Player state is a six-dimensional vector, $\{y, x, h, w, v_y, v_x\}$: two coordinates, y-x, their two velocities, a height, h, and a width, w. The dynamical model is a constant velocity model in y-x with a velocity limit, and a Brownian model in the size. A weak prior in

[1] `http://www.cvg.cs.rdg.ac.uk/datasets/index.html`

Table 1. Number and kind of failures in the VS-PETS sequence

Coalescence	Target lost	Label interchange	Total
0.2	1.2	1.4	2.8

h-w around typical human values is also considered. 50 particles per object were used. The y-axis points downwards. In the view of the pitch used, it is reasonable to assume that object A occludes object B if $y_A > y_B$ and their images overlap.

Whenever there is no overlapping between the estimated and the ground thruth bounding boxes, the tracking is resumed and an error is recorder. The number of failures and their classification averaged over 5 runs are shown in table 1. The computation time averaged over the sequence is shown in table 2. About 22% of the time is spent in the calculation of the initial particle weights and about 66% in the mean field iteration that refines them.

As we can deduce from table 1, our algorithm manages very well most occlusions that occur in this well-known sequence. In Fig. 2 we show one of the failures that often happens, representing the bounding box of the human ellipse model. Two players of the same team stay for several frames at a similar position and finally labels 29 and 38 are interchanged.

In [8], the same sequence was analysed using a Multiple Hypothesis Tracker (MHT). 179 occlusions events were identified. The tracking was regarded as successful if two corresponding tracked targets exist before the event and have centers that are each within their ground truth targets after the event. MHT successfully tracked targets in 136 events. Thus, a failure was found in 43 occlusions. Despite the different approach for evaluating, it is clear that our method reduces the number of failures. Running independent particle filters is hopeless in this sequence, giving rise to several tens of failures.

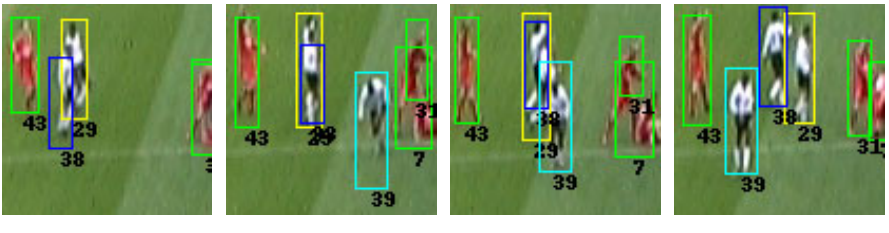


Fig. 2. Inset of some frames during a tracking failure in the VS-PETS football sequence. Frame numbers from left to right: 2215, 2227, 2235 and 2248. To be viewed in colour.

3.2 Indoor Surveillance Sequences

Next, we describe the results of our method on the indoor video set called CAVIAR[2]. In particular, we have used selected parts of some of the sequences

[2] `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`

Table 2. Computation time, Pentium D 3GHz, 2GB RAM, non optimized C code

Sequence	Number of objects	Time per frame and per object (s)	Time per frame (s)	Human height (pixels)
VS-PETS	4-20	0.029	0.407	35-60
CAVIAR-1	3	0.26	0.78	60-90
CAVIAR-2	5	0.63	3.15	60-140

Fig. 3. Selected frames of CAVIAR-2. Frame numbers from left to right and from top to bottom: 426, 446, 456, 470, 478 and 506. To be viewed in colour.

of the "shopping center corridor view", which show remarkable occlusions. We have treated this sequence as the previous one, except that we model human shape by a composition of three 2D ellipses corresponding to head, torso and legs. The state of each person has six components $\{y, x, h, w, \theta, v_y, v_x\}$: y-x are the coordinates of the centroid and v_y-v_x their velocities; h is a scale parameter that affects the three ellipses globally, while w is a width scale parameter that affects only the width of the ellipses; finally, θ is a global orientation angle. We use a constant velocity model in y-x and a Brownian model in the rest of parameters, in conjunction with limits in θ, v_y and v_x.

About five occlusions occur in the selected parts analysed, which we call CAVIAR-1 and CAVIAR-2, about 150 frames in total, and our algorithm does not fail in any of them. An example is shown in Fig. 3 (sequence CAVIAR-2), where we represent the average state too. Even though the labels are assigned correctly after the occlusion, the sequence of frames in Fig. 3 shows noticeable disturbance of the tracking. Image resolution provides a high detailed representation of a person that our simple model with three ellipses can not cope with. This is likely to cause the disturbance. A more detailed human model would help but at the expense of increasing the computation time.

The computation time averaged over the sequence is shown in table 2, together with some details about the number of people being tracked and their size. In these sequences we do not reach several frames per second in contrast to the previous experiment. There are several reasons for that. Firstly, the size of the objects is larger, so we have to compute many more pixels. Secondly, our selection of the region of interest for the likelihod computation is still non-optimal.

4 Conclusions

In this paper, we have dealt with occlusions using a Sequential Mean Field Monte Carlo Method, which was only used to cope with pairwise Markov Random Field in previous works. We have shown an approximation of the high dimensional integral that appears in the exact Mean Field calculation. The algorithm has been tested on several sequences, showing very good properties to avoid tracking failures during occlusions. However the application of our approach is limited by the high computation time if the target size is large. Thus our future work will focus on decreasing the computation time and on comparison with other approaches like Markov Chain Monte Carlo [3, 7].

Acknowledgments

This work was funded by the Spanish project TIN2010-20177.

References

1. Isard, M., MacCormick, J.: BraMBLe: A Bayesian Multiple-Blob Tracker. In: Proceedings of the IEEE ICCV, vol. 2, pp. 34–41 (2001)
2. Lanz, O.: Approximate Bayesian Multibody Tracking. IEEE Transactions on PAMI 9(28), 1436–1449 (2006)
3. Zhao, T., Nevatia, R., Wu, B.: Segmentation and tracking of multiple humans in crowded environments. IEEE Transactions on PAMI 30(7) (2008)
4. Hua, G., Wu, Y.: Sequential mean field variational analysis of structured deformable shapes. Computer Vision and Image Understanding 101, 87–99 (2006)
5. Hua, G., Wu, Y.: A decentralized probabilistic approach to articulated body tracking. Computer Vision and Image Understanding 108, 272–283 (2007)
6. Medrano, C., Herrero, J.E., Martinez, J., Orrite, C.: Mean field approach for tracking similar objects. Computer Vision and Image Understanding 113, 907–920 (2009)
7. Yao, J., Odobez, J.M.: Multi-Camera Multi-Person 3D Space Tracking with MCMC in Surveillance Scenarios. In: ECCV Workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications ECCV-M2SFA2 (2008)
8. Joo, S., Chellappa, R.: A Multiple-Hypothesis Approach for Multiple Visual Tracking. IEEE Transactions on Image Processing 11(16), 2849–2854 (2007)

New Approach for Road Extraction from High Resolution Remotely Sensed Images Using the Quaternionic Wavelet

Mohamed Naouai[1,2], Atef Hamouda[1], Aroua Akkari[1], and Christiane Weber[2]

[1] Faculty of Science of Tunis, University campus el Manar DSI 2092 Tunis Belvdaire-Tunisia Research unit Urpah
[2] Laboratory Image and Ville UMR7011-CNRS-University Strasbourg 3 rue de l'Argonne 67000 Strasbourg
naouai@polytech.unice.fr, atef_hammouda@yahoo.fr aroua.akkari@gmail.com, christiane.weber@lorraine.u-strasbg.fr

Abstract. Automatic network road extraction from high resolution remotely sensed images has been under study by computer scientists for over 30 years. In fact, Conventional methods to create and update road information rely heavily on manual work and therefore are very expensive and time consuming. This paper presents an efficient and computationally fast method to extract road from very high resolution images automatically. We propose in this paper a new approach for following roads path based on a quaternionic wavelet transform insuring a good local space-frequency analysis with very important directional selectivity. In fact, the rich phase information given by this hypercomplex transform overcomes the lack of shift invariance property shown by the real discrete wavelet transform and the poor directional selectivity of both real and complex wavelet transform.

Keywords: pattern recognition, features recognition, remote sensing, network road extraction, quaternionic wavelet, time frequency localization, high directional selectivity.

1 Introduction

With the development of remote sensing technology, especially the commercialization of high spatial resolution satellite and aerial images. Many interesting works have been devoted to road extraction. Among these, we could cite the adaptive template matching [9], snakes model[8][12], perceptual grouping [10], hyper-spectral approach [6], multi-scale or multi-resolution approach[3], GIS data guided [11]. The concept for road network extraction is relatively simple, but reliable processes remain a difficult challenge. There exists no generic algorithm sufficiently reliable for all practical use. Road extraction remains largely, at least in typical production environments, costly manual process [16]. Present main approaches are those like dynamic programming [1], texture analysis applied to a single layer, Snakes [19], Markov fields [13], neural networks [20], multiresolution analysis[4], multicriteria directional operator [14], mathematical

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 452–459, 2011.

morphology based on geometric shape [5]. All these models or algorithms are mainly based on radiometric characteristics and geometric constraints of road information in the imagery thus do not exploit fully the spectral information of roads[14]. So that, the road extraction methods still have problems in popularization and application, the extraction accuracy cannot satisfy the needs of engineering application, the automation is in a relatively low level and the performance is limited by either road materials or complex road networks. In this work we introduce a fully automatic and robust new approach for network road extraction based on spectral information in addition to radiometric characteristics and geometric constraints of roads. In fact, the use of the quaternionic wavelet underwrites a very important directional selectivity given by rich phase information, in addition to the good locality in space-frequency domain.

2 Quaternionic Wavelet Transform

Despite the advantage of the locality in both spatial and frequency domain, the wavelet pyramid of real-valued wavelets unfortunately has the drawback of being neither translation-invariant nor rotations invariant [17]. As a result, no procedure can yield phase information. This is one of the important reasons why researchers are interested in hypercomplex wavelet transforms like complex or the quaternion wavelet transforms. The quaternion wavelet transform is a natural extension of the real and complex wavelet transform, taking into account the axioms of the quaternion algebra, the quaternionic analytic signal [2] It overcomes the shift dependence problem caused by discritization of the continues real wavelet transform, in another hand, it really improve the power of the phase concept witch in the real wavelets is not possible, and in the case of the complex is limited to only one phase since the quaternion phase is composed from three angles. As proved in [2], the multiresolution analysis can also be straightforwardly extended to the quaternionic case. Eduardo proposed a quaternionic wavelet based on Gabor filter where the wavelet scale function h and the wavelet function g given by the equations:

$$h^q = g(x, y, \sigma_1, \epsilon) exp(i\frac{c_1 w_1 x}{\sigma_1}) exp(j\frac{c_2 \epsilon w_2 y}{\sigma_1}) \tag{1}$$

$$g^q = g(x, y, \sigma_2, \epsilon) exp(i\frac{\widetilde{c_1}\widetilde{w_1} x}{\sigma_2}) exp(j\frac{\widetilde{c_2}\epsilon\widetilde{w_2} y}{\sigma_2}) \tag{2}$$

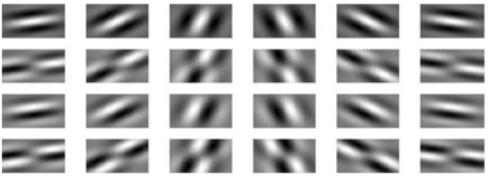


Fig. 1. Quaternion wavelet filters with selective orientations (from the left): 15, 45, 75, -75, -45, -15 degrees

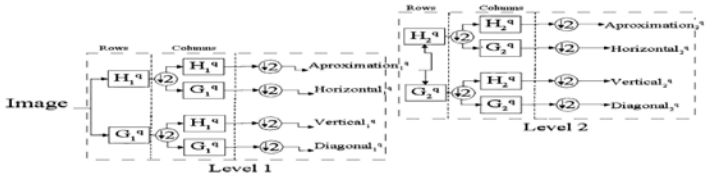


Fig. 2. Illustration of the pyramidal decomposition on quaternion wavelet: two levels

Note that the horizontal axis x is related with i and the vertical axis y is related with j, both imaginary numbers of the quaternion algebra fulfill the equation k=ji.

3 Our Approach for Network Road Extraction

The proposed work is composed of two main parts, the first consist in the localization step of road portion essentially based on radiometric information. As a result, this part provides us the control points. The second part, based on the control points, the frequency orientation representation given by the quaternion in wavelet decomposition in addition to the geometric information efficiently extracts the road map.

3.1 Road Detection Algorithm

In this work we introduce a road model in high spatial resolution remotely sensed images. This model is based on several properties with geometric and radiometric characteristics. Since this, in our case, the geometric and radiometric characteristics appear together, it is possible to apply a combination of these characteristics by representing a segmentation algorithm. We assume that each road segment is represented as an elongate rectangle has constant width and length, and they branches from often wide angles. The flow diagram of automatic road extraction process is shown in fig. 3. Automatic road extraction can be concentrated on road model, which embody the global features and local features of the road. So achieve to road detection, the key problem is correct description and understanding of the road and the establishment of appropriate road model. In this section, an innovative method (see fig. 8) is presented to guide the road extraction in an urban scene starting from a single complex high-resolution image (for details see)[15].

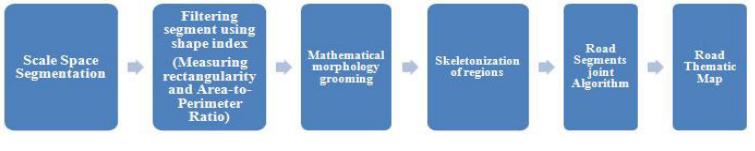


Fig. 3. Automatic road localisation process

Fig. 4. Result of algorithm detection

3.2 Control Point Selection

The alternative approach we follow consists in creating a version of the object that is as thin as possible, i.e. thinning the object to a set of idealised thin lines. The resulting thin lines are called the skeleton, or medial axis, of the input pattern and they are the thinnest representation of the original object that preserves the topology aiding synthesis and understanding. The methods to accomplish this are called thinning, or skeletonisation. The detection of end points, junction points and curve points of medial axis is important for a structural description that captures the topological information embedded in the skeleton. The thin lines can be converted into a graph associating the curve points with the edges, the end and the junction points with the vertices. Such a skeletal graph can then be used as an input to graph matching algorithms [18][**?**]. In this paper, we introduce algorithms for detecting controls points or skeletons characteristic points (see fig.7) (end points, junction points and curve points) based on a morphological approach. The skeleton image provided by the result of applying the detection algorithm described in [7].

End points and junction points. Formally we define the end points, the junctions points as follows:

Definition 1. *A point of one-pixel width digital curve is an end point if it has a single pixel among its 3×3 neighbourhood.*

Definition 2. *A point of one-pixel width digital curve is defined as a junction point if it has more than two curve pixels among its 3×3 neighbourhood.*

We propose another, purely morphological, method to detect end points and junction points from a skeleton as follows. To extract the end points, we perform erosion transform with the complement of each SE defining an end point, $\overline{A}$, and its rotations $\Theta_i(\overline{A})$ on the complement of X , $\overline{X}$, we take the union of all the results and then we intersect the union with X :

$$EndPoints(X) = [\cup_i \varepsilon\Theta_i\overline{A}(\overline{X})] \cap X \tag{3}$$

where$S = sum_i[\varepsilon\Theta_i(A)(X)]$ and $\varepsilon\Theta_i(A)(X)$ denote the erosion of X by $\Theta_i(A)$. at the same time we are able to detect.

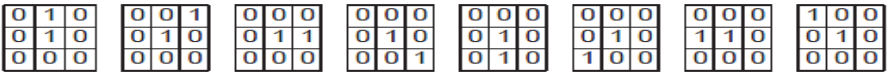


Fig. 5. SEs for the end points: the fundamental A (in bold) and its rotations $(\Theta_1(A), \Theta_2(A), \Theta_3(A), ..., \Theta_7(A))$in order

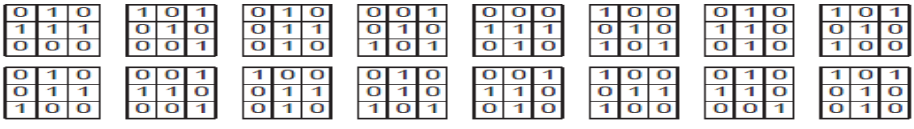


Fig. 6. SEs for the junction points: B (in bold) and its rotations $(\Theta_1(B), \Theta_2(B), \Theta_3(B), ..., \Theta_7(B))$ in the top row and C (in bold) and its rotations $(\Theta_1(C), \Theta_2(C), \Theta_3(C), ..., \Theta_7(C))$ in the bottom row, in order

According to the definition of junction points, only curve pixels are considered in the neighbouring configuration. In the eight-connected square grid, we can have two fundamental configurations corresponding to a junction point, B and C, and their seven rotations of 45° (see fig. 6). Thus, the extraction of the junction points from a skeleton is obtained by performing erosion transforms with each SE (B,C) and their rotations, $\Theta_i(B)$ and, $\Theta_i(C)$ and then taking the union of the results:

$$JunctionPoints(X) = [\cup_i \varepsilon \Theta_i B(X)] \cup [\cup_i \varepsilon \Theta_i C(X)] \quad (4)$$

All the curve points are trivially obtained by removing the end points and the junction points from the skeleton.

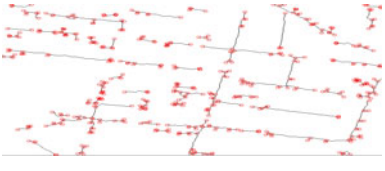


Fig. 7. Result of Control point detection

3.3 Frequency and Orientation Guided Joining Logic

Apart from the homogeneity of the grey value along the road, the assessment of the road hypotheses done in [16] is only based on geometric conditions. Spectral properties are not used. To use as much knowledge as possible for the verification of road hypotheses, we developed a new method for assessing road hypotheses based on the rich quaternionic phase information. Effectively, the transformation of the image using the Gabor filter based quaternionic wavelet insure us 16

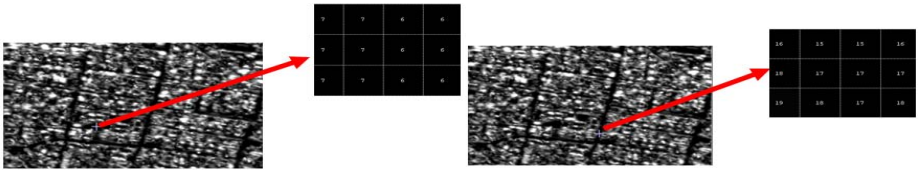


Fig. 8. The corresponding discrete frequency value of the 4×3 pixels for two similar radiometry region

frequency-orientation representation in each scale. Thus, we can observe the stability of the spacial frequency in different orientations. We then calculate the following parameters:

- Average of frequency in a given direction (fig. 8)
- Length Lmax<L<Lmin
- Average width Wmin<W<Wmax (fig.9)
- There exist parallel edges close to each other on both sides of the linear feature.

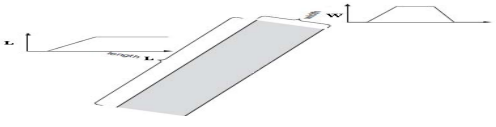


Fig. 9. Geometric criteria for the assessment of road hypotheses

4 Experimental Results

The quaternionic wavelet based on Gabor filters are in quadrature, they are implemented using parameters $\sigma_1 = \frac{\Pi}{6}$,$\sigma_2 = \frac{5\Pi}{6}$, c_1=c_2=3, $w_1 = 1$, $w_2 = 1$, ε=1 In our example we use 4 levels then we obtain 16 representations for each level showing the distribution of the frequency in different orientation.When calculating frequency average characterizing roads, experiments shows that is equal to 16. The proposed approach of automatic road extraction using quaternionic wavelet from high spatial resolution images can improve the accuracy of road extraction and reduce the effects of occlusions on roads such as shadows. The precision and the performance of road Network are validated through a Quickbird satellite image with very high-resolution. It allows also a precise follow-up of the roads and provides very good and concrete results(fig. 10).

Fig. 10. Final Road Network

5 Conclusion

We present here a robust approach for road network extraction using strong modeling tools, and giving pertinent results comparing with other approaches: as a matter of fact we use a radiometry based method for the detection of road portions then we introduce the mathematical morphology and squelitisation in order to extract basic control points and finally we opt to the innovative quaternionic wavelet transform to get a rich space frequency representation of the image in various orientation combined by the geometric information to finally join the basic control points and get the network road extracted efficiently. Experimental results show that our approach has given precise and accurate result for the detection of road networks from remote sensing images.

References

1. Gruen, A., Li, H.: Road extraction from aerial and satellite images by dynamic programming. Journal of Photogrammetry and Remote Sensing 50, 11–20 (1995)
2. Bayro-Corrochano, E.: The theory and use of the quaternion wavelet transform. J. Math. Imaging Vis. 24, 19–35 (2006)
3. Couloigner, I., Ranchin, T.: Mapping of urban areas: A multiresolution modeling approach for semi-automatic extraction of streets. Photogrammetric Engineering and Remote Sensing 66, 867–874 (2000)
4. Couloigner, I., Ranchin, T.: Mapping of urban areas: A multiresolution modeling approach for semiautomatic extraction of streets. Photogrammetric Engineering and Remote Sensing 66, 867–874 (2000)
5. Wang, Y.G., Zhu, C.Q., Ma, Q.H.: Automatic road extraction based on multiscale, grouping, and context. In: Proceedings of SPIE, the International Society for Optical Engineering, vol. 33, pp. 347–351 (2004)
6. Funk, C., Roberts, D., Gardner, M., Noonha, V.: Road extraction using spectral mixture and q-tree filter techniques (2001)
7. Gold, S., Rangarajan, A.: A graduated assignment algorithm for graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 377–388 (1996)
8. Gruen, A., Li, H.: Semi-automatic linear feature extraction by dynamic programming and lsb-snakes. Photogrammetric Engineering and Remote Sensing 63, 985–995 (1997)
9. Hu, X.: Automatic extraction of linear objects and houses from aerial and remote sensing imagery. Phd Dissertation, Wuhan University, China (2001)
10. Hu, X., Tao, C.V.: Automatic extraction of main-road from high resolution satellite imagery. International Archives of Photogrammetry and Remote Sensing 19, 293–298 (2002)
11. Klang, D.: Automatic detection of changes in road databases using satellite imagery. International Archives of Photogrammetry and Remote Sensing 32, 293–298 (1998)
12. Laptev, I., Lindeberg, T., Eckstein, W., Steger, C., Baumgartner, A.: Automatic extraction of roads from aerial images based on scale-space and snakes. Machine Vision and Application 12, 23–31 (2000)
13. Merlet, N., Zerubia, J.: New prospects in line detection by dynamic programming. IEEE Trans. Pattern Anal. Mach. Intell. 18, 426–431 (1996)

14. Naouai, M., Hamouda, A., Weber, C.: Detection of road in a high resolution image using a multicriteria directional operator. In: Computer Graphics, Visualization, Computer Vision and Image, vol. I (2009)
15. Naouai, M., Hamouda, A., Weber, C.: Urban road extraction from high-resolution optical satellite images. In: International Conference on Image Analysis and Recognition, vol. II, pp. 420–433 (2010)
16. Stefanidis, A., Doucette, P., Agouris, P., Musavi, M.: Self-organised clustering for road extraction in classified imagery. ISPRS Journal of Photogrammetry and Remote Sensing 55, 347–358 (2001)
17. Pan, H.-P.: Uniform full-information image matching using complex conjugate wavelet pyramids. In: XVIII ISPRS Congress, vol. 31 (1996)
18. Rodriguez, D.R., Rodriguez, G., Casta, L.: Recognition of shapes by morphological attributed relational graphs. In: Atti dell'VIII Convegno AIIA (2002)
19. Trinder, J.C., Li, H.: Semi-automatic feature extraction by snakes. In: Ascona 1995, pp. 95–104 (1995)
20. Bhattacharya, U., Parui, S.K.: An improved backpropagation neural network for detection of road-like features in satellite imagery. International Journal of Remote Sensing 18, 3379–3394 (1997)

On the Influence of Spatial Information for Hyper-spectral Satellite Imaging Characterization

Olga Rajadell, Pedro García-Sevilla, and Filiberto Pla

Depto. Lenguajes y Sistemas Informáticos
Jaume I University, Campus Riu Sec s/n 12071 Castellón, Spain
{orajadel,pgarcia,pla}@lsi.uji.es

Abstract. Land-use classification for hyper-spectral satellite images requires a previous step of pixel characterization. In the easiest case, each pixel is characterized by its spectral curve. The improvement of the spectral and spatial resolution in hyper-spectral sensors has led to very large data sets. Some researches have focused on better classifiers that can handle big amounts of data. Others have faced the problem of band selection to reduce the dimensionality of the feature space. However, thanks to the improvement in the spatial resolution of the sensors, spatial information may also provide new features for hyper-spectral satellite data. Here, an study on the influence of spectral-spatial features combined with an unsupervised band selection method is presented. The results show that it is possible to reduce very significantly the number of spectral bands required while having an adequate description of the spectral-spatial characteristics of the image for pixel classification tasks.

1 Introduction

Hyper-spectral images are the results of a detailed measurement of the spectra acquired by a special sensor. Currently, some sensors can easily cover a spectral resolution of 10nm with a considerably high spatial resolution that can reach 1m per pixel for satellite images. As a result, hyper-spectral images are composed by a very high number of correlated bands (between 200 and 500 spectral bands). Dealing with this type of images means facing a very high dimensional problem.

Since the usage of the whole hyper-spectral data set can fall into the course of dimensionality, several band selection methods have been studied in order to avoid the large amount of correlated data, while keeping the discrimination between land cover classes [1].

When the spatial resolution in hyper-spectral images was not high enough, major efforts to improve pixel classification were done focusing at the classification stage by simply using the spectral features provided by the sensors. These type of processing often used neural networks [2], decision trees [3], Bayesian estimation [4] and kernel-based methods [5] for the classification of the pixels in the images. In particular, Support Vector Machines proved to obtain good performances in this task [6].

Because of the increase in the spatial resolution, spectral-spatial analysis has been lately an issue of high interest for the improvement of hyper-spectral imaging characterization [7] which is widely used for tasks like land-cover classification and segmentation of remote sensing images. Some basic spatial features have been used like the

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 460–467, 2011.

mean value of a $N \times N$ window around a pixel, the standard deviation of the values in this window, and the combination of the mean and standard deviation for a series of window sizes [6]. On the other hand, textural analysis has been widely discussed to study the spatial relationships in an image. This sort of features could be applied over hyper-spectral images in order to have a better description of the spectral-spatial properties. There exists a huge variety of methods [8]: co-occurrence matrices, wavelet analysis, Gabor filtering, Local Binary Patterns, etc.

It is likely that improving the characterization of the image may help to reduce even more the amount of spectral bands required for the classification task. To pursue this goal, we have chosen two different spectral-spatial characterization methods. In first place, simple statistics (mean and standard deviation) of the $N \times N$ neighbors around a pixel will be considered for each spectral band. Later, a Gabor filter bank will be used to obtain features to describe the pixel in each band. Spectral-spatial feature extraction will be presented in Section 2. The hyper-spectral database used in our experiments is described in Section 3. The spectral-spatial methods proposed provide an improvement over the spectral classification as will be shown in Section 4. Finally, we draw out conclusions in Section 5.

2 Integration of Spatial Information in Imaging Characterization Methods

Pixel characterization aims at obtaining one feature vector for each pixel to be used in a pixel classification task in a multidimensional space. When only spectral data is used, the feature vector for every pixel is defined as the spectral curve provided by the sensor. The target of a spectral-spatial characterization method is to calculate a feature vector using the spectral data given and this can whether replace the spectral feature vector or being combined with it.

Let $I^i(x, y)$ be the i^{th} band of an image containing B bands. When the spectral curve is used as the feature vector for each pixel in the image this vector is simply composed of the values provided by the sensor, that is:

$$\psi_{x,y} = \left\{ I^i(x, y) \right\}_{i=1}^{B} \tag{1}$$

2.1 Basic Spatial Characterization

Spectral-spatial analysis of the image is based on a series of values extracted from spatial operations involving its neighbor pixels (spatial features) [9]. Frequently the two statistics used are the mean and the standard deviation of the neighborhood. This is a very simple method to include spatial information obtaining only 2 features per pixel [6].

Let $M_n^i(x, y)$ be the window $n \times n$ centered in the pixel (x, y) of the spectral band i. Then, the feature vector for this pixel is defined by:

$$\phi_{x,y} = \left\{ mean(M_n^i(x, y)), standard\_deviation(M_n^i(x, y)) \right\}_{i=1}^{B} \tag{2}$$

It is also possible to concatenate the features calculated from several window sizes (i.e. $n = 3, 5, 7, 9$) increasing the size of the vector ϕ depending on the number of windows used. This provides a multi-scale or multi-resolution description of the image.

2.2 Feature Extraction Based on Gabor Filters

Several features have been suggested in the literature for the description of texture information [8]. In this paper Gabor filtering will be used because, in general, they have provided the best results in different sort of texture characterization experiments [10] [11]. In this case, features are obtained by filtering the input image with a set of filters. The set of outputs obtained for each pixel in the image forms its feature vector. Here, the filter bank is defined to be a set of two-dimensional Gabor filters. Each Gabor filter is characterized by a preferred orientation and a preferred spatial frequency (scale). The filter acts as a local band-pass filter with optimal joint localization properties in the spatial domain and the frequency domain [12].

Gabor filters consist essentially of sine and cosine functions modulated by a Gaussian envelope. They can be defined by equation (3) where m is the index for the scale, n for the orientation and u_m is the central frequency of the scale [12].

$$f_{mn}(x,y) = \frac{1}{2\pi\sigma_m^2} \exp\left\{-\frac{x^2+y^2}{2\sigma_m^2}\right\} \times \cos(2\pi(u_m x \cos\theta_n + u_m y \sin\theta_n)) \tag{3}$$

Notice that set the condition $f_{mn}(0,0) = 0$ dismisses completely the effect of the measurements themselves and making the analysis independent from the pixel spectral values themselves.

Note that Gabor filters will be used in this case as a multi-dimensional extension of the technique designed for mono-channel images. In this way, multi-spectral images will be simply decomposed into separated channels and the same feature extraction process will be performed over each channel as shows equation (4).

$$h_{mn}^i(x,y) = I^i(x,y) * f_{mn}(x,y) \tag{4}$$

These responses are used to construct the final feature vector for each pixel.

$$\Upsilon_{x,y} = \{h_{mn}^i(x,y)\}_{\forall i,m,n} \tag{5}$$

3 Hyper-spectral Data Set

To pursue the experimental campaign a widely used hyper-spectral database has been used, 92AV3C, known as AVIRIS. Figure 1 show a color composition, its corresponding ground-truth and the classes in it.

Hyper-spectral image data 92AV3C was provided by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) and acquired over the Indian Pine Test Site in Northwestern Indiana in 1992. From the 220 bands that composed the image, 20 are usually ignored because of the noise (the ones that cover the region of water absorption or with low SNR [7]). The image has a spatial dimension of 145×145 pixels. Spatial resolution is 20m per pixel. Fig. 1 shows the sixteen available classes, ranging from 20 to 2468 pixels in size. In it, three different growing states of soya can be found, together with other

(a)

(b)

class	color	No. of samples
Stone-steel towers		95
Hay-windrowed		489
Corn-min till		834
Soybeans-no till		968
Alfalfa		54
Soybeans-clean till		614
Grass/pasture		497
Woods		1294
Bldg-Grass-Tree-Drives		380
Grass/pasture-mowed		26
Corn		234
Oats		20
Corn-no till		1434
Soybeans-min till		2468
Grass/trees		747
Wheat		212

(c)

Fig. 1. Hyper-spectral image AVIRIS (92AV3C). a)Color composition. b)Ground-truth. c)Target classes contained.

three different growing states of corn, woods, pasture and trees are the larger classes in terms of number of samples (pixels). Due to the small size of the rest of classes they are frequently dismissed in literature. In this paper, we will perform experiments using both 16 and 9 classes.

4 Spectral/Spatial Classification Results

As it has been pointed out, remote sensing has to deal with high dimensional feature vector where features are highly correlated. Consequently, band selection methods are frequently used. In our case, a band selection method presented by Martinez et al. in [1] has been used. Let D be a number of spectral bands such as $D \leq B$, where B is the total number of bands included in the database. This method provides the best set of D bands in term of uncorrelated information. It is based on a clustering approach that joins groups of bands depending on their mutual information. Once a partition of D groups is available, a representative band from each group is selected.

4.1 Classification Task

In Figures 2 and 3 a global view of the classification results using different spectral-spatial characterization methods can be found. The classification rates using only spectral information has also been included to be considered as a baseline reference. These results show the overall accuracy for four different sizes of windows to extract spatial information of the pixels ($3 \times 3, 5 \times 5, 7 \times 7, 9 \times 9$), the combinations of these spatial features which is just a concatenation of all of them, and the Gabor textural features

using two scales and four orientations. Every characterization method has been tested with the corresponding set of bands provided by the band selection algorithm from 1 to 15. Also the task with all bands in the dataset has been performed and can be observed in Table 1.

In these experiments, the pixels that form the whole image were divided into twenty non overlapping sets, keeping the a priori probability of each class. Therefore, no redundancies are introduced and each set is a representative set of the original image. The same sets of pixels are used in all experiments. Ten classification attempts were carried out with the k-nearest neighbor classifier with $k = 3$ and the mean of the error rates of these attempts was taken as the final performance of the classifier for this experiment. Each classification attempt uses one of these sets for training and another set for testing. Each set is never used twice, so the attempts are totally independent.

Experiments using all 16 available classes are shown in Figure 2. As an alternative, experiments excluding the classes with a reduced number of samples have also been carried out using the same criterion as in [6]. Their results are presented in Figure 3. Better results, as expected, were got in this case. Small classes represent small structures in the image that are hard to recognize since their size is not enough to be capture by spatial features. Furthermore, some neighborhoods may be too big that several spatial structures could be considered at a time.

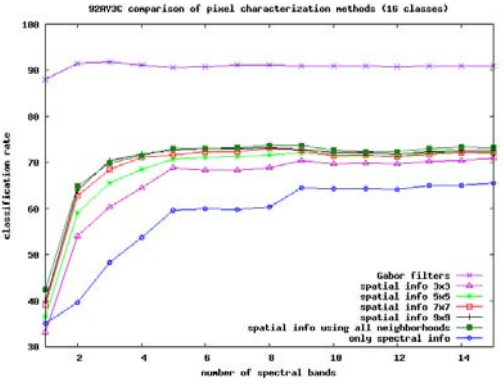


Fig. 2. Pixel classification rates for the 92AV3C database using all 16 classes. The number of spectral bands selected varies from 1 to 15.

Significant differences were obtained between spectral-spatial features and only spectral features even if the basic spatial features were used. Regarding these last sort of features, observe also that the larger the neighborhood used, the better classification results were obtained. Also, the concatenation of features obtained using different window sizes did not improve the results provided by using only the largest window. This means that, in this case, the spatial characterization is more reliable when we describe pixels by a fairly stable neighborhood. Furthermore, Gabor textural features outperformed all other methods very significantly. This points out that detailed spatial information is really discriminative for land use classification in this sort of images.

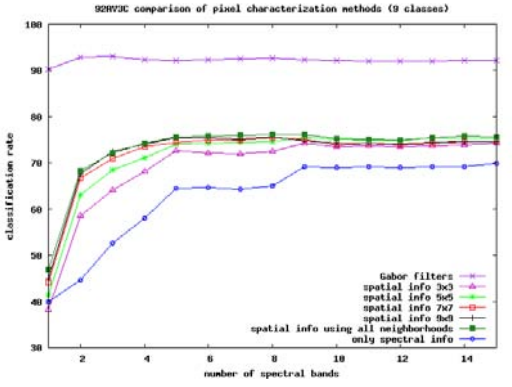


Fig. 3. Pixel classification rates for 92AV3C database using only the main 9 classes. The number of spectral bands selected varies from 1 to 15.

The differences between the characterization methods are not only due to the final classification rates obtained. Note also, that the number of spectral bands required to reach the stable behavior (where more spectral bands do not improve the classification results) is quite different. While spectral features require more than 12 bands, basic spatial features reach the stable zone with only 6–8 bands, while Gabor textural features required only 2–3 spectral bands.

In Table 1 the results obtained for several numbers of spectral bands can be compared with those obtain when using all 200 bands. Notice that, no matter the set of features used, no improvement is obtained by increasing the incoming data although the size of the problem is considerably increased.

Table 1. Accuracy for the 16 classes classification experiments of the 92AV3C dataset using different features. Results from the first sets of bands have been included together with the results obtained using the complete database (200 spectral bands).

	Characterization methods			
# of spectral bands	Spectral information	Spatial window 9×9	Spatial All windows	Gabor features
1	34.964	39.916	42.367	88.049
3	48.361	70.451	69.851	91.885
5	59.652	72.612	72.939	90.553
7	59.765	72.957	73.212	91.036
9	64.534	72.879	73.635	90.977
200	52.849	73.521	73.633	90.456

4.2 Segmentation Results

Since the problem we are tackling involve land-use pixel classification, the percentages of correct may not be enough to appreciate the goodness of the results. Pixel classification experiments assign a class label to each pixel in the test set. If we represent these

labels in the position of their corresponding pixels we will obtain a segmentation map of the processed image. In Figure 4 this representation of the results is shown where misclassified pixels (errors) are represented in white color, while the rest of pixels are represented by their own class color as presented in the ground-truth shown in Figure 1. The results shown correspond to classification experiments where only one set of pixels was used for training (5% of the pixels in the image) and the other 19 sets of pixels were used for testing, using 10 spectral bands. Only the results for three characterization methods are shown. On the left, the results using the basic spatial features extracted from all different window sizes ($3 \times 3, 5 \times 5, 7 \times 7, 9 \times 9$) are shown. Results using Gabor features are shown in the center of the figure. The results obtained using only spectral information are presented on the right.

Observe how the errors are distributed over the different classes. Spectral features (on the right) suffer from salt and pepper classification noise since the error is all over the areas an is not localized. However, when using Gabor textural features, the errors are located mainly in small areas and at the borders of the classes were the spatial features are mixing information from the heterogenous background. We could say that the areas recognized using these features are more homogeneous. In the case of the basic spatial features, the errors are distributed in a similar way to the ones obtained using Gabor features but the results are worse in this case, so the misclassified pixels extent deeper inside the classes.

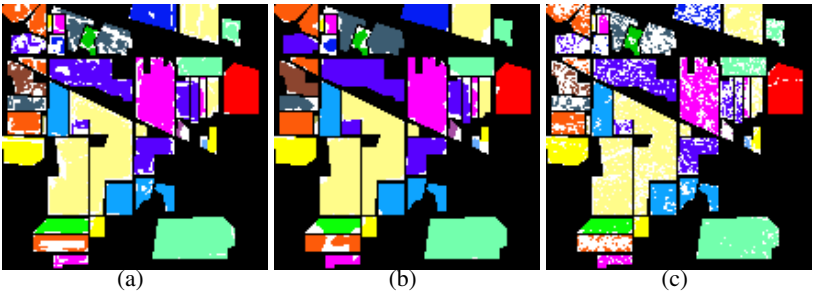


Fig. 4. Pixel classification results using 5% of the pixels for training for the 16 classes of the 92AV3C database, using 10 spectral bands. (a) Basic spatial features for all window sizes considered (b) Gabor textural features (c) Spectral features.

5 Conclusions

An experimental campaign over the 92AV3C dataset has been performed using several spectral-spatial characterization methods. Among them, the basic spatial features using simple statistics derived from a neighborhood and a Gabor textural features for a filter bank with two scales and four orientations have been used. Both basic and Gabor features outperform the naive spectral classification pointing out that taking advantage of the spatial resolution in the image is highly recommended for pixel classification tasks. Besides, Gabor textural features have provided very good classification results using a basic K-nearest neighbor classifier. Spectral features never provided results close to the

ones obtained using spatial information even when all two hundred spectral features were considered. In the segmentation experiments, spatial features have also proven their good performance providing quite homogenous regions and keeping the classification errors near the boundaries of the classes due to the influence of the heterogenous background. Furthermore, the good classification results obtained using spatial features required a minor number of spectral bands. Therefore, the use of spatial information can reduce the number of spectral bands required for pixel classification tasks and, at the same time, improve the rates of pixel classification.

Acknowledgments

This work has been partly supported by grant FPI PREDOC/2007/20 from Fundació Caixa Castelló-Bancaixa and projects CSD2007-00018 (Consolider Ingenio 2010) and AYA2008-05965-C04-04 from the Spanish Ministry of Science and Innovation.

References

1. Martínez-Usó, A., Pla, F., García-Sevilla, P.: Clustering-based hyperspectral band selection using information measures. IEEE Trans. on Geoscience & Remote Sensing 45, 4158–4171 (2007)
2. Yang, H., Meer, F., Bakker, W., Tan, Z.: A back-propagation neural network for mineralogical mapping from aviris data. International Journal of Remote Sensing 20, 97–110 (1999)
3. Zhou, H., Mao, Z., Wang, C.: Classification of coastal areas by airbone hyperspectral image. In: Proceedings of SPIE, pp. 471–476 (2005)
4. Chen, C., Ho, P.: Statistical pattern recognition in remote sensing. Pattern Recognition 41, 2731–2741 (2008)
5. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Trans. on Geoscience & Remote Sensing 43, 1351–1362 (2005)
6. Plaza, A., et al.: Recent advances in techniques for hyperspectral image processing. Remote Sensing of Environment 113, 110–122 (2009)
7. Landgrebe, D.A.: Signal Theory Methods in Multispectral Remote Sensing. Wiley, Hoboken (2003)
8. Petrou, M., García-Sevilla, P.: Image Processing: Dealing with Texture. John-Wiley and Sons, West Sussex (2006)
9. Jimenez, L., Landgrebe, D.: Hyperspectral data analysis and supervised feature reduction via projection pursuit. IEEE Trans. on Geoscience and Remote Sensing 37(6), 2653–2667 (1999)
10. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. IEEE Trans. on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)
11. Rajadell, O., García-Sevilla, P., Pla, F.: Filter banks for hyperspectral pixel classification of satellite images. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 1039–1046. Springer, Heidelberg (2009)
12. Fogel, I., Sagi, D.: Gabor filters as texture discriminator. Biological Cybernetics 61, 103–113 (1989)

Natural Material Segmentation and Classification Using Polarisation

Nitya Subramaniam, Gul e Saman, and Edwin R. Hancock*

Department of Computer Science, University of York, UK

Abstract. This paper uses polarisation information for surface segmentation based on material reflectance characteristics. Both polarised and unpolarised light is used, and the method is hence applicable to both specular or diffuse polarisation. We use moments to estimate the mean-intensity, polarisation and phase from images obtained with multiple polariser orientations. From the Fresnel theory, the azimuth angle of the surface normal is determined by the phase angle and for a limited range of refractive index the zenith angle is determined by the degree of polarisation. Using these properties, we show how the angular distribution of the mean intensity for remitted light can be parameterised using spherical harmonics. We explore two applications of our technique, namely a) detecting skin lesions in damaged fruit, and b) exploiting spherical harmonic co-efficients to segment surfaces into regions of different material composition using normalized graph cuts.

Keywords: Image classification, Image region analysis, Image segmentation, Image texture analysis.

1 Introduction

Polarisation is an important source of information conveyed by light, which is used to augment the visual capabilities of certain animals (e.g. the Mantis shrimp). Although the human vision system is insensitive to polarisation, it has proved to be a useful additional source of information in machine vision applications [13]. Polarisation imaging has been used to develop a variety of machine vision techniques, including surface quality inspection [10], shape recovery [1], [5] and material characterisation [9], [3]. Additionally, polarisation can also be used to infer information concerning the reflectance properties of surfaces. For instance, Atkinson and Hancock have shown in [2] how diffuse polarisation can be used to estimate the bidirectional reflectance function. However, their method is computationally demanding, using simulated annealing to estimate the BRDF.

Wolff has [9] used the Fresnel theory of light to develop a polarisation based method for differentiating between metal surfaces and dielectrics.The Fresnel theory of light (see [7], [4]) is a quantitative description of reflection and refraction at a smooth boundary between two media. The analysis is relatively straightforward for dielectrics, but the situation is less tractable for metals due to the induction of surface currents by the time

* Edwin Hancock is supported by a Royal Society Wolfson Research Merit Award.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 468–475, 2011.

varying electromagnetic field of light. In dielectrics, polarisation in remitted light may arise in two different ways. In the case of specular polarisation, initially polarised light is reflected in the specular direction. For diffuse polarisation, initially unpolarised light is refracted into the surface and the remitted light acquires a spontaneous polarisation. In both cases, the zenith angle of the reflected or remitted light is determined by the degree of polarisation and the azimuth angle determines the phase angle.

Metals and dielectrics can be differentiated on the basis of the phase of the polarisation as metal retards light waves and hence, changes the phase of the polarisation of the incident light wave on specular reflection while there is no change in the phase of the polarisation of the incident light for dielectrics [8].

In this paper, we adopt a simpler approach of the problem. For fixed light source direction and naturally occuring samples, provided that the range of refractive indices for different materials in a scene is limited, the polaristion image allows the angular distribution of reflected or remitted light to be estimated. Here we parameterise the distribution using spherical harmonics. Vectors of harmonic coefficients are then used to characterise the reflectance distribution on a pixel-by-pixel basis. We can then segment a scene into regions of different reflectance properties using the coefficient vectors. Here we compute, the difference in reflectance characteristics using the Mahalanobis distance between coefficient vectors and by using normalised cuts [12] to segment the scene into regions of different material composition.

2 Polarisation Image

When scattered light is measured through a linear polarising filter, the intensity changes as a sinusoidal function of the polariser angle α_p and the transmitted radiance sinusoid (TRS) is given by:

$$I(\alpha_p) = \frac{(I_{max} + I_{min})}{2} + \frac{(I_{max} - I_{min})}{2} cos(2\alpha_p - 2\phi) \quad (1)$$

where I_{max} is the maximum brightness, I_{min} the minimum brightness and ϕ the phase angle. It is more convenient to write the above formula in terms of the mean-intensity: $\hat{I} = (I_{max} + I_{min})/2$ and the degree of polarisation ρ:
$I(\alpha_p) = \hat{I}(1 + \rho \cos(2\alpha_p - 2\phi))$.

Suppose that we take N equally spaced polarisation images, so that the polariser angle index is $p = 1, 2, ..., N$. Let $x_p = (I(\alpha_p) - \hat{I})/\hat{I}$, $\hat{x} = 1/N \sum_{p=1}^{N} x_p$ and $\sigma^2 = 1/N \sum_{p=1}^{N} (x_p - \hat{x})^2$.

The moments estimators of the three components of the polarisation image are the mean intensity, degree of polarization and phase:
$\hat{I} = 1/N \sum_{p=1}^{N} I(\alpha_p)$, $\rho = \sqrt{2/\pi}\sigma$ and $\phi = 1/2 \cos^{-1}(\langle \hat{x} \cos(2\alpha) \rangle / \pi\rho)$.

We use a moment-based method together with least squares fitting to estimate the degree and phase of polarisation in the scattered light. To improve the robustness of our calculation, observations with large deviations (more than 25% of the TRS amplitude) are not used in the estimation.

From the Fresnel theory, it is straightforward to show that the azimuth angle for reflected polarised light or remitted diffusely polarised light is equal to the phase angle,

ϕ [9]. The zenith angle, θ depends on whether the polarisation is specular or diffuse. For diffuse polarisation, the polarisation degree is:
$\rho_d = \frac{(n-1/n)^2 \sin^2\theta}{2+2n^2-(n+1/n)^2 \sin^2\theta+4\cos\theta\sqrt{n^2-\sin^2\theta}}$ while the degree of specular polarisation is:$\rho_s = \frac{2\sin^2\theta\cos\theta\sqrt{n^2-\sin^2\theta}}{n^2-\sin^2\theta-n^2\sin^2\theta+2\sin^4\theta}$ where n is the refractive index.

Here we aim to use the above equations to analyse the distribution of relectance from approximately planar samples of different material. Provided we know whether we are measuring the specular polarisation of reflected polarised light, or the diffuse polarisation of remitted initially unpolarised light, then θ and ϕ are the zenith and azimuth angles of light with respect to the surface normal. We assume the range of refractive index is small, and can be treated as a constant $(n = 1.45)$, which is typical of a wide range of dielectrics.

3 Reflectance Distributions

The observation underpinning this paper is that under the restrictions of sample planarity and slowly varying refractive index, the polarisation image allows us to measure the distribution of mean intensity, $\hat{I}$ with the zenith and azimuth angle of remitted light, θ and ϕ. To provide some illustrative motivation, Fig.1 shows a scatter plot of the intensity versus the degree of polarisation and surface azimuth angle for real and plastic leaves. The leaves are approximately planar, and the angle of incidence is approximately $15\,^\circ$. Here we work with initially unpolarised light and use the formula for diffuse reflectance to estimate the zenith angle from the measured polarisation. There are a number of features to note from the plot. First, the distributions are quite different for the two materials. We attribute this to the fact that natural leaves have a layered sub-surface structure, which affects distribution of remitted light through subsurface refraction according to Snell's law. Artifical leaves do not exhibit such structure. Second, when the refractive index is changed within the known range for dielectrics, there is

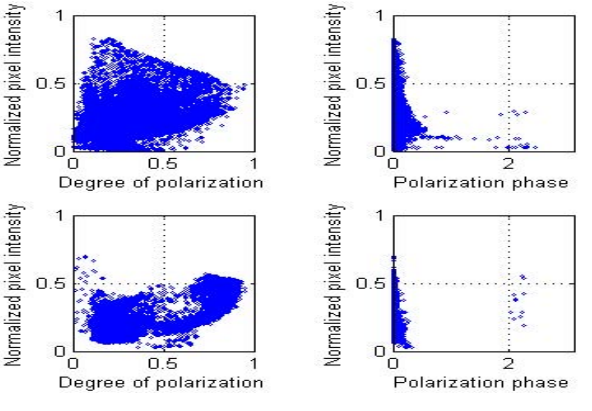


Fig. 1. (Top row) plastic and (Bottom row) real leaves, Scatter plots: The variations in pixel intensity plotted against ρ and ϕ

a small shift in the plots at all zenith angles. Since the shift is uniform, the effect of approximating refractive index in the feature calculations can be neglected. Thus, we demonstrate that polarisation information can be used for material-based classification in applications where colour-space classification may fail.

3.1 Calculating Spherical Harmonic Features

Our idea is to parameterise the distribution of mean intensity as a function of the azimuth and zenith angles. The polarisation image consists of a set of triples $[P = \{(\hat{I}_i, \rho_i, \phi_i), i = 1, ..., M\}]$ from which we compute the set $[D = \{(\hat{I}_i, \theta_i, \phi_i), i = 1, ..., M\}]$. Using the expression for diffuse polarisation in terms of zenith angle in (**??**). The distribution of mean image intensity at each pixel is expressed as a function of azimuth and zenith angles. Any such spherically symmetric function $f(\theta, \phi)$ can then be expressed as a weighted sum of the orthonormal basis functions Y_l^m (called the spherical harmonics of degree l and order m) as follows:

$$f(\theta, \phi) = \sum_{l=1}^{\infty} \sum_{m=-l}^{l} a_{l,m} Y_l^m(\theta, \phi), a \in \mathbb{R} \tag{2}$$

where $Y_l^m(\theta, \phi)$ is a function of the associated Legendre polynomials $P_l^m(z)$ with $z = \cos\theta$, given by $Y_l^m(\theta, \phi) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_l^m(\cos\theta) e^{im\phi}$.

Using the orthonormality properties of the spherical harmonics, the coefficients are given by

$$a_{l,m} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) Y_l^m(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi. \tag{3}$$

From (3), we obtain the following moments estimators of the spherical harmonic coefficients $a_{l,m} = \frac{1}{M} \sum_{i=1}^{M} \hat{I}_i Y_l^m(\theta_i, \phi_i)$ of the mean-intensity distribution.

In practice we estimate the set of coefficients over non-overlapping 10x10 blocks of pixels, and truncate the spherical harmonic expansion at $l = 8$ with m varying from $-l$ to l. As a result the mean intensity distribution in each pixel block is parameterised by a 81 element vector of spherical harmonic coefficients $A = [a_{0,0}, a_{1,-1}, a_{1,0}, a_{1,1}, ..., a_{8,8}]^T$.

The estimation of harmonic functions in the literature includes residual fitting approaches by [11] and the spherical Fast Fourier Transform by [6]. We use a MATLAB function to compute the Legendre polynomials and a moments based approach to estimate the coefficients, $a_{l,m}$. We divide the image into windows and calculate the average coefficients over each window. The window size is chosen to ensure that the instensity function is a reasonable representation of shape while taking care to not over-smooth the features.

3.2 Distribution of Information in the Feature Vector

We aim to use the coefficient vectors for both segmenting and classifying regions in scenes. To this end we commence by computing the variance matrix over blocks of the image. If the image blocks are indexed by $k = 1, ..., L$ and the k-th block has

coefficient A_k, then the mean coefficient vector is $\hat{A} = 1/L \sum_{k=1}^{L} A_k$ and the covariance matrix is $\Sigma_A = 1/L \sum_{k=1}^{L} (A_k - \hat{A})(A_k - \hat{A})^T$. The Mahalanobis distance between the coefficient vectors for the blocks indexed k_1 and k_2 is $D_{k_1,k_2} = (A_{k_1} - A_{k_2})^T \Sigma_A^{-1} (A_{k_1} - A_{k_2})$.

From the Mahalanobis distance we compute the $L \times L$ block affinity matrix S with elements $S(k_1, k_2) = \exp[-D_{k_1,k_2}]$. We segment the polarisation image into regions by recursively applying Shi and Malik's [12] algorithm to the affinity matrix.

4 Experiments

We explore two experimental applications of our new technique. First, we aim to use polarisation information to detect skin lesions in damaged fruit. Second, we aim to use the method to segment scenes into regions of different material composition. The images used in our study are recorded in a darkened room with matte black walls and working surfaces. The studied objects and the camera are positioned on the same axis and a halogen light source (visible spectrum) is positioned at approximately $15\,^\circ$ from the viewing axis for the leaves and $20\,^\circ$ for the fruits experiment, to reduce specular reflection. Linear polarising filters are placed in front of the source and the camera. The camera polaroid is rotated through $180\,^\circ$ at $30\,^\circ$ and $10\,^\circ$ intervals, respectively for the leaves and fruit. The images are captured with fixed aperture size and exposure time, using a Nikon D200 camera. Further experiments have been conducted in uncontrolled environments with natural light and outdoor scenes, with encouraging results.

4.1 Skin Lesions

We have taken a subimage of size 240×240 pixels to study, which shows the affected area of the plum. The degree and phase of diffuse polarisation are shown in Fig.2 for different stages of rotting. There are a number of features to note from the polarisation data. First, the degree and phase of polarisation reveal the boundaries between the undamaged and bruised surface regions in the scene.

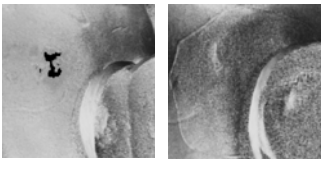


Fig. 2. (Left to Right)Polarisation phase on day 3 and day 4 showing changes in increasing stages of rotting for apple and plum, respectively

Figure 3 shows the spherical harmonic coefficients for the two subimages. The main feature to note is that the different objects define different regions in the data and that the co-efficient variation is greatest in the damaged areas.

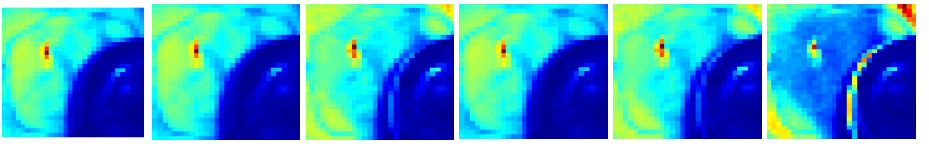


Fig. 3. Spherical harmonic coefficients for the section containing apple and plum for l=0,1 and 2 and positive values of m

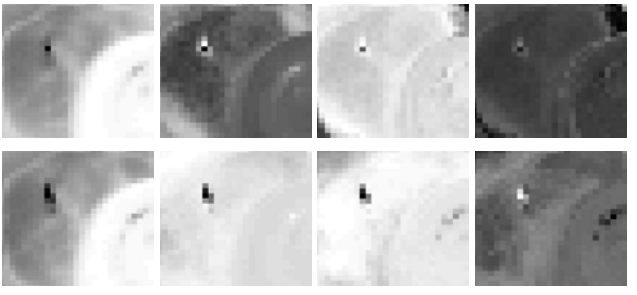


Fig. 4. The feature vector for the section of the test scene with apple and plum (Top row) relatively fresh (day 3) and (Bottom row) rotten (day 4)

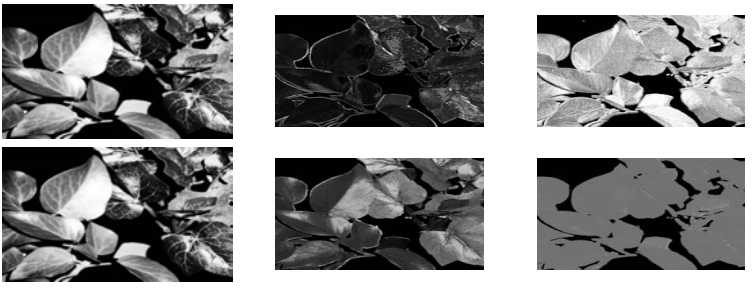


Fig. 5. Components of the polarisation image computed in unpolarised light (Top row) and polarised light(Bottom row)for the leaf scene are shown

The polarisation degree captures edges and fine surface texture in unpolarised light and coarse features in polarised light. The polarisation phase captures more surface detail in unpolarised light than in polarised light as demonstrated in Fig.5 for a mixed scene of objects. We have performed PCA on the coefficient vector covariance matrix Σ_A. The results are shown in Fig.6 and 4 which show the first four principal components which accout for 95% of the variance. These four components are used to compute a block-by-block feature vector. The features emphasize the vascular structure of real leaves and are weaker in polarised light. They also show the difference in the healthy and rotten fruit tissues. Different fruits can be seen in Fig.4 because of the spontaneous polarisation of light. Multiple scattering within the real leaf is harder to detect in strongly polarised light, however the shape vector is unaffected.

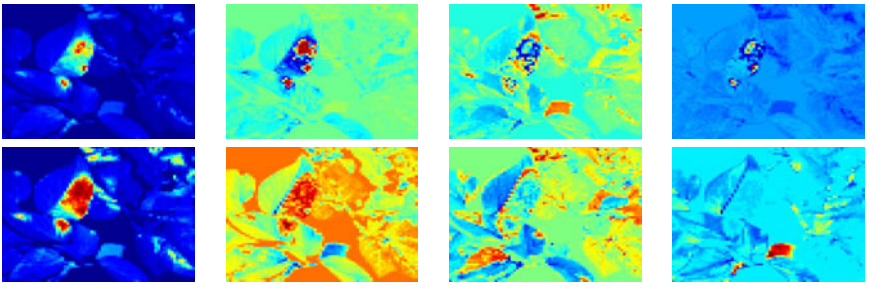


Fig. 6. Feature images (Left to Right): The first four features calculated from pca-mapped spherical harmonics coefficients

Fig. 7. Segmentation: The image is segmented using normalized cuts into (Left to Right) background, natural leaves and plastic leaves

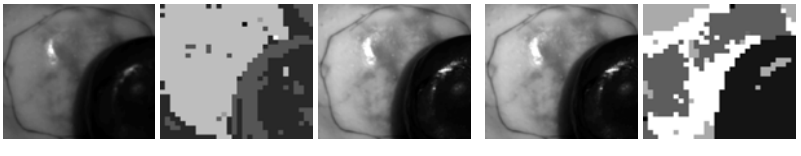


Fig. 8. Segmentation results: The segmentation for section of apple and plum, with each segment coded in a different graylevel. The first two images are for (day 3), while the last two are for (day 4).

4.2 Region Segmentation

To take this study one step further, we have explored whether the spherical harmonic co-efficients can be used to segment a scene into surfaces of different material composition. The results of segmenting the leaf scene, using the normalized graph cuts algorithm from [12] are shown in Fig.8. These results were obtained with 81 and 289 features from spherical harmonic coefficients up to degree 8 and 16 on 660x720 and 240x240 images for leaves and fruits, respectively. The affinity matrix was computed using the Mahalanobis distances between the feature vectors in blocks of 10x10 and 8x8 pixels, respectively for leaves and fruits. Specularities cause some difficulty in correct segmentation when using polarised light. The segmentation is better in unpolarised light even in the presence of specularities, due to stronger degree of spontaneous

polarisation which results in stronger discrimination in the coefficient features. The results shown represent a sample of the segmentation results obtained. Additional materials have been studied and these include plastic and natural leaves, military camouflage net, fresh and rotten fruits.

5 Conclusions

In this paper we have explored the use of polarisation information for surface segmentation and classification. Our idea is to parameterise the polarisation image using spherical harmonic co-efficients, and to use the co-efficients as a means of characterising surface properties. The characterisation has proved effective both in locating damaged regions of skin for fruit, and for segmenting scenes into regions of different material composition.

References

1. Atkinson, G., Hancock, E.R.: Recovery of Surface Orientation from Diffuse Polarization. IEEE TIP 15, 1653–1664 (2006)
2. Atkinson, G., Hancock, E.R.: Two-dimensional BRDF Estimation from Polarisation. Computer Vision and Image Understanding 111, 126–141 (2008)
3. Jones, B.F., Fairney, P.T.: Recognition of shiny dielectric objects by analyzing the polarization of reflected light. IVCJ 7 (1989)
4. Born, M., Wolf, E.: Principles of Optics, 7th edn. Cambridge University Press, Cambridge (1999)
5. Miyazaki, D., Kagesawa, M., Ikeuchi, K.: Transparent Surface Modeling from a Pair of Polarization Images. IEEE TPAMI 26, 73–82 (2004)
6. Saupe, D., Vranić, D.V.: 3-D Model Retrieval with Spherical Harmonics and Moments. In: Proceedings of the 23rd DAGM-Symposium on Pattern Recognition, pp. 392–397 (2001)
7. Hecht, E.: Optics, 4th edn. Addison-Wesley, Reading (2002)
8. Wolff, L.B.: Polarization-Based Material Calssification from Specular Reflection. IEEE TPAMI 12, 1059–1071 (1990)
9. Wolff, L.B., Boult, T.E.: Constraining Object Features using a Polarisation Reflectance Model. IEEE TPAMI 13, 635–657 (1991)
10. Morel, O., Stolz, C., Meriaudeau, F., Gorria, P.: Active lighting applied to three-dimensional reconstruction of specular metallic surfaces by polarization imaging. Applied Optics 45, 4062–4068 (2006)
11. Shen, L., Ford, J., Makedon, F., Saykin, A.: A Surface-based Approach for Classification of 3D Neuroanatomic Structures. Intelligent Data Analysis 8, 519–545 (2004)
12. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE TPAMI 22, 888–905 (2000)
13. Wolff, L.B.: Polarization vision: a new sensory approach to image understanding. In: IVC, vol. 15, pp. 81–93 (1997)

Reflection Component Separation Using Statistical Analysis and Polarisation

Lichi Zhang, Edwin R. Hancock*, and Gary A. Atkinson

Department of Computer Science, University of York, UK
Machine Vision Laboratory, University of the West of England, UK

Abstract. We show how to cast the problem of specularity subtraction as blind source separation from polarisation images. We commence by summarizing the relationships between the specular and diffuse reflection components for polarised images. We show how to use singular value decomposition for component separation. In particular, we show how reliable results can be obtained using three images acquired with different polariser angles under diffuse reflection. The proposed method can be used as the preprocessing step in shape from shading, segmentation, reflectance estimation and many other computer vision applications.

1 Introduction

Reflectance estimation is a key problem in computer vision and graphics. It has found widespread use in areas such as surface inspection and object rendering, with application in both the video game and film industries. Any light which is incident on a surface, will undergo two reflections which are specular and diffuse [11]. When performing reflectance estimation from images, the process can be simplified if the two components are separated beforehand. One way of doing so is to follow Ragheb and Hancock [12] and to fit a statistical mixture of specular and diffuse distributions across the object surface using shape-from-shading. Lin and Lee [7], on the other hand, separate the diffuse and specular components using photometric stereo, whilst simultaneously recovering surface height. In common with Ragheb and Hancock, they also use the Cook-Torrance model for reflectance model fitting, but eventually opt for the Lafortune model for reasons of greater flexibility when dealing with multiple reflection components. Wu and Tang [17] have extended this method to deal not only with specular and diffuse components, but also with a subsurface scattering component.

A more direct way to separate the reflectance components is to use polarisation images. Underpinning this approach is Fresnel theory, which explains the way polarised light interacts with surfaces. For dielectrics, the analysis is straightforward, but for metals the situation is less tractable due to the induction of surface currents by the time varying electromagnetic field. For dielectrics, polarisation may arise in two different ways. In the case of specular polarisation, initially polarised light is reflected in the specular direction. For diffuse polarisation, initially unpolarised light is refracted into the surface and the re-emitted light acquires a spontaneous polarisation. In both cases

* Edwin Hancock is supported by a Royal Society Wolfson Research Merit Award.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 476–483, 2011.

the zenith angle of the reflected or re-emitted light is constrained by the degree of polarisation, and the azimuth angle is constrained by the phase angle. Such results have been used to develop a number of machine vision techniques including methods for surface quality inspection [10] and surface shape recovery [1],[9],[13].

Polarisation can also be applied to infer information concerning the reflectance properties of surfaces. For instance, Atkinson and Hancock show [2] how diffuse polarisation can be used to estimate the bidirectional reflectance function. However, their method is computationally demanding, using simulated annealing to estimate the BRDF. There have been several attempts to use polarisation information for reflectance analysis. Nayar et al. [11] extend the dichromatic reflection model to incorporate both color and polarisation. Ma et al. [8] use structured light (four spherical gradient illumination patterns) to estimate surface normal direction, and show how polarised light can be used to estimate the diffuse and specular normal maps independently.

Here we aim to use statistical methods to separate the diffuse and specular components of reflection. Specifically, we pose the problem as Blind Source Separation (BSS). Stated succinctly, we aim to extract the underlying source signal from a set of linear mixtures, where the mixing matrix is unknown [6]. The technique of BSS has found applications in the removal of reflections from transparent glass surfaces. Examples include the work of Fraid and Adelson [4], which presents a simple method for solving the problems. Bronstein et al. [3] use sparse ICA to perform the separation and achieve better results at the expense of a more complicated algorithm, and Umeyama and Goldin [14] extend the method of Bronstein et al. using a single polariser. However, it cannot be used in the case where the light source is polarised, and the knowledge of phase angle is not considered, which is an important factor in the polarisation vision. Here, we extend the work in [14] and introduce a new method which addresses the above issues, and can perform the separation in a much faster way. We experimentally verify that our method gives good results for both polarised and unpolarised source illumination.

2 Polarisation

One disadvantage of many existing methods that apply polarisation information to reflectance estimation is that a sequence of images with varying polariser angle must be obtained in order to achieve robust estimation of the mean intensity, degree of polarisation and phase [16]. As there are three unknown variables, three images obtained with different polarisation orientations should be sufficient to perform the estimation. However, in practice more images are needed to achieve robustness against noise or experimental systematics [11]. The problem is exacerbated when the object under study is not static. For example, [15] introduces a method to separate the two reflectance components using only three photos taken in three different polarisation orientations as 0, 45 and 90 degrees. However, such method usually cannot be used in practice as a significant amount of noise is present, which drastically degrades results.

Here, we introduce the method to separate the reflectance components using Blind Source Separation and the theory of polarisation, which can tolerate reasonable amounts of noise from the images. The method can complete the separation without the

information of polariser angles. This is an improvement from the standard method in the past, and it can be applied to the experiments when using non-calibrated polarisation filters.

When scattered light is measured through a linear polarizing filter, the intensity changes as the polariser angle θ is rotated. The measured intensity follows the transmitted radiance sinusoid (TRS) given by:

$$I(\theta) = \frac{(I_{\max} + I_{\min})}{2} + \frac{(I_{\max} - I_{\min})}{2}\cos(2\theta - 2\phi)\ . \tag{1}$$

where $I_{\max}$ is the maximum brightness, $I_{\min}$ is the minimum, and ϕ is the phase angle, that corresponds to the angle of maximum transmission. In [11], part of the specular component, denoted as I_{sv}, is changed with the polarisation angles, while as the rest remains constant I_{sc}. Following the definition in [11] where the diffuse component is $I_d = I_{\min}$ and assuming the specular component $I_s = I_{\max} - I_{\min} = 2I_{sc} = 2I_{sv}$, (2) can be rewritten as

$$I(\theta) = I_d + \frac{I_s}{2} + \frac{I_s}{2}\cos 2\theta \cos 2\phi + \frac{I_s}{2}\sin 2\theta \sin 2\phi\ . \tag{2}$$

Our method requires three $M \times N$ images captured under different polariser orientations θ_1,θ_2 and θ_3. Each image is converted into a long-vector of length MN, and the observation matrix $\boldsymbol{X}$ is denoted as $(\boldsymbol{x}_1|\boldsymbol{x}_2|\boldsymbol{x}_3)$, where x_i is the long-vector by stacking the columns of the i-th polarisation image. We capture the two-components reflectance process using the matrix

$$\boldsymbol{C} = [(J_d + \frac{1}{2}J_s), (\frac{1}{2}J_s\cos 2\phi), (\frac{1}{2}J_s \sin 2\phi)] = (C_a|C_b|C_c)\ . \tag{3}$$

In the above, J_d and J_s are long-vectors of length MN that contain the diffuse and specular reflectance components I_d and I_s as elements. With these ingredients we can relate the observed polarised image data to the two component reflectance model via:

$$\boldsymbol{X} = \boldsymbol{C}\boldsymbol{A}^T\ ,\ \boldsymbol{A} = \begin{bmatrix} 1 & \cos 2\theta_1 & \sin 2\theta_1 \\ 1 & \cos 2\theta_2 & \sin 2\theta_2 \\ 1 & \cos 2\theta_3 & \sin 2\theta_3 \end{bmatrix}\ . \tag{4}$$

By solving this equation, we can estimate the elements of $\boldsymbol{C}$ and hence recover the phase angle together with separated reflectance components using $\phi = \frac{1}{2}\tan^{-1}(\frac{C_c}{C_b})$, $J_s = 2\sqrt{C_b^2 + C_c^2}$ and $J_d = C_a - 2\sqrt{C_b^2 + C_c^2}$.

3 Blind Source Separation

We commence by applying Singular Vector Decomposition (SVD) to the data matrix $\boldsymbol{X}$ which contains the stacked polarisation images. When performing component analysis, it is normal to center the data matrix. However, here we can not perform this operation since it will distort the diffuse reflectance component. Also, as [14] shows, using SVD for component separation without whitening remains valid according to experimental results. The SVD equation of the data matrix $\boldsymbol{X}$ gives $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{P}\boldsymbol{V}^T$, where

$\boldsymbol{U}$ is the $MN \times 3$ left eigenvector matrix, $\boldsymbol{D}$ the 3×3 diagonal matrix of singular values, and $\boldsymbol{V}$ the 3×3 right eigenvector matrix.

To simplify the equation we let $\boldsymbol{P} = \boldsymbol{U}\ \boldsymbol{D}$. To recover the matrix $\boldsymbol{C}$, we define $\boldsymbol{P} = k\boldsymbol{C}\boldsymbol{W}^{-1}$, and $\boldsymbol{V} = k^{-1}\boldsymbol{A}\boldsymbol{W}^{T}$, where k is a scaling parameter and $\boldsymbol{W}$ is a 3×3 weighting matrix satisfying the constraint $|\boldsymbol{W}| = 1$. To calculate the scaling parameter k we note that $C_a = (J_d + \frac{1}{2}J_s)$ is close to $E(i) = \frac{1}{3}[x_1(i) + x_2(i) + x_3(i)]$, which is the mean intensity at pixel i from $\boldsymbol{X}$. We define a region of interest, L, which will exclude all background areas of the image. Then we have a simple expression for our scaling factor:

$$k = \frac{\sum_{i \in L} E(i)}{\sum_{i \in L} C_a(i)}\ . \tag{5}$$

All that remains is to determine $\boldsymbol{W}$. Since $|\boldsymbol{W}| = 1$, we can have

$$|\boldsymbol{W}^T| = |\boldsymbol{W}| = k^{-1}\frac{|\boldsymbol{A}|}{|\boldsymbol{V}|} = 1\ . \tag{6}$$

In the above, the values of the elements of $\boldsymbol{P}$ are given by SVD and the unknown elements of the matrix $\boldsymbol{A}$ are determined by the values of θ_1, θ_2 and θ_3. However, from (6) if estimates of θ_1 and θ_2 are to hand, then the constraint that $|\boldsymbol{W}| = 1$ determines the value of θ_3. We may calculate the two angles θ_1 and θ_2 by exhaustive search [14], which will give a globally optimal solution, but is also time-consuming. Here instead we use Newton's method for estimating $\boldsymbol{W}$, because of its rapid (quadratic) convergence.

4 Iteration Process

We use Newton's method to estimate θ_1 and θ_2. The iteration process is initialised so that the two angles are determined by assuming $\boldsymbol{A}^{(0)} = k\boldsymbol{V}$, which we have

$$\theta_1^{(0)} = \frac{1}{2}\arcsin\left(\frac{\boldsymbol{V}_{(1,2)}}{\sqrt{\boldsymbol{V}_{(1,2)}^2 + \boldsymbol{V}_{(1,3)}^2}}\right),\ \theta_2^{(0)} = \frac{1}{2}\arcsin\left(\frac{\boldsymbol{V}_{(2,2)}}{\sqrt{\boldsymbol{V}_{(2,2)}^2 + \boldsymbol{V}_{(2,3)}^2}}\right)\ . \tag{7}$$

and then we use (6) to compute $\theta_3^{(0)}$. With the three angles to hand, an initial estimate of the matrix $\boldsymbol{A}$, i.e. $\boldsymbol{A}^{(0)}$, so we can have $\boldsymbol{C}^{(0)}$ of the matrix $\boldsymbol{C}$ from (4), which yields values for the diffuse and specular components $J_s^{(0)}$, $J_d^{(0)}$ together with the phase angle $\phi^{(0)}$ from (3).

To apply the Newton method, we require a measure of error. Here we use a smoothness criterion based on the local variance of the diffuse component of intensity. Firstly, we select a set of pixels within the region $R^{(0)}$, where the intensities in the specular component $J_s^{(0)}$ are higher than its mean value. That is $i \in R^{(0)}$, if $J_s^{(0)}(i) \geq \bar{J}_s^{(0)}$, $i = 1, ..., MN$, where $\bar{J}_s^{(0)} = \frac{1}{MN}\sum_{i=1}^{MN} J_s^{(0)}(i)$. Our error criterion is based on the variance of the intensities of the diffuse component over the image region $R^{(0)}$. For iteration t, we therefore have

$$\varepsilon^{(t)} = \sum_{i \in R^{(t)}} [J_d^{(t)}(i) - \bar{J}_d{}^{(t)}]^2\ . \tag{8}$$

where $\bar{J_d}^{(t)}$ is the arithmetic mean of $J_d^{(t)}(i)$ for all $i \in R^{(t)}$. When we have correct separation, the intensities of the diffuse component in that region should become approximately uniform. This means that there is neither specularity nor shading information present in $R^{(t)}$, and ε reaches a minimum. The Newton method for updating the two angles is

$$\Theta^{(t+1)} = \Theta^{(t)} - \gamma \boldsymbol{H}[\varepsilon^{(t)}]^{-1} \nabla \varepsilon^{(t)} \ . \tag{9}$$

where $\Theta^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})^T$, $\boldsymbol{H}[\varepsilon^{(t)}]$ is the Hessian of the error-function and $\nabla\varepsilon$ its gradient.

5 Separation Results

Here we present separation results for our method. Firstly, we have tested the method on three objects illuminated by polarised light from a collimated light source in the direction of the camera (frontal illumination). The objects studied are an apple, a plastic ball and a pear. We placed one polariser in front of the camera during acquisition, and for each object we collected images with polarisation angles of 0, 60 and 120 degrees. Settings from the camera have been tested, so that the resultant photos can avoid overexposure and other problems that might affect the accuracy of the results. The collimated light source is large and so is placed behind and a little above the camera, preventing any light from being partly blocked by the camera. The Image quality is set to be the highest, so that the fine details of the reflectances can be maintained. Before starting the separation process, all the images are eliminated of any background contents, their object locations are fully aligned, and their colors are converted to grey level.

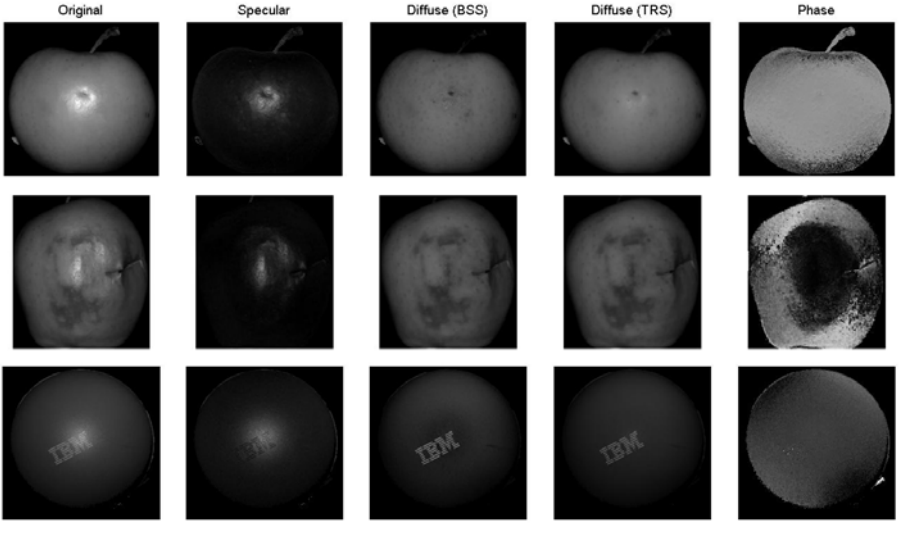


Fig. 1. Results for selected objects under polarised incident light. Columns from left-to-right, example input image (0 degree polariser), diffuse, specular and phase-angle. Images in third and fourth columns are the comparison of diffuse component between our method and the standard method. The image contrast has been adjusted to improve clarity.

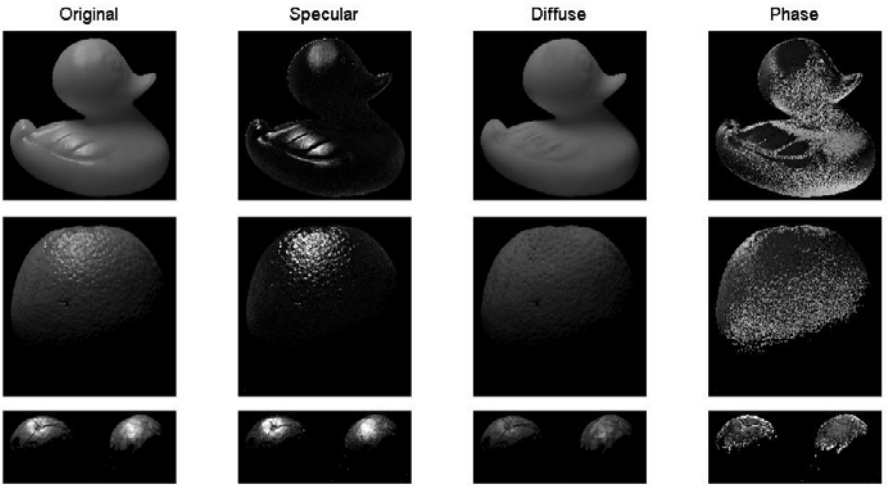


Fig. 2. Separation results for objects under unpolarised incident light at non-retroreflective illumination

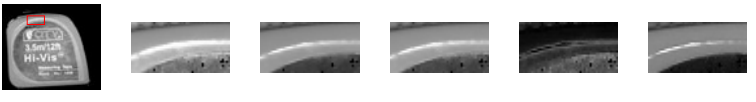


Fig. 3. Example for the result where the separation result is not satisfactory. The first image is the original photo, and images from the second to the fourth are the details of specularities from photos taken in 0, 60 and 120 polarisation angles, while as the fifth and sixth images are the estimated result for specular and diffuse components, respectively.

The results of this experiment are presented in Fig.1. The first column shows the image obtained with the incident light polariser at an angle of 0 degrees (vertical), the specular and diffuse separations are shown in the second, third and fourth columns, and the phase angles in the fifth column. The standard method uses the TRS equation (2) (put a citation here to confirm precisely which method is being used) to perform the separation, which requires knowledge of both the input images and corresponding polariser angles.

Comparing the results of our method to that of the standard one, it can be seen that some specularity remained on the surface of apple using TRS fitting. Using our method by contrast, the shininess has been fully eliminated. The results clearly demonstrate therefore, that using our method the knowledge of polariser angle is not required, while as the result of separation remains quite satisfactory. Next we present the results for objects imaged using unpolarised light, where the polariser detects diffuse polarisation. Fig.2 shows the results with the same column ordering as above. As before the three camera polariser angles are 0, 60 and 120 degrees. The objects studied are a plastic duck, a nut and an orange. The light source is located above the objects, so part of the image surface is in shadow. However, it is clear that the separation process still works well.

It is noted that, while our method performs well on relatively matte surfaces such as fruit or plastic, it does not produce satisfactory results on very shiny materials, such as what is shown in Fig.3. This is attributable to camera saturation. Use of TRS fitting and other polarisation separation methods also fail to address this issue. There are two solutions to this problem. We can either reduce the camera exposure time or adjust the polariser angles to emphasis the variation in highlight intensity. Alternatively, if colour information is used then the dichromatic reflection model can be used to improve the separation of the specular component [11], and this is a current research goal.

Finally, we explore the application of our method to color images. Here we have investigated samples of human skin. The color images are decomposed into red, green and blue channels, and the separation process is applied to each channel in turn. The specular and diffuse components in the different channels are then recombined to give composite separations [5]. Fig.4 shows the results obtained, with the usual column ordering. The specular separation is good, removing shininess from the surface of the skin.

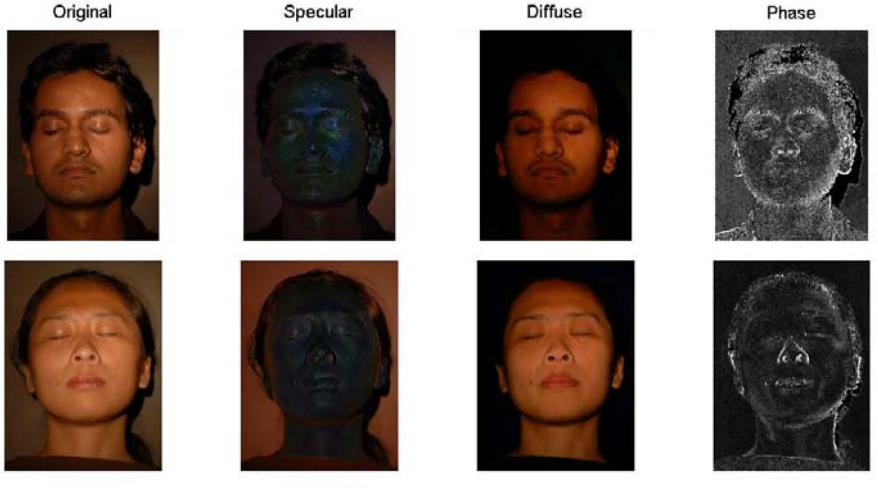


Fig. 4. Example results for color images

6 Conclusion

In this paper we have shown how Blind Source Separation can be applied to the separation of diffuse and specular reflection components using polarised images. The new method is proved to be effective. It can be used as the preprocessing steps in many applications of polarisation vision, such as shape reconstruction and surface reflectance estimation. Future research will aim to exploit the method in these two domains.

References

1. Atkinson, G., Hancock, E.: Shape estimation using polarization and shading from two views. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(11), 2001–2017 (2007)
2. Atkinson, G., Hancock, E.: Two-dimensional BRDF estimation from polarisation. Computer Vision and Image Understanding 111(2), 126–141 (2008)
3. Bronstein, A., Bronstein, M., Zibulevsky, M., Zeevi, Y.: Sparse ICA for blind separation of transmitted and reflected images. International Journal of Imaging Systems and Technology 15(1), 84–91 (2005)
4. Farid, H., Adelson, E.: Separating reflections and lighting using independent components analysis. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 262–267 (1999)
5. Forsyth, D., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall, New York (2003)
6. Kisilev, P., Zibulevsky, M., Zeevi, Y., Pearlmutter, B.: Multiresolution framework for blind source separation. Technical Report CCIT 317 (June 2001)
7. Lin, S., Lee, S.: Estimation of diffuse and specular appearance. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 855–861 (1999)
8. Ma, W., Hawkins, T., Peers, P., Chabert, C., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In: Rendering Techniques 2007: 18th Eurographics Workshop on Rendering, pp. 183–194 (June 2007)
9. Miyazaki, D., Kagesawa, M., Ikeuchi, K.: Transparent surface modeling from a pair of polarization images. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(1), 73–82 (2004)
10. Morel, O., Meriaudeau, F., Stolz, C., Gorria, P.: Polarization imaging applied to 3D reconstruction of specular metallic surfaces. In: Electronic Imaging, vol. 5679, pp. 178–186 (2005)
11. Nayar, S., Fang, X., Boult, T.: Separation of Reflection Components using Color and Polarization. International Journal of Computer Vision 21(3), 163–186 (1997)
12. Ragheb, H., Hancock, E.: Highlight removal using shape-from-shading. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2351, pp. 626–641. Springer, Heidelberg (2002)
13. Rahmann, S., Canterakis, N.: Reconstruction of specular surfaces using polarization imaging. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 149–156 (2001)
14. Umeyama, S., Godin, G.: Separation of Diffuse and Specular Components of Surface Reflection by Use of Polarization and Statistical Analysis of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5), 639–647 (2004)
15. Wolff, L.: Polarization vision: a new sensory approach to image understanding. Image and Vision Computing 15(2), 81–93 (1997)
16. Wolff, L., Boult, T.: Constraining object features using a polarization reflectance model. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 635–657 (1991)
17. Wu, T.-P., Tang, C.-K.: Separating specular, diffuse, and subsurface scattering reflectances from photometric images. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3022, pp. 419–433. Springer, Heidelberg (2004)

Characterizing Graphs Using Approximate von Neumann Entropy

Lin Han, Edwin R. Hancock, and Richard C. Wilson

Department of Computer Science, University of York, YO10 5GH, UK

Abstract. In this paper we show how to approximate the von Neumann entropy associated with the Laplacian eigenspectrum of graphs and exploit it as a characteristic for the clustering and classification of graphs. We commence from the von Neumann entropy and approximate it by replacing the Shannon entropy by its quadratic counterpart. We then show how the quadratic entropy can be expressed in terms of a series of permutation invariant traces. This leads to a simple approximate form for the entropy in terms of the elements of the adjacency matrix which can be evaluated in quadratic time. We use this approximate expression for the entropy as a unary characteristic for graph clustering. Experiments on real world data illustrate the effectiveness of the method.

1 Introduction

One of the key problems that arises in the analysis on non-vectorial pattern data such as strings, trees and graphs is how to characterize such data for the purposes of clustering and classification. Unlike pattern vectors, when the analysis of tree or graph data is attempted then there is frequently no labelling or ordering of the nodes of the structure to hand.

Broadly speaking, there are three ways by which to overcome this problem. The first is to extract characteristics from the graph or tree data to-hand, and then to cluster graphs on the basis of vectors of structural characteristics [4]. The second method is to use a measure of pairwise distance between structures and resort to pairwise clustering methods [8]. The third method involves constructing a class prototype through the union or intersection of different structures[17] [11][12]. These latter two methods can prove very time consuming and even fragile since they require reliable node correspondences to hand [15][16], and this invariably requires inexact graph matching over the dataset to hand.

It is for this reason that the use of graph characteristics has proved to be an attractive one. Although there are a number of simple alternatives that can be used, such as node or edge frequency, edge density, diameter and perimeter, these have proved to be ineffective as a means of characterizing variations in intrinsic structure. Instead, it has proved necessary to resort to more complex representations. One of the most successful of these has been to use graph-spectral methods [13][14]. Here the distribution of the eigenvalues and eigenvectors can be used to construct permutation invariants that do not require node correspondences. Examples here include Laplacian spectra and characteristic polynomials. This

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 484–491, 2011.

study has recently been taken one step further by Xiao, Wilson and Hancock [4] who have performed an analysis of the heat kernel for graphs, and have shown that the Riemann zeta function can be used to generate a number of powerful invariants from the normalized Laplacian spectrum. One interesting conclusion of this work was that the gradient of the zeta function at the origin yields a unary characteristic which can be used to cluster quite complex data-sets of graphs. Another route to a unary characterization of graph structure is to define measures of intrinsic complexity. The characterization of graph complexity is a long standing problem, but recently measures based on the heat kernel have proved effective, and these include the use of Birkoff polytopes [10] and heat-flow complexity [9].

Unfortunately, both graph-spectral and heat flow complexity methods can prove computationally burdensome. The reason for this is that the computation of the graph-spectrum is cubic in the number of nodes. Our aim in this paper is to explore whether more efficient complexity characterizations are to hand and whether they can compete with the unary characterization provided by the Riemann zeta function.

Our approach is as follows. We commence from the von Neumann entropy of a graph. This is simply the Shannon entropy associated with the spectrum of the normalized Laplacian matrix. We explore how to simplify and approximate the calculation of von Neumann entropy. Our first step is to replace the Shannon entropy by its quadratic counterpart. An analysis of the quadratic entropy reveals that it can be computed from a number of permutation invariant matrix trace expressions. This leads to a simple expression for the approximate entropy in terms of the elements of the adjacency matrix, and which can be computed without evaluating the normalized adjacency matrix. The expression is quadratic in the number of nodes in a graph. Moreover, we illustrate experimentally that it outperforms the gradient of the Riemann zeta function as a unary attribute.

2 Graph Representation and the von Neumann Entropy

To commence, we denote the graph under study by $G = (V, E)$ where V is the set of nodes and $E \subseteq V \times V$ is the set of edges. Further, we represent the structure of the graph using a $|V| \times |V|$ adjacency matrix whose elements are

$$A(u,v) = \begin{cases} 1 & \text{if } (u,v) \in E \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{1}$$

The degree matrix of graph G is a diagonal matrix D whose elements are given by $D(u,u) = d_u = \sum_{v \in V} A(u,v)$. From the degree matrix and the adjacency matrix we can construct the Laplacian matrix $L = D - A$, i.e. the degree matrix minus the adjacency matrix. The elements of the Laplacian matrix are

$$L(u,v) = \begin{cases} d_v & \text{if } u = v \ , \\ -1 & \text{if } (u,v) \in E \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{2}$$

The normalized Laplacian matrix is given by $\hat{L} = D^{-1/2}LD^{-1/2}$ and has elements

$$\hat{L}(u,v) = \begin{cases} 1 & \text{if } u = v \text{ and } d_v \neq 0 \ , \\ -\frac{1}{\sqrt{d_u d_v}} & \text{if } (u,v) \in E \ , \\ 0 & \text{otherwise .} \end{cases} \tag{3}$$

The spectral decomposition of the normalized Laplacian matrix is $\hat{L} = \Phi\Lambda\Phi^T$ where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_{|V|})$ is a diagonal matrix with the ordered eigenvalues as elements $(0 = \lambda_1 < \lambda_2 < ... < \lambda_{|V|})$ and $\Phi = (\phi_1|\phi_2|...|\phi_{|V|})$ is a matrix with the corresponding ordered unit-norm eigenvectors as columns. The normalized Laplacian matrix is positive semi-definite and so has all eigenvalues non-negative. The number of zero eigenvalues is the number of connected components in the graph. For a connected graph, there is only one eigenvalue which is equal to zero. The normalization factor means that the largest eigenvalue is less than or equal to 2, with equality only when G is bipartite. Hence all the eigenvalues of the normalize Laplacian matrix are in the range $0 \leq \lambda \leq 2$. The normalized Laplacian matrix is commonly used as a graph representation and the eigenvector ϕ_2 associated with the smallest non-zero eigenvalues λ_2 referred to as the Fiedler-vector [1] is often used in graph cuts [2][3].

The von Neumann entropy of the graph associated with the Laplacian eigenspectrum is defined as [6]

$$H = -\sum_{i=1}^{|V|} \frac{\lambda_i}{2} \ln \frac{\lambda_i}{2} \ . \tag{4}$$

We approximate the entropy $-\frac{\lambda_i}{2} \ln \frac{\lambda_i}{2}$ by the quadratic entropy $\frac{\lambda_i}{2}(1 - \frac{\lambda_i}{2})$, to obtain

$$H = -\sum_i \frac{\lambda_i}{2} \ln \frac{\lambda_i}{2} \simeq \sum_i \frac{\lambda_i}{2}(1 - \frac{\lambda_i}{2}) = \frac{\sum_i \lambda_i}{2} - \frac{\sum_i \lambda_i^2}{4} \ . \tag{5}$$

Using the fact that $Tr[\hat{L}^n] = \sum_i \lambda_i^n$, the quadratic entropy can be rewritten as

$$H = \frac{Tr[\hat{L}]}{2} - \frac{Tr[\hat{L}^2]}{4} \ . \tag{6}$$

Since the normalized Laplacian matrix $\hat{L}$ is symmetric and it has unit diagonal elements, then according to equation (3) for the trace of the normalized Laplacian matrix, we have

$$Tr[\hat{L}] = |V| \ . \tag{7}$$

Similarly, for the trace of the square of the normalized Laplacian, we have

$$\begin{aligned} Tr[\hat{L}^2] &= \sum_{u \in V} \sum_{v \in V} \hat{L}_{uv} \hat{L}_{vu} = \sum_{u \in V} \sum_{v \in V} (\hat{L}_{uv})^2 \\ &= \sum_{\substack{u,v \in V \\ u=v}} (\hat{L}_{uv})^2 + \sum_{\substack{u,v \in V \\ u \neq v}} (\hat{L}_{uv})^2 \\ &= |V| + \sum_{(u,v) \in E} \frac{1}{d_u d_v} \ . \end{aligned} \tag{8}$$

Substituting Equation (7) and (8) into Equation (6), the entropy becomes

$$H = \frac{|V|}{2} - \frac{|V|}{4} - \sum_{(u,v)\in E} \frac{1}{4\ d_u d_v} = \frac{|V|}{4} - \sum_{(u,v)\in E} \frac{1}{4\ d_u d_v}\ . \tag{9}$$

As a result, we can approximate the von Neumann entropy using two measures of graph structure. The first is the number of nodes of the graph, while the second is the degree of the nodes of the graph. The approximation bypasses calculating the Laplacian eigenvalues of a graph to estimate its von Neumann entropy.

3 Riemann Zeta Function Derivative

Before we proceed to the experimental evaluation of the approximate entropy, we explain how the unary representation based on the analysis of the Riemann zeta function arises. The Riemann zeta function associated with Laplacian eigenvalues is defined to be

$$\zeta(s) = \sum_{\lambda_i \neq 0} \lambda_i^{-s}\ . \tag{10}$$

which is the result of exponentiating and summing the reciprocal of the non-zero Laplacian eigenvalues.

According to [4], the zeta function is related to the Mellin moment of the heat kernel trace, i.e. the sum of the diagonal elements of the heat kernel matrix of the graph

$$Tr[h_t] = \sum_{i=1}^{|V|} \exp[-\lambda_i t]\ . \tag{11}$$

where $h_t = e^{-t\hat{L}}$ is the heat kernel and t is time. The heat kernel can be viewed as describing the flow of the information across the edges of the graphs with time and the rate of flow is determined by the Laplacian of the graph.

By making use of the Mellin transform, i.e.

$$\int_0^\infty t^{s-1} e^{-\lambda_i t} dt = \Gamma(s)(-\lambda_i)^{-s}\ . \tag{12}$$

where $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$, we can establish the link between the zeta function and the trace of the heat kernel trace, i.e.

$$\zeta(s) = \sum_{\lambda_i \neq 0} \lambda_i^{-s} = \frac{1}{\Gamma(s)} \int_0^\infty (Tr[h_t] - C) t^{s-1} dt\ . \tag{13}$$

The zeta function is also linked to the determinant of the Laplacian. To show this, we rewrite the zeta function in terms of the natural exponential with the result

$$\zeta(s) = \sum_{\lambda_i \neq 0} \lambda_i^{-s} = \sum_{\lambda_i \neq 0} \exp[-s \ln \lambda_i]\ . \tag{14}$$

The derivative of the zeta function is given by

$$\zeta'(s) = \sum_{\lambda_i \neq 0} \{-\ln \lambda_i\} \exp[-s \ln \lambda_i] \ . \tag{15}$$

At the origin the derivative takes on the value

$$\zeta'(0) = \sum_{\lambda_i \neq 0} \{-\ln \lambda_i\} = \ln\{\prod_{\lambda_i \neq 0} (\frac{1}{\lambda_i})\} \ . \tag{16}$$

Mckay [5] has shown that the derivative of the zeta function at the origin is link to the number of spanning trees in a graph G through

$$\tau(G) = \frac{\prod_{u \in V} d_u}{\sum_{u \in V} d_u} \exp[-\zeta'(0)] \ . \tag{17}$$

As a result, the derivative of the zeta function at the origin is determined by the number of spanning trees in the graph together with the degree of its vertices.

4 Experiments

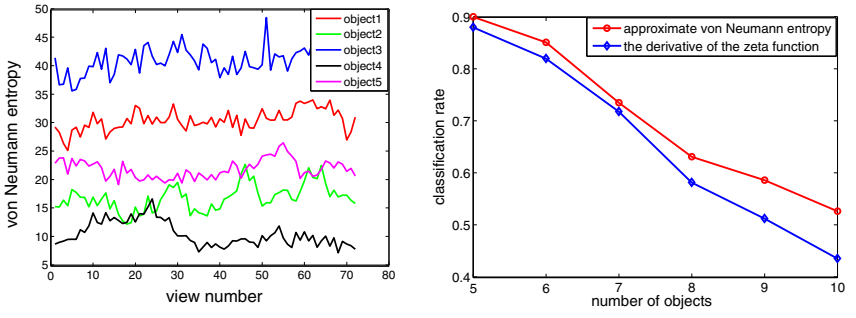
In this section, we provide some experimental evaluation of the approximate expression for the von Neumann entropy on a real-word dataset. The dataset used is the COIL[7] which consists of images of 10 objects, with 72 views of each object from equally spaced directions over 360°. We extract corner features from each image and use the detected feature points as nodes to construct sample graphs by Delaunay triangulation. Example images of the objects are given in Figure 1.

Fig. 1. Example images of the objects in COIL dataset

We commerce by illustrating the behavior of the approximate expression for the von Neumann entropy. To do this, we select 5 objects from the COIL dataset and plot the approximate von Neumann entropy of their Delaunay graphs as a function of view number (1 to 72) in Figure 2(a). In the plot, different curves are for different objects and the x-axis is the view number of the image from which the relevant Delaunay graph was extracted. From the figure it is clear

that although the individual values of the entropy fluctuate for each object, the curves for the different object clusters can be separated and do not overlap.

To further evaluate the use of the approximate von Neumann entropy as a graph characteristic on an object classification task, we apply a K-nearest neighbor classifier to the approximate von Neumann entropy of the Delaunay graphs of the objects in COIL dataset and observe how the classification rate changes as we increase the number of objects to be classified. For comparison, we have also investigated the result obtained using the derivative of Riemann zeta function at the origin (as outlined in section 3). Figure 2(b) shows the variation of the classification rates for the two graph characteristics. In our experiment, we set k to 3 and the classification rate is the average fraction of graphs that are correctly identified, computed using 10-fold cross-validation. From the plot, it is clear that the approximate von Neumann entropy measure (red line) always achieves a higher classification rate than the derivative of zeta function at the origin (blue line). Moreover, the classification rate of the latter decreases faster as we increase the number of objects. However, the classification rates for both characterization methods decay to around 50% when the number of objects used increases to 10, and the attributes become overlapped.



(a) (b)

Fig. 2. (a) The value of the approximate von Neumann entropy for five objects. (b) The comparison of classification rates of the two graph characteristics.

We now turn our attention to analyzing how well the approximation of the von Neumann entropy holds. Recall that in section 2 we show there is a link between the value of the von Neumann entropy of a graph and the number of nodes in the graph together with the node degree. This link is realized by approximating the Shannon entropy $-\frac{\lambda_i}{2}\ln\frac{\lambda_i}{2}$ by the quadratic entropy $\frac{\lambda_i}{2}(1-\frac{\lambda_i}{2})$. Here we explore how the distribution of the Laplacian eigenvalues affects the approximation. To do this, we experiment with both a synthetic eigenvalue dataset and the eigenvalues of the Delaunay graphs from the COIL dataset. The eigenvalues in the synthetic dataset are sampled from a uniform distribution between 0 and 2. We select at random different sets of eigenvalues and compute the von Neumann entropy together with its quadratic approximation. The two plots on the upper row of Figure 3 respectively show the uniform distribution of

eigenvalues together with a scatter plot of the von Neumann entropy (y-axis) versus its quadratic approximation (x-axis) for the synthetic eigenvalue distribution. Figure 3(c) shows the distribution of the eigenvalues of the normalized Laplacian matrix of the graphs from the COIL dataset. Here we observe that there is a peak around 1.4 and most eigenvalues are in the range [1,1.5]. Based on this distribution, the scatter plot of von Neumann entropy versus the quadratic approximation gives the linear pattern in Figure 3(d). The reason for this is that for the eigenvalues in the peak, there is a linear relationship between the two entropies. Moreover, the distribution of colors along the line (which indicate the object classes) indicate that the object separation is very good.

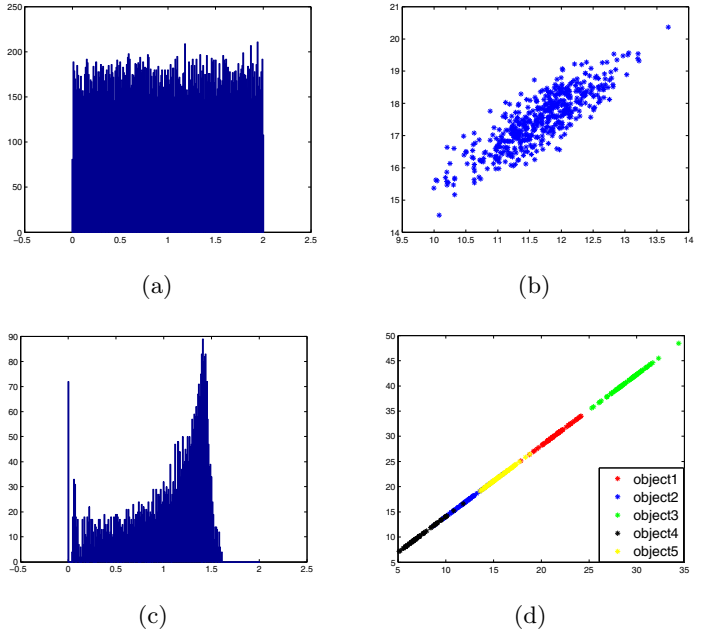


Fig. 3. The distribution of the eigenvalues and the approximation of the von Neumann entropy for the synthetic dataset and COIL dataset

5 Conclusion

In this paper we show how to use of the von Neumann entropy computed from the Laplacian eigenspectrum to characterize graphs. We approximate the Shannon term in the definition of the von Neumann in a quadratic manner. This approximation leads to an expression for the von Neumann entropy in terms of the number of nodes and node degrees. We experiment with the proposed von Neumann entropy measure on an object classification task, and show it outperforms the derivative of the zeta function at the origin (which is also a very successful unary attribute of graphs). Experimental results also reveal that there is a linear relationship between the von Neumann entropy and its approximation for Delaunay graphs.

References

1. Chung, F.R.K.: Spectral Graph Theory. American Mathematical Society, Providence (1997)
2. Robles-Kelly, A., Hancock, E.R.: A Riemannian Approach to Graph Embedding. Pattern Recognition, 1042–1056 (2007)
3. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. In: Computer Vison and Pattern Recognition, pp. 731–737 (1997)
4. Xiao, B., Hancock, E.R., Wilson, R.C.: Graph Characteristic from the Heat Kernel Trace. Pattern Recognition 42, 2589–2606 (2009)
5. McKay, B.D.: Spanning Trees in regular Graphs. Eur. J. Combin. 4, 149–160
6. Passerini, F., Severini, S.: The von Neumann entropy of networks. arXiv:0812.2597 (2008)
7. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (coil100). Technical Report, Department of Computer Science, Columbia University (1996)
8. Torsello, A., Robles-Kelly, A., Hancock, E.R.: Discovering Shape Classes using Tree Edit-Distance and Pairwise Clustering. IJCV 72(3), 259–285 (2007)
9. Escolano, F., Lozano, M.A., Hancock, E.R., Giorgi, D.: What is the complexity of a network? The heat flow-thermodynamic depth approach. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR&SPR 2010. LNCS, vol. 6218, pp. 286–295. Springer, Heidelberg (2010)
10. Escolano, F., Hancock, E.R., Lozano, M.A.: Polytopal graph complexity, matrix permanents, and embedding. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 237–246. Springer, Heidelberg (2008)
11. Torsello, A., Hancock, E.R.: Graph Embedding Using Tree Edit-union. Pattern Recognition 40(5), 1393–1405 (2007)
12. Suau, P., Escolano, F.: Bayesian Optimization of the Scale Saliency Filter. Image Vision Comput. 26(9), 1207–1218 (2008)
13. Luo, B., Wilson, R.C., Hancock, E.R.: A Spectral Approach to Learning Structural Variations in Graphs. Pattern Recognition 39(6), 1188–1198 (2006)
14. Wilson, R.C., Hancock, E.R., Luo, B.: Pattern Vectors from Algebraic Graph Theory. IEEE Trans. Pattern Anal. Mach. Intell. 27(7), 1112–1124 (2005)
15. Ferrer, M., Valveny, E., Serratosa, F., Bunke, H.: Exact median graph computation via graph embedding. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 15–24. Springer, Heidelberg (2008)
16. Ferrer, M., Serratosa, F., Valveny, E.: On the relation between the median and the maximum common subgraph of a set of graphs. In: Escolano, F., Vento, M. (eds.) GbRPR. LNCS, vol. 4538, pp. 351–360. Springer, Heidelberg (2007)
17. Riesen, K., Neuhaus, M., Bunke, H.: Graph embedding in vector spaces by means of prototype selection. In: Escolano, F., Vento, M. (eds.) GbRPR. LNCS, vol. 4538, pp. 383–393. Springer, Heidelberg (2007)

A Distance for Partially Labeled Trees

Jorge Calvo, David Rizo, and José M. Iñesta

Dept. Lenguajes y Sistemas Informáticos, Universidad de Alicante,
E-03080 Alicante, Spain
jcz4@alu.ua.es, {drizo,inesta}@dlsi.ua.es

Abstract. In a number of practical situations, data have structure and the relations among its component parts need to be coded with suitable data models. Trees are usually utilized for representing data for which hierarchical relations can be defined. This is the case in a number of fields like image analysis, natural language processing, protein structure, or music retrieval, to name a few. In those cases, procedures for comparing trees are very relevant. An approximate tree edit distance algorithm has been introduced for working with trees labeled only at the leaves. In this paper, it has been applied to handwritten character recognition, providing accuracies comparable to those by the most comprehensive search method, being as efficient as the fastest.

Keywords: Tree edit distance, approximate distances, qtrees.

1 Introduction

In contrast to scalar or vectors for representing measures, data exhibit inner structures by nature in many applictions. Trees are able to code hierarchical relations in their structure in a natural way and they have been utilized in many tasks, like text document analysis [10], protein structure [14], image representation and coding [3], or music analysis and retrieval [12], to name just a few.

The computation of a measure of the similarity between two trees is a subject of interest in these areas. Different approaches have been posed in order to perform this comparison. Some of them impose a restriction on how the comparison is performed, while others establish valid mappings. Some methods pay more attention on the tree structure, and others pay it on the content of the nodes and the leaves. Most of them are designed to work with fully labeled trees.

Recently, we have proposed an algorithm designed to work with partially labeled trees, more precisely with those labeled only at the leaves [11]. This feature, focus more on the coded content and the relations within its context. One of the fields where this situation is relevant is music comparison and retrieval. Trees have been used for this task and a number of representation and comparison schemes have been applied based on tree edit distances [12] or probabilistic similarity schemes [5].

In any case, the computation of these measures is usually a time consuming task and different authors have proposed algorithms that are able to compute

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 492–499, 2011.

them in a reasonable time [16], through approximated versions of the similarity measure. In this paper, a new algorithm is presented, able to deal with trees labeled only at the leaves that runs in $O(|T_1| \times |T_2|)$ time.

2 Tree Comparison Methods

In general, trees can be divided in *orderered* and *not ordered trees*, and in *evolutionary* and *not evolutionary* trees. In this work only ordered trees are considered. Regarding the *evolutionary trees*, they are often used to conceptually represent the evolutionary relationship of species or organisms in biology, evolution of works in linguistics, statistical classifications, or even tracking computer viruses. An *evolutionary tree* can be defined as a tree with distinct labels at leaves. Several algorithms have been given to solve the comparison of evolutionary trees [13,2,9]. In this work we deal with trees that have not distinct labels, and thus they are not *evolutionary trees*, so those algorithms are not applicable for our problem.

A number of similarity measures for ordered non-evolutionary trees have been defined in the literature [1]. Some of them measure the sequence of operations needed to transform one tree in another one, others look for the longest common path from the root to a tree node. There are methods that allow wildcards in the matching process in the so-called *variable-length doesn't care* (VLDC) distance. Several taxonomies have been proposed. The interested reader can look up a hierarchy of tree edit distances in [8] and [18].

Table 1. Some tree edit and alignment distance algorithms and their time complexities

Tai [17]	$O(\|T_1\| \times \|T_2\| \times \text{depth}(T_1)^2 \times \text{depth}(T_2)^2)$
Shasha & Zhang [19]	$O(\|T_1\| \times \|T_2\| \times \min\{\text{depth}(T_1), \|\text{leaves}(T_1)\|\} \times \min\{\text{depth}(T_2), \|\text{leaves}(T_2)\|\})$
Jiang [7]	$O(\|T_1\| \times \|T_2\| \times (\text{rank}(T_1) + \text{rank}(T_2))^2)$
Selkow [16]	$O(\|T_1\| \times \|T_2\|)$
Valiente [18]	$O(\|T_1\| \times \|T_2\| \times \log(\|T_1\| + \|T_2\|))$

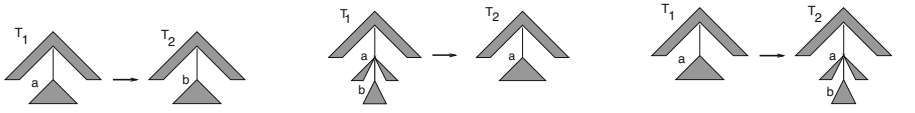
Some of the most relevant tree edit and alignment distances have been compiled in Table 1 together with their time complexities. These measures are the ones utilized for comparison in this paper, except that of Tai due to its very high complexity.

The classical edit distance between two trees $d(T_1, T_2)$ is computed using an edit script $e = e_1 \cdots e_n$ that is a sequence of edit operations allowing the transformation of a tree T_1 to a tree T_2. The cost of an edit script $C(e)$ is the sum of the costs of the edit operations involved in the script: $C(e) = \sum_{i=1}^{n} c(e_i)$. Given $S(T_1, T_2)$, the set of all the scripts that enable the emission of T_2 given T_1, the edit distance between T_1 and T_2 is defined by: $d(T_1, T_2) = \min_{e \in S(T_1,T_2)} C(e)$. The edit operations allowed over two trees are any of the following:

- *relabel* the label l of a node v of T_1 by the label l' of another node w of T_2, denoted by (v, w) (Fig. 1a).

– *deletion* of a non-root node v from T_1, denoted by (v, λ), consists of deleting it, making the children of v become the children of its parent node, in the position that was occupied by v, preserving this way the left to right ordering of leaves (Fig. 1b).
– *insertion* of a non-root node w in T_2, denoted by denoted by (λ, w). Given a sequence $w_i \cdots w_j$ of subtrees of a common parent w, the insertion of node w' makes those $w_i \cdots w_j$ subtrees children of w', and w' child of w (Fig. 1c).

Note that the operation (λ, λ) is not allowed.



(a) Substitute operation (b) Delete operation (c) Insert operation

Fig. 1. Tree edit operations (from [6])

The edit cost of each operation, $c(e_i)$ is given based on that of the edit cost of the symbols for the labels, $c\colon \Sigma \times \Sigma \to \mathbb{R}$ that depends on the particular application. Therefore, $c(e_i)$ denotes the cost of applying the edit operation (v, w).

The approaches by Selkow [16], Valiente [18], and Jiang [7] are restricted version of this general methodology.

3 Proposed Tree Comparison Algorithm

The edit distances presented above are designed to work with fully labeled trees. In order to apply those algorithms to trees labeled only at the leaves, the non-labeled inner nodes can be assigned a special label "empty". However, it is expected they don't work as well as they do with fully labeled trees.

In order to overcome this situation two approaches are possible. The first one consists of labeling all nodes using any bottom-up propagation scheme using specific knowledge of the application domain. The main drawback of that option is that any intermediate process will condition the resulting trees, with a loss of generality. The second approach is to design a distance function able to compare partially labeled trees.

The *partially labeled tree comparison algorithm* (s_p) is based on the assumption that the similarity value between a labeled leaf and a non-labeled inner node should be the average of the chances of finding that leaf in the descendants of that inner node. Fig. 2a shows the simplest case of having two leaf trees: $s_p(T_A, T_B) = \delta(a, b)$, where $\delta(a, b) = 1 \iff a = b$, and 0 elsewhere. For comparing the trees shown in Fig. 2b, the chances of finding the label a in T_B are computed as $s_p(T_A, T_B) = (\delta(a, x) + \delta(a, y))/2$. If, instead of a label, y another tree is placed there, the function should be computed recursively. Finally, when none of the trees is composed by a single leaf (Fig. 2c), the similarity function is computed like an edit distance between sequences wx and yz, where each symbol is a tree.

This measuring method omits the accounting of the insertion or deletion of nodes and only measures the chance of finding matching labels, giving more importance to the information hierarchically contained in the tree than to the tree structure.

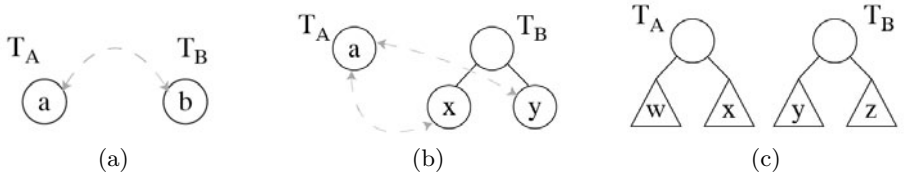


Fig. 2. Similarity function s_p representative cases

This method is designed for working with partially labeled trees, but we can slightly adapt the original idea to work with fully labeled trees. The case of comparing a leaf to a non-leaf tree (Fig. 3a) is computed as $s_p(T_A, T_B) = (\delta(a,b) + \delta(a,x) + \delta(a,y))/3$. And in the same way as in the case of non-labeled nodes, the similarity $s_p(T_A, T_B)$ between two fully labelled trees (Fig. 3b) is computed as the edit distance between the sequences wx and yz, where each symbol is a tree, plus now the similarity between the labels a and b.

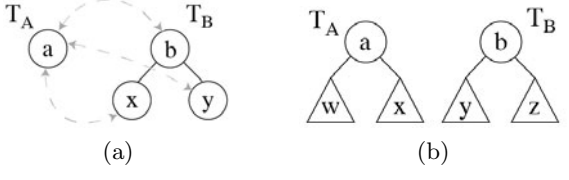


Fig. 3. Similarity function s_p working on fully labelled trees

The algorithmic details of this method can be found at [11]. In that paper it is also shown that the time complexity of the proposed algorithm is $O(|T_1| \times |T_2|)$, like that fastest of the methods cited in Table 1.

4 Experiments and Results

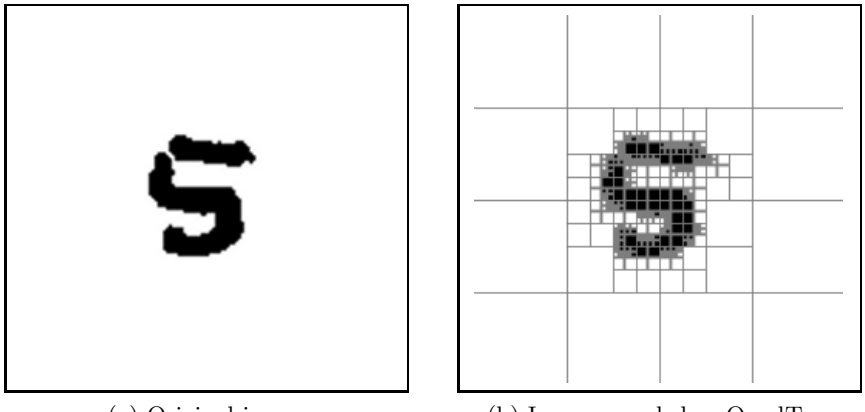
The experiments are devised to show that the proposed tree distance is able to provide good results in a reasonable time when compared to other classical tree comparison algorithms. For that, we need to select an application where the data can be described in terms of trees.

In general, images can be coded as QuadTrees [15]. In a QuadTree, each inner node has four children. Each leaf represents a region in the image, labeled with any of its properties. If the considered property in the region covered by that leaf is not homogeneous enough (the deviation of their pixel values are over a threshold), the leaf is converted to a inner node and 4 leaves are created, splitting

the region into 4 sub-regions. This procedure is repeated recursively until al the regions are homogeneous. At this point the tree is built, containing labels only at the leaves.

In particular, for binary images, the alphabet for the labels is $\Sigma = \{0, 1\}$. That is the case for the images of the isolated hand-written character images that have been used and coded as QuadTrees (Fig. 4) for our experiments. The problem to solve is to identify a character or digit from a set made by different writers.

It should be pointed out that this evaluation is not oriented to outperform the state of the art in hand-written OCR, but to check whether the proposed tree edit algorithm is useful for solving tree-encoded structure comparisons efficiently. In order to this, its performance has been compared to those of other tree distances, not to other handwritten OCR methods.



(a) Original image (b) Image encoded as QuadTree

Fig. 4. Encoding an image as a QuadTree

Corpora. The NIST image gallery [4] was utilized. To code the character images with QuadTrees the usual procedure has been followed, splitting recursively the image in four parts until they reach a level where all the pixels in each new part have the same value (foreground or background). That value will be placed only in the leaves.

Two corpora have been used in our experimental setup. The first consists of a set of 19,540 digits, the same number for each one. The second corpus consists of a set of 50,400 letter-and-digits images, with the same number for each class again.

Character classification accuracy. The experiments have been performed using the following scheme. The corpora have been split into 10 folds. Then, given each fold, for each class, a prototype has been taken as a query, being compared to all prototypes in the fold, obtaining a list L of prototypes sorted by similarity value, L_k being the k-th position in the list.

The accuracy of the system has been computed using the *mean reciprocal rank* (MRR). Let q be a query, p a prototype, and let the function $c(p)$ return the class of the prototype p, the reciprocal rank RR for a query is computed as:

$$RR(q) = \frac{1}{\arg\min_k \{c(L_k(q)) = c(q)\}} \tag{1}$$

The denominator should be read as the minimum position in the list, k, so that the predicted class matches that of the query.

The *mean reciprocal rank* is the mean of all $RR(q)$ for all the queries.

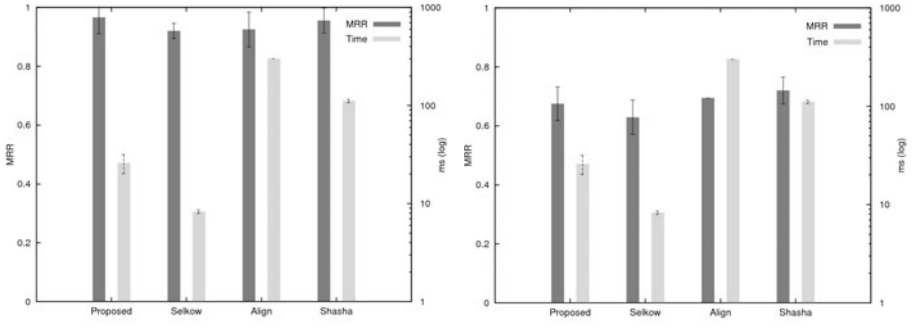
Running times were measured in milliseconds, taking into account only the test phase, leaving aside the construction of the representations that may be done off-line. All experiments were performed using a Sun machine with 8 Gb RAM and 8 Intel(R) Xeon(R) CPU X5355 running at 2.66GHz, with a SUSE Linux with kernel version 2.6.

4.1 Results

The achieved results are shown in Fig. 5. For each tree comparison method, the mean reciprocal rank (MRR) and processing time have been computed. Each magnitude has been represented in a different y-axis (MMR left, time right). Dark gray bars represent MRR and light gray correspond to times. Fig. 5(a) displays the results obtained using only digits, while Fig. 5(b) plots those obtained with the alphanumeric data.

The results show that the proposed algorithm (left-most bars in each graph) achieves success rated comparable to the best algorithms but keeping processing times comparable to the faster method (note that the time axis is logarithmic). With the digits corpus it performed the best and with the alphanumeric one it reached MRR values that where comparable to those achieved by the best one, with no significant differences.

Thus, it seems that the proposed algorithm is able to compare hand-written character images encoded as QuadTree better than the other classical tree comparison algorithms in terms of trade-off between time and success rate.



(a) Digits corpus (b) Alphanumeric corpus

Fig. 5. Success rates of proposed method compared to classical tree edit distances

5 Conclusions

An approximate tree edit distance algorithm has been introduced for working with trees labeled only at the leaves. In order to assess its performance, it has been applied to handwritten character recognition using the well-known NIST database. Our aim was not to outperform the state of the art but to show that the proposed tree distance is able to provide good results in a reasonable time when compared to other tree distances.

The performance has been compared to classical tree similarity algorithms from the literature. The results show that, in the used context, the proposed algorithm achieved accuracies comparable to those by the most comprehensive search method and it is as efficient as the fastest.

Acknowledgements

This work is supported by the Spanish Ministry projects DRIMS (TIN2009-14247-C02), and Consolider Ingenio 2010 (MIPRCV, CSD2007-00018), partially supported by EU ERDF. The authors want to thank Javier Gallego for his help in providing and pre-processing the data.

References

1. Bille, P.: A survey on tree edit distance and related problems. Theorical Computer Science 337(1-3), 217–239 (2005)
2. DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J., Zhang, L.: On distances between phylogenetic trees. In: SODA 1997: Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 427–436. Society for Industrial and Applied Mathematics, Philadelphia (1997)
3. Finkel, R.A., Bentley, J.L.: Quad trees: A data structure for retrieval on composite keys. Acta Inf. 4, 1–9 (1974)
4. Garris, M.D., Wilkinson, R.A.: Nist special database 3: Handwritten segmented characters. NIST, Gaithersburg, Md
5. Habrard, A., Iñesta, J.M., Rizo, D., Sebban, M.: Melody recognition with learned edit distances. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 86–96. Springer, Heidelberg (2008)
6. Isert, C.: The editing distance between trees (1999)
7. Jiang, T., Wang, L., Zhang, K.: Alignment of trees – an alternative to tree edit. Theoretical Computer Science 143(1), 137–148 (1995)
8. Kuboyama, T., Shin, K., Miyahara, T.: A hierarchy of tree edit distance measures. Theoretical Computer Science and its Applications (2005)
9. Lee, C.-M., Hung, L.-J., Chang, M.-S., Shen, C.-B., Tang, C.-Y.: An improved algorithm for the maximum agreement subtree problem. Information Processing Letters 94(5), 211–216 (2005)
10. Marcus, M., Kim, G., Marcinkiewicz, M.A., Macintyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The penn treebank: Annotating predicate argument structure. In: ARPA Human Language Technology Workshop, pp. 114–119 (1994)

11. Rizo, D., Iñesta, J.M.: New partially labelled tree similarity measure: a case study. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR&SPR 2010. LNCS, vol. 6218, pp. 296–305. Springer, Heidelberg (2010)
12. Rizo, D., Lemström, K., Iñesta, J.M.: Tree representation in combined polyphonic music comparison. In: Ystad, S., Kronland-Martinet, R., Jensen, K. (eds.) CMMR 2008. LNCS, vol. 5493, pp. 177–195. Springer, Heidelberg (2009)
13. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences 53(1-2), 131–147 (1981)
14. Russell, R.B., Barton, G.J.: Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. Proteins: Structure, Function, and Bioinformatics 14, 309–323 (2004)
15. Samet, H.: The quadtree and related hierarchical data structures. ACM Comput. Surv. 16(2), 187–260 (1984)
16. Selkow, S.M.: The tree-to-tree editing problem. Information Processing Letters 6(6), 184–186 (1977)
17. Tai, K.-C.: The tree-to-tree correction problem. J. ACM 26(3), 422–433 (1979)
18. Valiente, G.: An efficient bottom-up distance between trees. In: International Symposium on String Processing and Information Retrieval, pp. 212–219. IEEE Computer Society, Los Alamitos (2001)
19. Zhang, K., Shasha, D.: Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput. 18(6), 1245–1262 (1989)

An Online Metric Learning Approach through Margin Maximization

Adrian Perez-Suay*, Francesc J. Ferri, and Jesús V. Albert

Departament d'Informàtica. Universitat de València
{adrian.perez,francesc.ferri,jesus.v.albert}@uv.es

Abstract. This work introduces a method based on learning similarity measures between pairs of objects in any representation space that allows to develop convenient recognition algorithms. The problem is formulated through margin maximization over distance values so that it can discriminate between similar (intra-class) and dissimilar (inter-class) elements without enforcing positive definiteness of the metric matrix as in most competing approaches. A passive-aggressive approach has been adopted to carry out the corresponding optimization procedure. The proposed approach has been empirically compared to state of the art metric learning on several publicly available databases showing its potential both in terms of performance and computation results.

Keywords: Metric Learning, Dimensionality Reduction, Classification, Nearest Neighbor, Online Learning, Passive-Aggressive.

1 Introduction

The problem of classifying and/or conveniently representing sets of data is of key importance in different fields such as pattern recognition, data mining and image analysis in their different application domains and, for example in image retrieval (CBIR). This problem is particularly critical when the objects under study are characterized by very high dimensional descriptors. The classical approach to deal with these problems lies in the application of some form of dimensionality reduction in order to search numerical stability, improved performance or to obtain appropriate recognition results in a reasonable amount of time [1, 2].

Dimensionality reduction has been studied from different points of view. In particular, linear methods like Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) are very well known and are used very often in practice [1–3]. Moreover, most linear approaches to reduce the dimensionality can be extended to the nonlinear case by the familiar kernel trick [4].

In recent years [5–9], it has been highlighted the use of quadratic distances, $d^M(x,y) = (x-y)^T M(x-y)$ between pairs of points $x, y \in \ ^D$ where M is a positive (semi)definite matrix (PSD). The (squared) Euclidean distance constitutes a particular case of such distance.

* Work partially funded by FEDER and Spanish and Valencian Governments through projects TIN2009-14205-C04-03, ACOMP/2010/287, GV/2010/086 and Consolider Ingenio 2010 CSD07-00018.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 500–507, 2011.

Such distances have been widely used also to find a transformation into a linear subspace in which the original data can be conveniently represented. In particular, certain methods are based on the general idea of grouping the points of space as compact clusters [5, 7].

Most approaches proposed to date explicitly put the restriction of dealing with PSD matrices by explicitly tackling the corresponding constraints or approximating it [5–10] or by formulating the problem in such a way that the constraint is implicitly taken into account [9]. In the present work, the PSD constraint is not included in the formulation and this can lead to non-Euclidean distances as in other previous works as [11].

The formulation presented in this work consist of a margin maximization problem applied to non Euclidean distances between pairs of training points. Instead of a direct (batch) solution to this quadratic problem, an online approach using a passive-aggressive algorithm is considered. The performance of the algorithm and the computational burden it implies is studied and assessed in the experimentation carried out.

2 Maximizing the Margin on Distances

We start with a set of n labelled points $X = \{(x_i, c_i)\}_{i=1}^{n}$ where c_i indicates the class label associated with $x_i \in \ ^{D}$ and $c_i \in \{1, \ldots, c\}$, where c is the total number of classes.

Let us to consider a quadratic distance parametrized by a symmetric matrix $M \in \ ^{D \times D}$

$$d_{ij}^{M} = d^{M}(x_i, x_j) = (x_i - x_j)^{T} M (x_i - x_j) \tag{1}$$

In our particular case we will only require the matrix $M \in \ ^{D \times D}$ be symmetric which may give rise to a degenerate metric (zero eigenvalues of M) or even to a non Euclidean metric (negative eigenvalues of M). Even so, values of the "distance" could in principle be interpreted as measures of similarity that will be low or high (as when comparing objects of the same or different classes) regardless of whether or not these values are negative.

Our goal is to learn the matrix M from the set of samples X in order to appropriately discriminate between the distances to points in the same class (intraclass) and the distances between points of different classes (interclass).

A way to achieve this goal in its simplest version is to require that the intra- and inter-class distances belong to different ranges of the similarity measure. In other words, we can define a margin of separation between the values these take. This idea can be expressed by the following constraints, in terms of a parameter b that refers to the center of the margin.

$$\begin{aligned} d_{ij}^{M} &< b - 1, \text{ if } c_i = c_j \\ d_{ij}^{M} &\geq b + 1, \text{ if } c_i \neq c_j \end{aligned} \tag{2}$$

The objective is to assign distances intra-class values less than $b-1$ and inter-class distance values greater than $b+1$ in order to discriminate between pairs of similar and dissimilar points respectively. Obviously, the fact that the distances are negative only plays a marginal role in this formulation and simply introduces more degrees of freedom. In fact, a convenient and logical solution to this problem will imply that only few distances will effectively achieve negative values. The value of b will be determined by the solution to the problem.

Introducing indicator variables y_{ij}, inequalities (2) can be written for $i, j = 1, \ldots, n$ in a more compact way as

$$y_{ij}(d_{ij}^M - b) \geq 1 \tag{3}$$

where $y_{ij} = -1$, if $c_i = c_j$ and $y_{ij} = 1$, if $c_i \neq c_j$. In other words, low values of the distance are sought for pairs of points in the same class while higher values are meant for different class pairs. In particular, an ideal margin of 2 is enforced between these two ranges of values [10].

Under this formulation the problem of finding a generic metric matrix M can be solved and tackled in a similar way as with Support Vector Machines [10, 12]. In particular, it can be formulated as the minimization of the squared Frobenius norm [10] of matrix M subject to the above conditions (3) appropriately modified by adding the corresponding slack variables ξ_{ij}.

$$\begin{aligned} &\min_{M,b,\xi} \frac{1}{2}\|M\|_{Fro}^2 + C\sum_{k=1}^{N}\xi_{ij} \\ &\text{s.t.} \quad y_{ij}((x_i - x_j)^T M(x_i - x_j) - b) \geq 1 - \xi_{ij} \\ &\qquad \xi_{ij} \geq 0, \quad i, j = 1, \ldots, n \end{aligned} \tag{4}$$

This formulation leads to a quadratic dual problem which admits a solution that is functionally equivalent to the one corresponding to SVM. A very important difference is that in our case, $\mathcal{O}(n^2)$ restrictions are obtained which makes the problem much harder.

3 A Passive-Aggressive Approach to Distance Margin Maximization

In order to solve the above problem in a convenient way both from the point of view of computation and robustness, an online learning approach has been considered. In particular, a passive-aggressive approach has been chosen. This family of learning algorithms was introduced by Crammer [13] to solve SVM-like problems and has been successfully used to solve large scale learning problems [14] to learn a similarity function which closely relates to metric learning. Starting from the inequality (3) and introducing the hinge loss function

$$l_M(x_i, x_j, b) = \max\{0, 1 - y_{ij}((x_i - x_j)^T M(x_i - x_j) - b)\}, \tag{5}$$

we can set our goal as the minimization of a global loss, L_M, that accumulates hinge losses (5) over all possible pairs of points x_i, x_j in the training set. This can be written as

$$L_M = \sum_{i<j} l_M(x_i, x_j, b) \tag{6}$$

Instead of a global criterion as in the previous formulation, we consider a passive-aggressive approach in which only a constraint (a pair) of points at a time is taken into account and in which the matrix M (and the threshold b) is progressively modified at each iteration. First, M and b are initialized to some initial values M^0, b^0. Then, at each training iteration k, we select a unique pair of instances x_i, x_j, and solve the following soft margin optimization problem:

$$\begin{aligned}(M^k, b^k) = &\arg\min_{M,b} \frac{1}{2}\|M - M^{k-1}\|^2_{Fro} + \frac{1}{2}(b - b^{k-1})^2 + C\xi_{ij} \\ &s.t.\ \ l_M(x_i, x_j, b) \leq \xi_{ij}, \quad \xi_{ij} \geq 0\end{aligned} \tag{7}$$

By introducing two Lagrange multipliers, τ, λ, and equating to zero partial derivatives of the Lagrangian as in [14], it is possible to obtain the following updating rules

$$M^k = M^{k-1} + \tau y_{ij}(x_i - x_j)(x_i - x_j)^T \tag{8}$$

$$b^k = b^{k-1} - \tau y_{ij} \tag{9}$$

where

$$\tau = \min\left\{C, \frac{l_{M^{k-1}}(x_i, x_j, b^{k-1})}{1 + \|(x_i - x_j)(x_i - x_j)^T\|^2_{Fro}}\right\}. \tag{10}$$

Equations 8, 9 and 10 summarize the final updating that must be applied at each iteration k. The corresponding online algorithm will start from initial values of M and b and will select (randomly in principle) pairs of points at each iteration in order to apply the above update on M and b. This method can be extended to the nonlinear case with the well known kernel trick.

Although the final matrix M that solves the problem (7) is not PSD in general, it is always possible to find the closest PSD matrix (with respect to the standard Frobenius norm) [5, 6, 8, 10]. This is equivalent to projecting the matrix on the called PSD cone in the space of matrices $^{D\times D}$. In particular, the problem: $\arg\min_{A\succeq 0} \|M-A\|^2_F = \sum_{i=1}^{D} \max\{\lambda_i, 0\} v_i v_i^T$ where $\{\lambda_i\}_{i=1}^{D}$ are the eigenvalues of matrix M in decreasing order and $\{v_i\}_{i=1}^{D}$ represents the set of associated eigenvectors. As a sub-product, it can be obtained in this case a linear projection to a lower dimension, m, since we can define matrices W from only the largest eigenvalues. $\hat{M} = \sum_{i=1}^{m} \max\{\lambda_i, 0\} v_i v_i^T = W^T W$. In this case, the matrix $\hat{M}$ is merely an approximation to the solution to the problem (7). This solution will be more or less appropriate depending on the value of m.

In our proposal, instead of enforcing positive semidefiniteness at each iteration and inspired by the work in [14] we chose to force this only after convergence of the proposed online algorithm.

4 Empirical Evaluation

A number of experiments has been performed in order to empirically assess the proposed method. In this work, a comparative experimentation taking into account an Information Theoretic Metric Learning (ITML) is presented. ITML can be considered an state of the art approach that has been compared to many other approaches recently [5, 9, 15]. In this work, ITML only is considered under an experimental setting similar to [9].

Comparative experimentation has been carried out and results have been collected using $k-$nearest neighbor ($k-$NN) classification. Experimentation has been kept as realistic as possible and an empirical setup as close as possible to the one presented in [9] has been fixed. In particular, algorithms have been assessed using the 4-nearest neighbor classification via two-fold cross validation over 5 runs. Binomial confidence intervals are given at the 95%.

Parameter C in the proposed method was tuned over a set of different values logarithmically distributed in the range 10^{-5} to 10^{-1}. Tuning has been done by taking 50% of train to train and the remaining 50% to evaluate the 4-nearest neighbor error rate.

In the experiments considered, only a fraction of all possible pairs have been considered to train the algorithms. In the particular case of ITML and according to [9], only $20c^2$ pairs are given to the algorithm. In other words, the number of iterations of the algorithm is bounded in any case by $20c^2$. In the case of the proposed algorithm, a fixed number of iterations equal to 20% of the total number of pairs has been considered.

To study the behavior of the method we based the experimentation on classification over public databases. In particular we have selected six databases from the UCI repository (Wine, Ionosphere, Balance, Iris, Soybean Big, Soybean Small) and the database Nist16 (from PRTOOLS [16]).

Finally, when proposed algorithms have been trained, we use the respective matrices obtained to transform the data into a Euclidean subspace on which we apply the classification explained above. Results presented about ITML have been obtained using the code made available by the authors in [17].

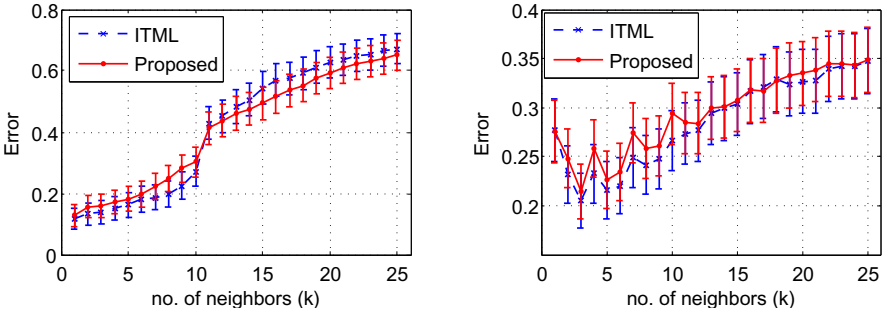


Fig. 1. $k-$NN Error rate obtained with the two methods proposed (left) Soybean Big, (right) Balance

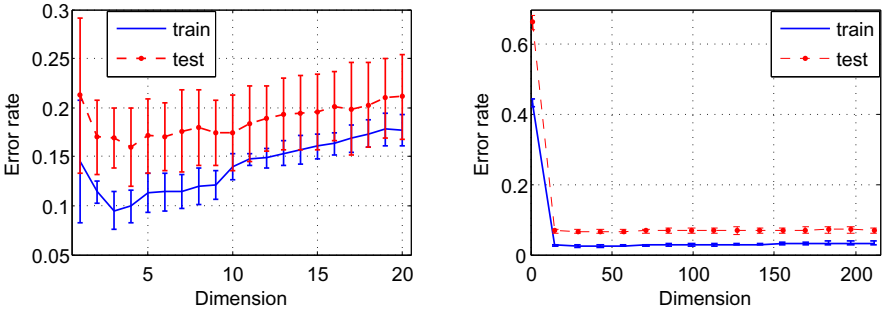


Fig. 2. Error rate obtained with the proposed method in terms of final dimension kept from matrix M. (Left) Ionosphere database, (right) Nist16 database.

To fully show the behavior of the algorithms, we have computed classification error rates for increasing values of k in the $k-$NN classifier. In particular, Figure 1 shows this curve for the particular databases Soybean Big and Balance. As it can be seen in this figure, both algorithms obtain very similar results across all range of k values. This behavior is kept in the other databases considered in this work. Consequently, it can be said that both the proposed approach and the ITML algorithm give equally good results on a wide range of problems. Furthermore, these results are as good as or better than the ones that can be expected from other metric learning algorithms that have been previously compared to ITML [9]. The same conclusions can be arrived at with regard to the behavior of the algorithms when the final dimension of the linear projections obtained are progressively decreased. In Figure 2 the behavior of the 4-NN classification rule after projecting the data using the proposed algorithm in terms of final dimensionality is shown. Again, results using the corresponding projections obtained from ITML (not shown) give not significantly different rates. As can be seen from the figure, both training and test results exhibit a very close behavior (even closer in the Ionosphere database as with most other small databases). The general conclusion is that dimension can be safely decreased to very small values (depending on databases) without a significant decrease in performance. This is true both for the proposed algorithm as with ITML even taking into account that the matrix obtained from the proposed approach has been first projected onto the PSD cone.

Even more interesting than these preliminary classification results are the behavior of the algorithms with regard to computational burden. Table 1 shows CPU time spent by both algorithms on learning the corresponding matrices (including tunning which has been done exactly in the same way). Standard deviations are also shown in the same table. As can be observed, the proposed algorithm needs only a small fraction of the time spent by ITML to obtain fully comparable results. This can be considered a very interesting result taking into account that ITML is considered as a fast algorithm in comparison to other metric learning algorithms [9]. In order to illustrate this difference in time in a more

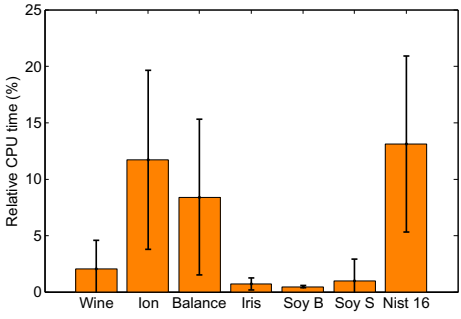


Fig. 3. Relative averaged CPU time execution

Table 1. CPU time obtained when using the two learning algorithms considered on all databases(in seconds) with standard deviations (in parentheses)

	Wine	Ionosphere	Balance	Iris	Soybean B	Soybean S	Nist16
Proposed	0.04(0.01)	0.42(0.09)	0.27(0.02)	0.01(0.00)	0.24(0.00)	0.08(0.01)	263(00)
ITML	2.12(2.49)	3.59(2.30)	3.29(2.67)	2.47(1.94)	52.27(9.25)	8.55(16.06)	2005(1190)

graphical way, Figure 3 displays relative averaged CPU times on all databases considered in this work. Reduction percentages of about 1% are observed, going up to about 15% for some of the databases including the largest ones considered.

5 Conclusions

A new metric learning method based based on discrimination between distances using margins has been presented. In particular, a passive-aggressive algorithm has been used to achieve the proposed goal. An empirical evaluation has been presented comparing our method with a state of the art method. Comparable results in performance have been obtained when considering classification tasks on public databases.

With regard to computational burden, the proposed approach outperforms the competing algorithm. All together, the proposed approach constitutes a very challenging option as there is still much room to improve it both in performance, robustness and even execution time by controlling the convergence of the algorithm. Further work is directed towards these improvements. In particular, reformulation of the online learning algorithm in such a way that positive definiteness is weighted (but not enforced) is being considered.

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience, Hoboken (2000)
2. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell. 22, 4–37 (2000)

3. Fukunaga, K.: Introduction to statistical pattern recognition, 2nd edn. Academic Press Professional, Inc., San Diego (1990)
4. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
5. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Neural Information Processing Systems (NIPS 2005), vol. 18, pp. 451–458 (2005)
6. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems, vol. 17, pp. 513–520. MIT Press, Cambridge (2004)
7. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems 15, vol. 15, pp. 505–512 (2002)
8. Shalev-Shwartz, S., Singer, Y., Ng, A.Y.: Online and batch learning of pseudometrics. In: ICML 2004: Proceedings of the Twenty-first International Conference on Machine Learning, p. 94. ACM, New York (2004)
9. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML 2007: Proceedings of the 24th International Conference on Machine Learning, pp. 209–216. ACM, New York (2007)
10. Nguyen, N., Guo, Y.: Metric learning: A support vector approach. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 125–136. Springer, Heidelberg (2008)
11. Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence). World Scientific Publishing Co., Inc., River Edge (2005)
12. Vapnik, V.N.: Statistical Learning Theory. Wiley Interscience, Hoboken (1998)
13. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. Journal of Machine Learning Research 7, 551–585 (2006)
14. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research 11, 1109–1135 (2010)
15. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research 10, 207–244 (2009)
16. Duin, R., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D.: Prtools4, a matlab toolbox for pattern recognition (2004)
17. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information Theoretic Metric Learning. UT, Austin (2007), `http://www.cs.utexas.edu/users/pjain/itml/`

Graph Matching on a Low-Cost and Parallel Architecture*

David Rodenas, Francesc Serratosa, and Albert Solé

Universitat Rovira i Virgili, Departament d'Enginyeria Informàtica i Matemàtiques
david.rodenas@gispert-rodenas.com,
{francesc.serratosa,albert.sole}@urv.cat

Abstract. This paper presents a new parallel algorithm to compute the graph-matching based on the Graduated Assignment. The aim of this paper is to perform graph matching in a current desktop computer, but, instead of executing the code in the generic processor, we execute a parallel code in the graphic processor unit. This computer can be embedded in a low-cost pattern recognition system or a mobile robot. Experiments show a speed-up of the run time around 400 times, which makes the use of attributed graphs to represent objects a valid solution.

1 Introduction

Classification is a task of pattern recognition that attempts to assign each input value to one of a given set of classes. Pattern recognition algorithms generally aim to provide a reasonable answer for all possible inputs and to do inexact matching of inputs. Pattern recognition is studied in many fields such as psychology, cognitive science, computer science and so on. Depending on the application, inputs of the pattern recognition model or objects to be classified are described by different representations. The most usual representation is a set of real values but other common ones are strings, trees or graphs. These structures have more capacity to capture the knowledge of the model but their comparison or matching is also more computationally expensive. The distance between a pair of strings or trees is computed in polynomial time; nevertheless, the computation of the distance between a pair of graphs is exponential respect the number of vertices. For this reason, some algorithms that compute the distance between graphs have been presented that obtain a sub-optimal distance [1]. Although these last algorithms have a polynomial computational cost, the run time is not acceptable for some applications.

In some cases, the computer that has to process the pattern recognition task is embedded in a mobile robot or is a personal computer. The tendency of personal computers is to have a multi-core generic processor and also a special Graphics Processor Unit (GPU) dedicated to intensive computations [2].

* This research was partially supported by Consolider Ingenio 2010; project CSD2007-00018 and by the CICYT project DPI 2010-17112.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 508–515, 2011.

This paper presents a new research project that aims to compute pattern recognition tasks in a up-to-date desktop computer. Intensive computation tasks are computed in the GPU, such as the graph-matching algorithms. In this framework, some pattern recognition applications computed in desktop computers or mobile robots can make use of graphs to represent the involved objects in an acceptable run time. The bases of our work are commented in the next section and the new parallel algorithm is explained in section 3. Section 4 shows the runtime of the sequential algorithm in comparison to the new parallel algorithm with graphs of order up to 1024 vertices. Section 5 concludes the paper.

2 Graph Matching and Computer Architecture

In this section, we introduce a graph-matching algorithm and we relate it to up-to-date desktop computer architecture. The aim is to establish the basis for computing pattern recognition algorithms of high computational cost in current computers.

2.1 Attributed Graphs and Graph Matching

The Graduated Assignment algorithm [1] may by the most widespread graph matching algorithm. The algorithm approximates a distance and a labelling between two graphs using a polynomial time method respect the order of the graphs. The result of the Graduated Assignment algorithm is a probability matrix M that represents, in each element, the probability of matching a node of one of the graphs to a node of the other graph. Since matrix M values are continuous, to obtain the final labelling between nodes of both graphs, a discretisation process of the probability matrix [3] is needed. This process is out of the scope of this paper.

Given a pair of graphs G_S and G_D (that have A vertices) and their respective adjacency matrices S and D, the general outline of the Graduated Assignment is shown in algorithm 1.

Algorithm 1. General diagram of the Graduated Assignment.

```
repeat
    repeat
        M := Update(M, S, D, β)
        M := Normalise(M)
    until M convergence
    β := next β
until M convergence return M
```

Function *Update* updates the probability matrix as follows,

$$\forall_{a=1}^{A}\forall_{i=1}^{A} M_{ai} := exp(\beta Q_{ai}) \text{ given } \begin{cases} Q_{ai} = \sum_{b=1}^{A}\sum_{j=1}^{A} M_{bj} C_{aibj} \\ C_{aibj} = S_{ab} D_{ij} C_{ai} C_{bj} \end{cases} \quad (1)$$

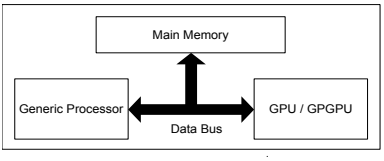


Fig. 1. General view of a desktop computer architecture

Where C_{aibj} represents the compatibility of labelling edge (a, b) of graph G_S to edge (i, j) of graph G_D and their respective ending nodes. Several options to define this function have been presented. Since in our applications edges do not have attributes, we define the following one: $C_{aibj} = C_{ai}C_{bj}$ if both edges exists and $_{aibj} = 0$, otherwise. Compatibilities C_{ai} and C_{bj} represent the compatibilities of matching the respective nodes of both graphs and they are usually pre-computed and stored in a matrix C.

Function *Normalise* obtains a double stochastic matrix [1] using the Sinkhorn method [4] as follows,

Algorithm 2. Graduated Assignment matrix normalisation algorithm.

repeat
$\forall_{a=1}^{A}\forall_{i=1}^{A} M_{ai}^{1} := \frac{M_{ai}}{\sum_{k=1}^{A} M_{ak}}$
$\forall_{a=1}^{A}\forall_{i=1}^{A} M_{ai} := \frac{M_{ai}^{1}}{\sum_{k=1}^{A} M_{ki}^{1}}$
until M convergence **return** M

Sinkhorn method has been parallelised for many high-performance architectures, such as vector machines [5] and connection machines [6].

2.2 Desktop Computer Architecture

The current desktop computers are composed by 2 processors: A generic multicore processor (composed by few cores) and a Graphics Processing Unit (GPU, composed by tens of small cores). Both processors have access to main memory (figure 1).

Current GPUs are General Purpose GPUs (GPGPU) dedicated to intensive computations, mainly addressed to graphic tasks. They are able to execute simple functions usually called *kernels*. GPGPUs are massively multi-threaded architectures. They are composed by several multiprocessors (figure 2), each of which has multiple cores and a shared memory. *Cores* are the processing units that compute thread instructions. Shared memory has multiple banks, they can serve data simultaneously to multiple threads. This shared memory has small size but very low latency.

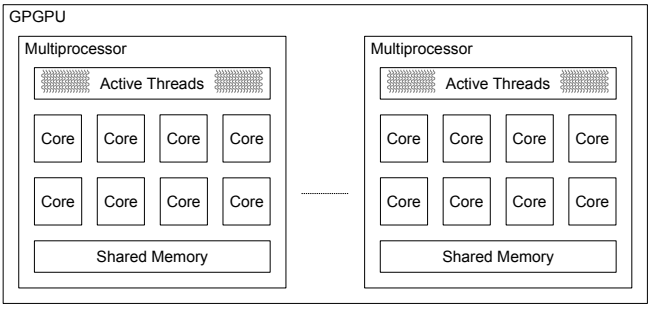


Fig. 2. Generic GPGPU Architecture

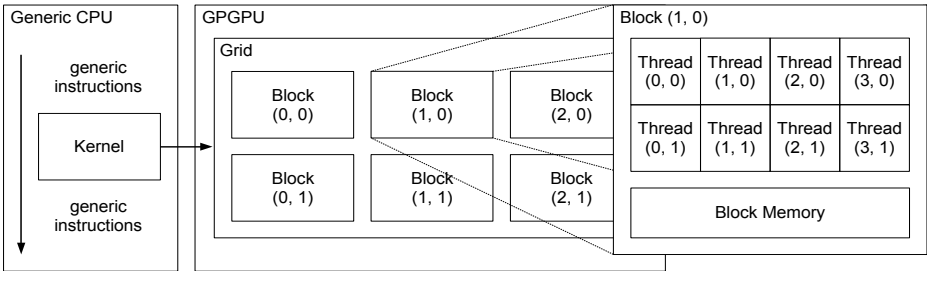


Fig. 3. Generic GPGPU Architecture

2.3 Parallel Programming Model

Our parallel programming model is CUDA [7]. This programming framework allows to mix sequential C code, executed in the generic processor, with kernels, executed in the GPGPU. When the sequential code reaches a kernel, it configures a logical grid of blocks (figure 3) and launches its execution on the GPGPU. A kernel code is executed concurrently by all threads of the grid of blocks. Each block is physically mapped into a GPGPU multiprocessor. The threads of a block are executed in the cores of one multiprocessor and the block memory is mapped into the shared memory of the multiprocessor (figure 2).

We define the following four directives to express parallel tasks:

Parallel Map$^{(X2,Y2)}_{(x,y)=(X1,Y1)}$ **Into Blocks** { *func* }: The same function is executed by all blocks *(x,y)* in the grid of blocks between block $(X1, Y1)$ and block $(X2, Y2)$.
Parallel Map$^{(X2,Y2)}_{(x,y)=(X1,Y1)}$ **Into Threads** { *func* }: The same function is executed by all threads *(x,y)* of specific block between thread $(X1, Y1)$ and thread $(X2, Y2)$.
Parallel Reduction *func* : A function is performed by an specific block.
Parallel Fetch$^{(X2,Y2)}_{(x,y)=(X1,Y1)} U_{x,y}$ Given a matrix U stored in the main mem-

ory, an specific thread of an specific block reads a sub-matrix of elements of U between position $(X1, Y1)$ and position $(X2, Y2)$ and writes it to the shared memory of the block. Moreover, this directive acts as a barrier for all threads of the same block.
Fetch $U_{x,y}$: Given a matrix U stored in the main memory, an specific thread of an specific block reads the element $U_{x,y}$ and writes it to the shared memory of the block.

3 A Parallel Solution for the Graph Matching problem

This section shows a parallel solution of the two main functions of the Graduated Assignment (algorithm 1): *Update* and *Normalise*.

3.1 Parallel Update of the Probability Matrix

From the theoretical point of view, *Update* function computes (1) and it is highly parallel, nevertheless, an algorithm that aims to compute this function has to consider the restriction of number of threads, memory capacity and bandwidth. The probability matrix M adjacency matrices S, D and the compatibility matrix between nodes C could be stored in the main memory (figure 1) or in the shared memory of the multiprocessors (figure 2). Considering the first option, the cardinality of matrices (that is, the number of vertices of the graphs) can be very high but the communication overhead also becomes very high, decreasing the run time of the algorithm. Considering the second option, there is an important restriction of the number of nodes of the graphs since the shared memories are much smaller than the main memory.

We present Algorithm 3 as a parallel solutionto update the probability matrix M, which balances the communication overhead and scalability of the number of nodes of the graphs. Matrices M, S, D and C are stored in the main memory but they are divided into sub-matrices of $B \times B$ elements, which are temporarily stored in the shared memories. In the algorithm, coordinates (ha, hi) points to the current block and coordinates (la, li) points to the current thread inside this block (ha, hi). This mapping is carried out in lines 1 and 2. Coordinates (a, i) points to elements of M, S, D and C being $a \in [1 \dots A]$ and $i \in [1 \dots A]$ (line 3). Variables a, i, b and j of the algorithm represent indexes a, i, b and j respectively in (1). Figure 4.1 shows the distribution of M in blocks and the variables above mentioned. The value $M_{a'i'}$ is computed from line 4 to line 18 by the thread (la', li') in the block (ha', hi'). Figures 4.2, 4.3 and 4.4 show the distribution in blocks of matrices S, D and M and C, respectively. Thread (a', i') in block (ha', hi') accesses to the grey and black cells. Suppose the execution of the algorithm is in the point that $hb = hb'$ and $hj = hj'$, then, black cells in these figures represent the following: Figure 4.2: Fetch of S_{ab} (line 7). Figure 4.3: Fetch of D_{ij} (line 9). Figure 4.4: Fetch of M_{bj} (line 10) or C_{bj} (line 11).

Algorithm 3. General diagram of the Graduated Assignment.

```
1: Parallel Map^{(A/B-1, A/B-1)}_{(ha,hi)=(0,0)} Into Blocks {
2:     Parallel Map^{(B,B)}_{(la,li)=(1,1)} Into Threads {
3:         a := ha · B + la; i := hi · B + li
4:         Q_ai := 0
5:         Fetch C_ai
6:         For^{(A/B-1)}_{(hb)=(0)} {
7:             Parallel Fetch^{(a,(hb+1)·B)}_{(a,b)=(a,hb·B+1)} S_ab
8:             For^{(A/B-1)}_{(hj)=(0)} {
9:                 Parallel Fetch^{(a,(hj+1)·B)}_{(i,j)=(i,hj·B+1)} D_ij
10:                Parallel Fetch^{((hb+1)·B,(hb+1)·B)}_{(b,j)=(hb·B+1,hj·B+1)} C_bj
11:                Parallel Fetch^{((hb+1)·B,(hb+1)·B)}_{(b,j)=(hb·B+1,hj·B+1)} M_bj
12:                For^{(B,B)}_{(lb,lj)=(1,1)} {
13:                    b := hb · B + lb; j := hj · B + lj
14:                    Q_ai := Q_ai + M_bj S_ab D_ij C_ai C_bj
15:                }
16:            }
17:        }
18:        M^0_ai := exp(βQ_ai)
19:    }
20: }
```

3.2 Parallel Normalisation Using Sinkhorn Method

Algorithm 4 is a parallel solution of the Sinkhorn method. In this second function, matrix M is not stored in the shared memories and it is updated on the main memory. Only the variables sum_a and sum_i are stored in the shared memories. Each block process a row (line 2) and a column (line 6) of matrix M. The normalisation process is done by the threads of each block (lines 3 and 4) and (lines 7 and 8). Note that we define a temporal matrix M^1, which is also stored in the main memory.

4 Practical Evaluation

We have implemented the sequential algorithm [1] and the proposed parallel algorithm. Both algorithms have been tested over a GPGPU parallel architecture and over a generic Intel architecture. Table 1 shows the four experiments. Columns 2 and 3 show the algorithm and the used desktop computer, respectively. Column 4 shows the processor (Generic processor or GPGPU in which the code was executed, see figure 1). The other columns show the architecture characteristics.

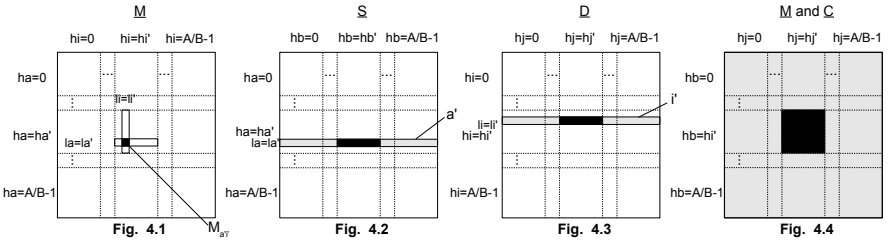


Fig. 4. Fig. 4.1 shows the blocks of M. Fig. 4.2, 4.3 and 4.4 show fetch of thread (a', i').

Algorithm 4. Parallel normalisation algorithm.

repeat
  Parallel Map$^{(A,i)}_{(a,i)=(1,i)}$ **Into Blocks** {
    Parallel Reduction $sum_a := \sum_{i=1}^{A} M_{ai}$
    Parallel Map$^{(a,A)}_{(a,i)=(a,1)}$ **Into Threads** { $M^1_{ai} := \frac{M_{ai}}{sum_a}$ }
  }
  Parallel Map$^{(a,A)}_{(a,i)=(a,1)}$ **Into Blocks** {
    Parallel Reduction $sum_i := \sum_{a=1}^{A} M_{ai}$
    Parallel Map$^{(A,i)}_{(a,i)=(1,i)}$ **Into Threads** { $M^1_{ai} := \frac{M_{ai}}{sum_i}$ }
  }
until M convergence

Table 1. Architectures used in the comparative and their characteristics

Test	Alg.	Computer	Proc./GPGPU	GHz	Power	Cores	Threads	Bandwith
Test1	[1]	ViewSonic VT132	Intel Atom 330	1.6	8W	2	4[1]	5GB/s
Test2	New	ViewSonic VT132	NVIDIA 9400M	1.1	10W	16	1536	5GB/s
Test4	New	MacBook Air	NVIDIA 320M	0.97	14W	48	4608	10GB/s
Test5	New	GA-965P-DS4	NVIDIA 8800GT	1.65	>50W	112	10752	53GB/s

The test set is composed by random graphs with a connectivity percentage between 10% and 50%. Each graph has cardinality from 16 to 1024 vertices. Given a randomly generated graph, the other graph to be compared to is obtained from it and then it is modified by randomly changing the order of nodes, removing and adding edges, changing node values, and removing or adding some nodes.

Figure 5 shows the mean run time of the four experiments given different cardinalities of the graphs. The obtained distance is not shown since the sequential and parallel algorithm obtains exactly the same result. It can be observed a clear improvement on the run time when the parallel algorithm is used.

[1] There are 4 threads available but algorithm [1] uses only one thread.

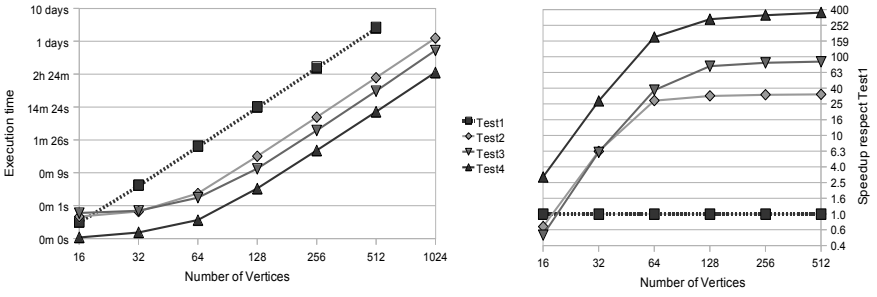


Fig. 5. Run time of the 4 tests respect to the number of vertices and speed-up of the parallel solutions (Test2, Test3, Test4) respect to the serial solution (Test 1). Both are in log. scale.

5 Conclusions and Future Work

We have presented a parallel algorithm which can take advantage of present computational resources on current desktop computers. Results show a significant speed-up of the run time of graph-matching algorithms. The aim is to demonstrate that it is possible to perform graph-matching with hundreds of vertices in a current and low-cost computer or a laptop embedded in a mobile robot. Future efforts will focus on parallelise other graph-matching algorithms, and take advantage of other available resources, like generic multi-core.

References

1. Gold, S., Rangarajan, A.: A Graduated Assignment Algorithm for Graph Matching. IEEE TPAMI 18(4), 377–388 (1996)
2. Owens, J.: Streaming architectures and technology trends. GPU Gems 2, 457–470 (2005)
3. Kuhn, H.W.: The Hungarian method for the assignment problem Export. Naval Research Logistics Quarterly 2(1-2), 83–97 (1955)
4. Sinkhorn, R.: A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. The Annals of Mathematical Statistics 35(2), 876–879 (1964)
5. Zenios, S.A., Iu, S.-L.: Vector and parallel computing for matrix balancing. Annals of Operations Research 22, 161–180 (1990)
6. Zenios, S.A.: Matrix balancing on a massively parallel Connection Machine. ORSA Journal on Computing 2, 112–125 (1990)
7. NVIDIA CUDA, `http://developer.nvidia.com/object/cuda.html`

A Probabilistic Framework to Obtain a Common Labelling between Attributed Graphs*

Albert Solé-Ribalta and Francesc Serratosa

Universitat Rovira i Virgili (URV)
Departament d'Enginyeria Informàtica i Matemàtiques
43007 Tarragona, Catalonia, Spain
{albert.sole,francesc.serratosa}@urv.cat

Abstract. The computation of a common labelling of a set of graphs is required to find a representative of a given graph set. Although this is a NP-problem, practical methods exist to obtain a sub-optimal common labelling in polynomial time. We consider the graphs in the set have a Gaussian distortion, and so, the average labelling is the one that obtains the best common labelling. In this paper, we present two new algorithms to find a common labelling between a set of attributed graphs, which are based on a probabilistic framework. They have two main advantages. From the theoretical point of view, no additional nodes are artificial introduced to obtain the common labelling, and so, the structure of the graphs in the set is kept unaltered. From the practical point of view, results show that the presented algorithms outperform state-of-the-art algorithms.

1 Introduction

In some applications, it is required a representative of a set of Attributed Graphs. To obtain this representative, it is necessary to compute first a common labelling between all nodes of all the graphs in the set. This common labelling incorporates the knowledge of the local parts of the structures and their similarities and relations. Reference applications could be found in [1] and [2].

Unfortunately, only few techniques to compute these correspondences have been developed. They can be divided in two different classes. The first ones [3], [4] and [5] share the same weaknesses: the pairwise labellings between nodes of the graphs are computed at the first step and are kept unaltered; therefore, a simple pairwise error taken at the first step could derive in a bad global result. The second class compresses [2] and [6], however its use is not feasible with large graphs sets due to the computational cost. In this article, we model the common labelling problem using a probabilistic framework and we describe two alternatives to obtain the final common labelling.

This document is structured as follows. Section 2 defines the graph-matching and common-labelling problem. Section, 3 introduces a Probabilistic Framework. Section 4 and 5 present two methodologies to obtain the common labelling. Sections 6 and 7 evaluate the algorithms from the theoretical and practical point of view. Finally, Section 8 draws conclusions and gives directions for further work.

* This research is supported by "Consolider Ingenio 2010": project CSD2007-00018, by the CICYT project DPI2010-17112 and by the Universitat Rovira I Virgili through a PhD grant.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 516–523, 2011.

2 Graph Matching and Common Labelling of a Set of Graphs

An Attributed Graph over Δ_v and Δ_e is defined by a tuple $AG = (\Sigma_v, \Sigma_e, \gamma_v, \gamma_e)$, where $\Sigma_v = \{v_k | k = 1, ..., R\}$ is the set of vertices, $\Sigma_e = \{e_{ij} | i, j \in \{1, ..., R\}; i \neq j\}$ is the set of arcs and $\gamma_v : \Sigma_v \rightarrow \Delta_v$ and $\gamma_e : \Sigma_e \rightarrow \Delta_e$ assign attribute values to vertices and arcs respectively. Given two graphs G^p and G^q, there are several error-tolerant graph matching algorithms that aim to obtain the best isomorphism $f^{p,q}$ between them, given a minimization criteria. This isomorphism assigns each vertex from G^p to only one vertex of G^q, $f^{p,q} : \Sigma_v^p \rightarrow \Sigma_v^q$. The cost of matching graphs G^p and G^q given the isomorphism $f^{p,q}$ is represented as $C(G^p, G^q, f^{p,q})$. The lower the cost, the better the isomorphism. Considering that these graphs have a degree of disturbance and also the exponential complexity of the problem [7], some of these algorithms [8], [9], [10], [11], [12] do not return exactly the isomorphism $f^{p,q}$ but a probability matrix related to it. We represent this matrix by $P_f^{p,q}$ where each cell contains:

$$P_f^{p,q}[i,j] = Prob(f^{p,q}(v_i^p) = v_j^q) \tag{1}$$

Usually, these iterative algorithms (that we will name K) obtain a more accurate approximation of $P_f^{p,q}$ at each step until a convergence criterion is reached. Each step can be represented like:

$$(P_f^{p,q})^{t+1} = K^{step}(G^p, G^q, (P_f^{p,q})^t) \tag{2}$$

To obtain the final isomorphism, $f^{p,q}$, it is necessary to convert $P_f^{p,q}$ into $f^{p,q}$. There are several techniques to find these isomorphisms ([11], [14]) which are out of the scope of this paper. We call this discretisation process Λ, in this way, $f^{p,q} = \Lambda(P_f^{p,q})$.

Given a set of N attributed graphs $\Gamma = \left\{G^1, G^2, ..., G^N\right\}$, we define the set of all possible pairwise isomorphisms as a multiple isomorphism: $\varphi = \left\{f^{1,2}, ..., f^{2,1}, ..., f^{N,N-1}\right\}$. We consider a multiple isomorphism to be a consistent multiple isomorphism if:

$$f^{q,k}(f^{p,q}(v_i^p)) = f^{p,k}(v_i^p); 1 \leq p, q, k \leq N; 1 \leq i \leq R \tag{3}$$

The main idea of a consistent multiple isomorphism is that in all the graphs, there is a vertex that represents the same local part of a general structure (or global object). In our model, this local part is represented by a vertex l_k that belongs to a virtual vertex set, $L = \{l_1, l_2, ..., l_N\}, l_k \in \Sigma_v$. In this way, we define a Common Labelling $\psi = \left\{h^1, h^2, ..., h^N\right\}$ between all the vertices of the graphs to this virtual set, L, in the following way:

$$\begin{aligned} &h^1(v_i^1) = l_i; 1 \leq i \leq R \\ &h^p(v_i^p) = h^{p-1}(v_j^{p-1}); 1 \leq i, j \leq R; 2 \leq p \leq N \\ &\text{where } f^{p-1,q}(v_j^{p-1}) = v_i^p \end{aligned} \tag{4}$$

The common labelling relates each vertex of each graphs to a vertex of the virtual structure that represents the same local part of a general structure. Figure 1 shows the relation between the common labelling and the consistent multiple isomorphism.

The cost of a consistent multiple isomorphism φ given a set of graphs Γ is defined as,

$$C(\Gamma, \varphi) = \sum_{\forall G^p, G^q} C(G^p, G^q, f^{pq}); G^p, G^q \in \Gamma; f^{pq} \in \varphi \tag{5}$$

Again, the lower the cost, the better the multiple isomorphism. The most well know algorithm that solves the common labelling problem was presented by Bonev *et al.* [4]. Their method has two main steps. In the first one, they obtain all the probability matrices for each pair of graphs, $P_f^{p,q}$. To obtain these matrices, they can use any graph-matching algorithm K. In the second step, all the matrix cells are sorted in descending order to compute, what they call, a super graph. The cost of their algorithm is $O(Cost(K)N^2 + N^2R^2Log(N^2R^2))$.

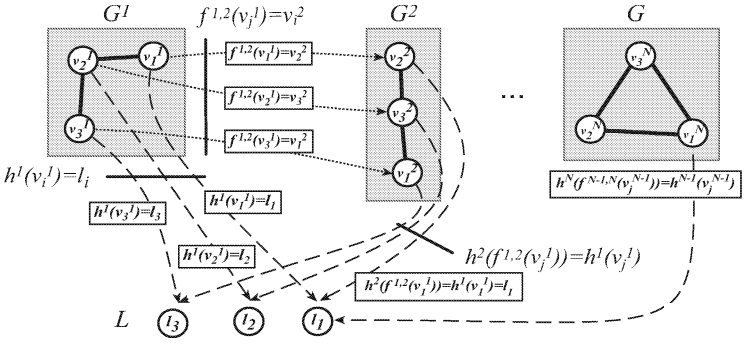


Fig. 1. A set of graphs on the top and a virtual set on the bottom. A common labelling h relates the vertices of the graphs with the virtual nodes l_k throughout the consistent multiple isomorphism composed by all the $f^{p,q}$.

3 A Probabilistic Framework to Obtain a CL

Similarly to the graph-matching algorithms, we define the probability of matching a graph vertex v_i^p to a virtual vertex l_k as,

$$P_h^p[i,k] = Prob(h^p(v_i^p) = l_k) \tag{6}$$

We consider that the probability of matching vertex v_i^p of graph G^p to vertex l_j of the virtual node set L is the probabilistic union of all the paths that goes through the nodes $v_{1..R}^q$ of a third graph G^q. In other words,

$$Prob(h^p(v_i^p) = l_k) = Prob\begin{pmatrix} [f^{p,q}(v_i^p) = v_1^q \wedge h^q(v_1^q) = l_k] \vee \\ [f^{p,q}(v_i^p) = v_2^q \wedge h^q(v_2^q) = l_k] \vee \\ ... \vee \\ [f^{p,q}(v_i^p) = v_R^q \wedge h^q(v_R^q) = l_k] \end{pmatrix} \tag{7}$$

Assuming independence in (7) and considering definitions (4) and (6), we deduce:

$$P_h^p[i,j] = \sum_{k=1}^{R} P_f^{p,q}[i,k] \cdot P_h^q[k,j] \text{ from where } P_h^p = P_f^{p,q} \cdot P_h^q \tag{8}$$

In a similar way, we could infer that:

$$P_f^{p,q} = P_h^p \cdot (P_h^q)^{Trans} \tag{9}$$

Fig. 2 shows how the probability P_h is obtained through the set of probabilities P_f. If the multiple isomorphism obtained by the discretisation of the set of probabilities,

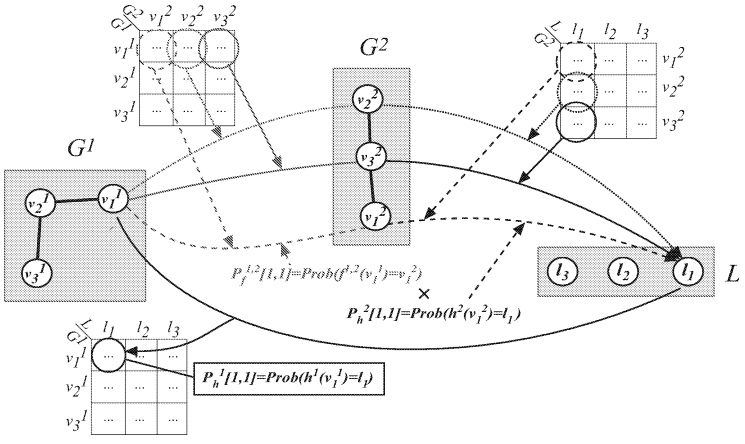


Fig. 2. Probability of a common labelling h obtained through the probabilities of a multiple isomorphism f

$f^{p,q} = \Lambda(P_f^{p,q}), 1 \leq p,q \leq N$, is consistent, then we could obtain the common labelling applying (4) and (6) as follows.

$$\begin{array}{l} h^1 = \Lambda(P_h^1) = \Lambda(\text{identity\_matrix}) \\ h^2 = \Lambda(P_h^2) = \Lambda(P_f^{2,N}) \cdot \Lambda(P_h^N) = ... = \Lambda(P_f^{2,1}) \cdot \Lambda(P_h^1) = \Lambda(P_f^{2,1}) \\ ... \\ h^N = \Lambda(P_h^N) = \Lambda(P_f^{N,N-1}) \cdot \Lambda(P_h^{N-1}) = ... = \Lambda(P_f^{N,1}) \cdot \Lambda(P_h^1) = \Lambda(P_f^{N,1}) \end{array} \tag{10}$$

However, in real data, it is usual the obtained multiple isomorphism is not consistent. Consequently, some of the equalities of (10) do not hold and labellings h^p can not be computed directly through the discretisation of the probability matrices P_f and P_h as in (10). An option is to consider P_h^p as the sample mean:

$$P_h^p \approx \bar{P}_h^p = \frac{1}{N-1}\left(P_f^{p,1} \cdot P_h^1 + P_f^{p,2} \cdot P_h^2 + ... + P_f^{p,N} \cdot P_h^N\right) \tag{11}$$

The two following sections describe two different algorithms to compute $\bar{P}_h^p$.

4 Common Labelling via Minimum Least Square Method (LSM)

The first method is based on considering (11) as a system of linear equations with some restrictions. We could represent the sample mean of P_h (11) in matrix form as:

$$\begin{array}{l} \begin{bmatrix} P_h^1[1,1] & P_h^1[1,2] \\ P_h^1[2,1] & P_h^1[2,2] \end{bmatrix} = \begin{bmatrix} P_f^{1,2}[1,1] & P_f^{1,2}[1,2] \\ P_f^{1,2}[2,1] & P_f^{1,2}[2,2] \end{bmatrix} \begin{bmatrix} P_h^2[1,1] & P_h^2[1,2] \\ P_h^2[2,1] & P_h^2[2,2] \end{bmatrix} + P_f^{1,3} \cdot P_h^3 \\ ... \\ \begin{bmatrix} P_h^3[1,1] & P_h^3[1,2] \\ P_h^3[2,1] & P_h^3[2,2] \end{bmatrix} = \begin{bmatrix} P_f^{3,1}[1,1] & P_f^{3,1}[1,2] \\ P_f^{3,1}[2,1] & P_f^{3,1}[2,2] \end{bmatrix} \begin{bmatrix} P_h^1[1,1] & P_h^1[1,2] \\ P_h^1[2,1] & P_h^1[2,2] \end{bmatrix} + P_f^{3,2} \cdot P_h^2 \end{array} \tag{12}$$

that corresponds to a linear system with $N \cdot R$ equations and $N \cdot R$ unknowns:

$$\begin{aligned}
&P_h^1[1,1] = P_f^{1,2}[1,1]\,P_h^2[1,1] + ... + P_f^{1,3}[1,1]\,P_h^3[1,1] + P_f^{1,3}[1,2]\,P_h^3[2,1] \\
&P_h^1[1,2] = P_f^{1,2}[1,1]\,P_h^2[1,2] + P_f^{1,2}[1,2]\,P_h^2[2,2] + ... \\
&.... \\
&P_h^3[2,2] = P_f^{3,1}[2,1]\,P_h^1[1,2] + ... + P_f^{3,2}[2,1]\,P_h^2[1,2] + P_f^{3,2}[2,2]\,P_h^2[2,2]
\end{aligned} \tag{13}$$

However, due to the system does not have any independent term, it just has one trivial solution $\bar{P}_h^p = [0]$. This trivial solution has a comprehensible meaning due to the virtual set L has neither attributes nor structure. To solve this problem, we impose the restriction that P_1^p is a constant matrix. In this way, the discretisation process produces $h^1(v_i^1) = l_i, 1 \leq i \leq R$ as in (4). With this restriction, the system of linear equations becomes an over determined with $N \cdot R$ equations and $(N-1) \cdot R$ variables. We find an approximation of the solution using the Minimum Least Square Methodology (MLS). The cost of computing P_h given all pairwise matrices P_f is $O(((N-1) \cdot R)(N \cdot R)(N \cdot R+1))$ (to perform the QR factorization using the Gram-Schmidt method) plus $O(\frac{1}{2}(N \cdot R)(N \cdot R+1))$ (to solve the system) [15].

5 Common Labelling via Probabilistic Iterative Algorithm

The second method is based on considering the common labelling as a generalization of the graph matching problem. Our algorithm computes a common labelling given two parameters: a set of graphs and an iterative graph matching algorithm K. We need the graph-matching algorithm to be iterative (2) since we compute only one step of the graph-matching algorithm at each step of our algorithm. The pseudocode of the algorithm is listed below:

0: Start Algorithm(G)
1: P_h^i = inicializeCL();
2: repeat until$(P_h^p)^t\ \forall p \in \{1..N\}$converges or $t \geq K^{Max}$
3: for all $G^p \in \{\Gamma - G^1\}$
4: $(P_h^p)^{t+1} = [0]$;
5: for all$G^q \in \{\Gamma - G^p\}$
6: $(P_f^{p,q})^{t+1} = K^{step}(G^p, G^q, (P_h^p)^t \cdot ((P_h^q)^t)^{Trans})$
7: $P_h^{'} = enhanceEntropy((P_f^{p,q})^{t+1}, (P_f^{p,q})^{t+1} \cdot (P_h^q)^t)$
8: $(P_h^p)^{t+1} = (P_h^p)^{t+1} + \dfrac{1}{N-1} P_h^{'}$
9: end
10: $(P_h^q)^{t+1} = normalize((P_h^q)^{t+1})$
11: end
12: $t = t+1$
13: end
14: End Algorithm returns $\Lambda(P_h^p), \forall p = 1..N$

Where K^{Max} indicates the maximum number of iterations that algorithm K requires. Function $InitializeCL$, initializes matrices P_h to $1/N$ and P_1 to the identity matrix (4). Function $enhanceEntropy()$ applies the Softmax method [16] to keep a similar entropy at each step. This operation is necessary because the product of two stochastic matrices tends to converge to a stationary distribution. Finally, function $normalize()$ applies the method presented in [13] to ensure that P_h is an stochastic matrix. The algorithm computes N^2 calls to the K^{step} function plus a constant cost (lines 7, 8 and 10) at each iteration. The global cost is:

$$O([Cost(K^{step}) + Cost(line7, line8)] \cdot N^2 \cdot K^{Max} + Cost(line10) \cdot N \cdot K^{Max} \quad (14)$$

6 Evaluation of the Computational Cost

Table 1 shows the upper bound of the computational cost of the three algorithms commented above. These equations have been obtained through the specific computational cost explained in the respective sections. The first term of the three equations is similar, which is the cost of computing all the pairwise graph labellings. Since $Cost(K)$ is usually equal or higher than $K^{Max} \cdot N^4$, the second term does not has an important impact on the computational cost.

7 Evaluation

We compare the two presented algorithms with the one presented in [4]. We applied the Graduated Assignment method [11] as a graph-matching algorithm K, in the three algorithms. We used two datasets of graphs that represent 2D objects. Nodes are defined over a 2D domain that represents its plane position (x, y). Edges have a binary attribute that represents the existence of a line between two terminal points. The former dataset is a synthetic one composed by 30 classes. The number of graphs per class is $N \in [3, 5, 7, 9, 10]$ and the noise level between class elements is $\nu \in [10, 20, 40...70]$. Therefore, we defined 5 x 6 = 30 different classes. 6 classes with 3 graphs (with different noise levels), 6 classes with 5 graphs (with different noise levels) and so on. Each class was created as follows. We randomly generate a base graph composed of $R = 10$ nodes with random attributes in the range $\Delta_v = [0..100, 0..100]$. Edges were defined by the Delaunay triangulation. Then, with this base graph, we created $10 \cdot N$ other graphs by: 1, generating Gaussian noise at every node with standard deviation $\sigma = \nu/100$. 2, removing $\nu\%$ nodes randomly. 3, inserting $\nu\%$ nodes (with random attributes) and 4, changing the binary attribute of $\nu\%$ edges. The latter dataset is the Letter dataset (high distortion) created at the University of Bern [17]. It is composed of 15 classes and 150 graphs per class representing different letters of the Roman alphabet. From each class, we randomly selected $N \in [3, 5, 7, 9, 10]$ graphs, to generate the common labelling. With the aim of obtaining non-biased results, the experiments were performed 10 times in both datasets. Table 2 and 3 show the mean cost of a multiple isomorphism (5) applied to the synthetic dataset and grouped by the level of noise introduced to the graphs and the size of the graphs, respectively. Table 4 shows the mean cost of a multiple isomorphism (5) applied to the Letter dataset. Results show that when the order of the graphs is small ($N \leq 4$) or graphs are almost similar ($noiselevel \leq 10$) the three algorithms obtain similar costs. Nevertheless, when the number of vertices increases or the

Table 1. Computational cost upper bounds

Algorithm	Computational Cost (upper bound)
Bonev *et al.* [4]	$O((N^2Cost(K)) + (N^2R^2Log(N^2R^2)))$
Least Square Method (LSM)	$O((N^2Cost(K)) + (N^3R^3 + N^2R^2))$
Common Labelling	$O((N^2Cost(K)) + (N^2R^2 + NR^2))$

Table 2. Mean cost obtained from the synthetic dataset and grouped by Noise Level

Group by noise level	Iterative Alg.	LSM	[4]
10	2280,74	2280,74	2280,74
20	8298,69	8694,84	11066,05
40	12533,24	12912,65	17535,61
50	13108,18	13382,58	18153,18
60	14612,35	14886,47	20629,15
70	14777,80	15056,06	21347,65

Table 3. Mean cost obtained from the synthetic dataset and grouped by the set size (N)

Group by set size (N)	Iterative Alg.	LSM	[4]
3	1442,74	1400,46	1534,03
5	4716,44	4838,36	6081,73
7	9899,13	10200,46	13474,84
9	17118,57	17608,91	23641,58
10	21498,96	21962,91	31111,47

Table 4. Mean cost obtained from the Letter dataset, different number of graphs per class and different algorithms

Group by set size (N)	Iterative Alg.	LSM	[4]
3	239,422	239,648	236,885
5	788,867	798,992	805,897
7	1661,22	1670,141	1692,506
9	2850,235	2852,134	2903,003
10	3554,106	3585,076	3682,933

noise introduced to the graphs increases, our iterative algorithm obtains better results (lower the cost, better the common labelling). The run time of these three algorithms is similar since most of the run time is spent computing the graph-matching algorithm K or each step of this algorithm K^{step}.

8 Conclusions and Further Work

Known algorithms to compute a common labelling consist on first finding the labelling between any pairs of graphs and then combining this information to compute the common labelling. The first algorithm that we present computes the common labelling using the minimum least squares method. The second one is an iterative algorithm that computes the common labelling at the same time as the pairwise labellings, mixing the

local and global knowledge at each step of the algorithm. We have compared the presented algorithms with the most popular one in the literature and we have shown that our both methods obtain a better common labelling with similar computational cost than the reference one. As a future work, we will apply this new technique to the representative of a set of graphs called Structurally-Defined Random Graph [1] and we will analyze its ability to keep the structural and semantic knowledge of the set.

References

1. Solé-Ribalta, A., Serratosa, F.: A structural and semantic probabilistic model for matching and representing a set of graphs. In: Torsello, A., Escolano, F., Brun, L. (eds.) GbRPR 2009. LNCS, vol. 5534, pp. 164–173. Springer, Heidelberg (2009)
2. Williams, M.L., Wilson, R.C., Hancock, E.R.: Multiple Graph Matching with Bayesian Inference. PRL 18(11-13), 1275–1281 (1997)
3. Wong, A.K.C., et al.: Entropy and distance of random graphs with application to structural pattern recognition. IEEE TPAMI 7, 599–609 (1985)
4. Bonev, B., Escolano, F., Lozano, M.A., Suau, P., Cazorla, M.A., Aguilar, W.: Constellations and the unsupervised learning of graphs. In: Escolano, F., Vento, M. (eds.) GbRPR. LNCS, vol. 4538, pp. 340–350. Springer, Heidelberg (2007)
5. Serratosa, F., et al.: Synthesis of function-described graphs and clustering of attributed graph. IJPRAI 16(6), 621–655 (2002)
6. Solé-Ribalta, A., Serratosa, F.: On the Computation of the Common Labelling of a set of Attributed Graphs. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 137–144. Springer, Heidelberg (2009)
7. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness (1979)
8. Messmer, B.T., Bunke, H.: Fast Error-correcting Graph Isomorphism Based on Model Precompilation. In: Del Bimbo, A. (ed.) ICIAP 1997. LNCS, vol. 1310, pp. 693–700. Springer, Heidelberg (1997)
9. Rosenfeld, A., Hummel, R.A., Zucker, S.W.: Scene labeling by relaxation operators. IEEE Transactions on Systems, Man and Cybernetics 6, 420–443 (1976)
10. Feng, J., Laumy, M., Dhome, M.: Inexact matching using neural networks. In: PR in Practice IV: Multiple Paradigms, Comparative Studies and Hybrid Systems, pp. 177–184 (1994)
11. Gold, S., Rangarajan, A.: A Graduated Assignment Algorithm for Graph Matching. IEEE TPAMI 18(4), 377–388 (1996)
12. Christmas, W.J., Kittler, J., Petrou, M.: Structural matching in computer vision using probabilistic relaxation. IEEE TPAMI 17(8), 749–764 (1995)
13. Sinkhorn, R.: A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. The Annals of Mathematical Statistics 35(2), 876–879 (1964)
14. Kuhn, H.W.: The Hungarian method for the assignment problem Export. Naval Research Logistics Quarterly 2(1-2), 83–97 (1955)
15. Dahlquist, G., Bjrck, K.: Numerical Methods, section 5.7. Prentice-Hall Inc., Englewood Cliffs (1974) ISBN 0-13-627315-7
16. Bridle, J.S.: Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: Advances in Neural Information Processing Systems, vol. 2, pp. 211–217. Morgan Kaufmann Publishers Inc., San Francisco (1990)
17. Riesen, K., Bunke, H.: IAM graph database repository for graph based pattern recognition and machine learning. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 287–297. Springer, Heidelberg (2008)

Feature Selection with Complexity Measure in a Quadratic Programming Setting

Ricardo Sousa, Hélder P. Oliveira, and Jaime S. Cardoso

INESC Porto, Faculdade de Engenharia, Universidade do Porto, Portugal
{rsousa,jaime.cardoso}@inescporto.pt,
helder.oliveira@fe.up.pt
http://www.inescporto.pt/~{rsousa,hfpo,jsc}

Abstract. Feature selection is a topic of growing interest mainly due to the increasing amount of information, being an essential task in many machine learning problems with high dimensional data. The selection of a subset of relevant features help to reduce the complexity of the problem and the building of robust learning models. This work presents an adaptation of a recent quadratic programming feature selection technique that identifies in one-fold the redundancy and relevance on data. Our approach introduces a non-probabilistic measure to capture the relevance based on Minimum Spanning Trees. Three different real datasets were used to assess the performance of the adaptation. The results are encouraging and reflect the utility of feature selection algorithms.

Keywords: Feature Selection, Pattern Recognition, Quadratic Optimisation.

1 Introduction

With access to new means of information becoming increasingly easier, there has been a rapid growth of data. Providers gather useful information on customers' tastes and preferences in order to offer products and services similar to penchants of target consumers; bank analysts gather client's profiles based on their bank history and experience; and, with the advent of gene technology, it is now possible to sequence genes with innumerable possibilities of applications. Nevertheless, with the information growth and, consequently, the increased number of characteristics that define the data, new technologies are required to deal with this amount of information efficiently. One way is through the use of feature selection (FS) techniques that try to capture only the relevant information from data.

FS is an important issue in many applications, and in some cases, it is a necessary step for the construction of the predictive model (classification or regression). This became more important especially for data with an high dimensional nature. Moreover, the use of unimportant features can add undesirable complexity to the model and can negatively influence its behaviour by introducing errors in the prediction. Therefore, FS techniques were designed with different

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 524–531, 2011.

purposes, such as preventing overfitting, improving classification performance, design of faster models and to improve understanding of the processes that generate the data. Furthermore, the usage of FS techniques can induce a reduction of the complexity of the dataset.

Other approaches used for dimensionality reduction, such as principal component analysis, differ from the FS problem. While the first one changes the original representation of the data, the second only uses a subset (containing hopefully the most important information) of the whole feature set. This means that the original semantics are preserved.

FS algorithms can be divided into three different categories depending on how the FS is incorporated into the predictive model construction task. The categories are mainly defined as filter, wrapper and embedded methods. The first one filters the data by removing features with low relevance. This is performed as a preprocessing step and it is independent of the construction phase of the predictive model. Wrapper methods iteratively select a subset of features until a sufficiently 'good' model is constructed. This evaluation is usually performed during the training and validation phase of the model. Finally, the embedded techniques are intrinsically designed with the learning model. A best subset of features is obtained during the training phase.

In [8] a new FS algorithm that captures in just one step the correlated features and those related to the classes is presented. This is achieved with a quadratic programming formulation where a term α controls how much importance the algorithm should give to the within features correlation or to feature-class correlation. To measure this correlation the authors in [8] use either the Pearson or Mutual Information (MI) to measure linear or non-linear relations, respectively. In this article an adaptation is presented motivated by [10] where the major advantage is a non-probabilistic approach for data estimation.

In Section 2 a recent method which tackles the problem of FS through a quadratic programming formulation will be briefly discussed. Some issues regarding the use of MI in the FS context will also be outlined. This strategy tries to identify the non-redundant features and those that are correlated with the class labels. In Section 3 an adaptation of this strategy of capturing the instances correlated with the class labels through a non-probabilistic scheme will be outlined. Finally, in Section 4 some results will be presented and some conclusions will be drawn in Section 5.

2 Quadratic Programming for Feature Selection

The authors in [8] proposed a new way of performing FS using a quadratic programming formulation. The FS problem can be described as a two-fold problem. First, one tries to eliminate the similar variables (redundancy) and secondly, to capture how correlated each feature is with the class (relevance). In [8] the authors propose to tackle the FS problem in one-step process through quadratic

programming. The quadratic term (Q in Equation (1)) would capture the redundancy whereas the linear term (F in Equation (1)) would capture the relevance.

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}(1-\alpha)\mathbf{x}'Q\mathbf{x} - \alpha F'\mathbf{x} \right\} \tag{1}$$

The α constant can be considered as a trade-off between redundancy and relevance. The $\mathbf{x}$ magnitude values show how important each feature is to the problem. In order to capture these two different pieces of information, the authors first used the Pearson correlation measure. In doing so, the quadratic term (Q) and linear term (F) would be defined as follows:

$$Q_{ij} = \rho_{ij} = \frac{\sum_{m=1}^{M}(v_{mi}-\overline{v_i})(v_{mj}-\overline{v_j})}{\sqrt{\sum_{m=1}^{M}(v_{mi}-\overline{v_i})^2 \sum_{m=1}^{M}(v_{mj}-\overline{v_j})^2}} \qquad F_i = \sum_{k=1}^{C} P(K=k)|\rho_{iC_k}| \tag{2}$$

where M is the number of samples, v_{ki} is the k^{th} sample of random variable v_i and $\overline{v_i}$ is the sample mean of the random variable v_i. P is the empirical prior probability and C_k is the class represented by a 1-of-K coding scheme.

However, this measure only captures the linear relations among instances and between instances and classes. To retrieve the non-linear relations the authors opted to use Mutual Information (MI). This method is borrowed from the information theory fields and measures how much information two random variables share. In this setting, this type of metric would be used to assess redundancy and relevance. It is therefore not surprising that entropy measures such as MI can also be viewed as feature dependence estimators where its broadly use can be associated to the good performances achieved.

In spite of the concept of independence being mathematically well defined, the use of the opposite for dependence assessing does not provide how much in value one variable depends on the other. Hence, the dependence concept in the feature selection setting raises some questions. In [9] a set of postulates proposed by Rényi to define dependence is revised. In this work it is also argued that a measure that holds on those postulates is valid for capturing dependence in the feature selection context. In other words, given a specific set of statistical properties to evaluate dependence, a good estimator would have to satisfy its corresponding properties. It is also argued that this is not valid for MI, since despite most commonly used MI techniques are non-negative and symmetric, they are not invariant to one-to-one transformations and do not reach a maximum when features are highly correlated. Moreover, in [9] it is also argued that highly complex learning machines could easily cope with the data complexity and infer a linear relation with the features and output, or more precisely, perform overfitting on the data.

Due to the issues mentioned above, a more pragmatic approach could be more convenient. For this purpose we have selected MST (Minimum Spanning Trees). Even though MST was introduced as a generalisation of the Wald-Wolfowitz run test, it provides the means for our measure to assess the increase of complexity when a subset of features is removed. Furthermore, such measure can easily be fitted into the Equation (1) to assess data relevance.

Despite the elegant approach [8], the quadratic programming problems can be highly computational heavy. To overcome this difficulty, the authors [8] also proposed an iterative optimisation strategy. Due to time restrictions, it was not possible to explore this approach. Nonetheless, for the proposal delved in the following section and with regards to the experiments performed, a generic quadratic programming solver is sufficient.

3 Feature Selection Based on Redundancy and Relevance

In [8], the Pearson correlation was first used to measure the linear relation on each pair of features (*redundancy*). In other words, features that are highly correlated can be discarded since they do not provide a meaningful informative gain. However, this measure may not be appropriate if features are not linearly correlated. To overcome this problem, the MI was used. With regards to the correlation of features with the class (*relevance*), the Pearson and MI were used in the same manner. When measuring the linear relation between features and classes, a 1-of-K coding scheme for the labels was used [8].

The capture of the non-linear relations through information theory techniques such as mutual information can be very appealing (and widely used, e.g. [4, 1] to name a few). However, and as stated in Section 2, the use of MI for feature selection can, in theory, encompass some issues as argued in [9].

Several works presented in the literature tackle this situation by introducing new techniques to capture the non-linear relations. In the next section, the incorporation of a non-probabilistic based technique in the formulation presented in Equation (1) will be outlined.

3.1 Capturing Relevance through Classifier Complexity

This study was initially motivated by the work presented in [10]. Here, FS is defined as a method which aims to select a subset of features that minimises the overall classification complexity. To achieve this goal, the authors [10] propose a multiresolution approach on the feature space. However, such approach requires high computational effort. Therefore, they suggest techniques such as MST as an alternative [10].

Distribution free procedures are already well covered in the literature. In [6] a technique is proposed to measure the multivariate randomness by using a MST. This method constructs a minimum weight graph that connects all of the N nodes (instances) with E edges. The weights in our study were defined as the euclidean distance between a pair of instances. By using MST over all of our instances, the edges that connect nodes with different classes labels can afterwards be counted [6]. The rationale is, the more edges connecting nodes with different labels one has, the more complex the dataset is. The complexity of an MST is, in the worst case, of the order $\mathcal{O}(E + X \log(N))$, where X is the number of edges no longer than the longest edge in the MST. Our first trial encompassed the direct implementation of this setting to estimate features relevance. This

analysis is performed feature by feature in the linear term in Equation (1) where F_i is inversely proportional to the number of nodes with different labels which share an edge.

Based on this complexity analysis, one can also think how less (or more) complex our dataset can become if we discard a feature, say f_i. For this approach, one starts by measuring the complexity of the dataset with the whole features, $\mathbf{f}$. The gain would be expressed by the complexity increase with the removal of one feature. Formally,

$$F_{-i} = \frac{f_i - \mathbf{f}}{\mathbf{f}} = \Delta f_i$$

Although these techniques can be very complex in computational terms, they do not assume any probabilistic distribution of the data. For both adaptations, this problem was tackled from a wrapper perspective were the parameter α needed to be tuned.

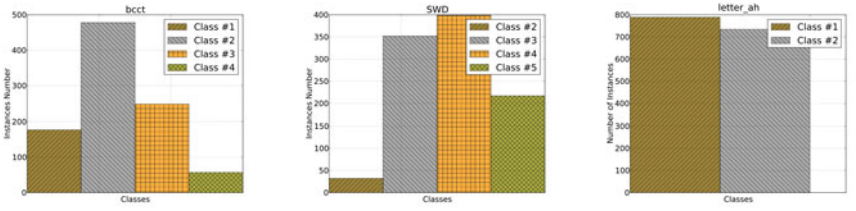
4 Experiments

The aim of this experimental study was to evaluate the usage of other measures to capture the relevance of each feature. Aiming towards the identification of the best features that describe the data, the question of which classifier was used is not considered as the authors did in [8]. Nevertheless, the same learning scheme was used in all experiments.

For the real data we tested the new measures on the Letter, SWD and BCCT datasets [5,2,3]. The first dataset, which is publicly available on the UCI machine learning repository, is composed of $20,000$ instances with 16 features describing the 26 capital letters of the English alphabet. Each instance is mainly defined by statistical moments and edge counts. In our experiments we used a subset of the whole dataset comprehending only the discrimination of the letter A versus the letter H. The SWD dataset contains real-world assessments of qualified social workers regarding the risk facing children if they stayed with their families at home and is composed by 10 features and 4 classes. The last dataset encompasses on 113 patients and expresses the aesthetic evaluation of Breast Cancer Conservative Treatment (BCCT) [3]. For each patient there were 8 observers that manually identified several fiducial points. Based on these points, an automated system automatically recorded 30 features, capturing visible breast alterations such as: breast asymmetry, skin colour changes due to the radiotherapy treatment and surgical scar appearance. The aesthetic outcome of the treatment for each patient was classified in one of the four categories: Excellent, Good, Fair and Poor. For this specific set, the same experiment procedure was conducted as described in [7].

In Fig. 1 the class frequency distribution for each dataset is depicted.

The training was performed on 20% of the data. The splitting of the data into training and test sets was repeated 5 times in order to obtain more stable results for accuracy by averaging and also to assess the variability of this measure. The best parametrisation of each model was found by a 'grid-search' based on a 5-fold cross validation scheme conducted on the training set.



(a) Frequency for BCCT dataset (b) Frequency for SWD dataset (c) Frequency for Letter AH dataset

Fig. 1. Real datasets frequency values

For these experiments the wrapper approach of the original method was selected . Therefore the best α value had to be tuned. We performed a search over the range 0.2 — 0.6 with a 0.2 step. Finally, the error of the model was estimated on the test set using the Misclassification Error Rate (MER) measure.

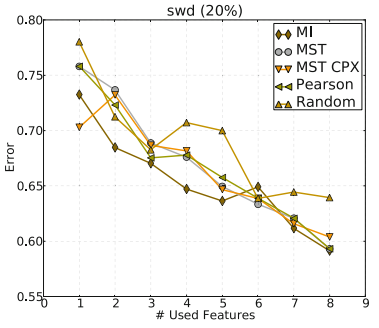
4.1 Results

We started our analysis on the SWD dataset. One can easily see that MI achieves the best results on the major subset of features—see Fig. 2a. However, when using more than 6 features from the whole set of features, MST, MST CPX and MI obtained similar results. The same behaviour can be stated on the letter dataset—see Fig. 2b. Here MST CPX attains a slightly better performance than all of the other techniques when using more than 8 features.

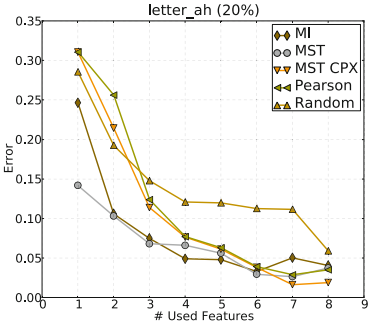
Afterwards, the performance of all feature selection techniques on the BCCT dataset was analysed. One can see that the minimum error attained was with the MI using two features—see Fig. 2c. However, the performance gain was subtle being the difference from the second best method of approximately 3%.

Being already defined some prior knowledge for this problem [7], an experiment exploring the benefit of such technical information was conducted. To achieve this goal, the prior knowledge was introduced as a postprocessing phase of the FS algorithm. This postprocessing can be described as follows. There are three important consequences of the BCCT treatment and they must all be considered. However, there are some features that translate these differences better than others. In order of importance, from the the most to the least important, we have the asymmetry alteration, colour change and scar appearance. For example, if four features had to be selected, we would use 2 asymmetry features, 1 colour feature and 1 of scar feature, in order of importance given by the FS algorithm.

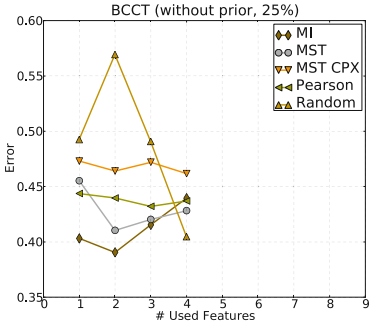
From the prior knowledge perspective, it clearly attained the best results—see Fig. 2d. Even with the prior knowledge, MI and MST based approaches were not able to attain such results. The best selected features were ρBRA and BCD for MST, and BCD and LBC for MI. For the Pearson measure, the best selected features were ρLBC, sX2L, cX2b and ρBCE.



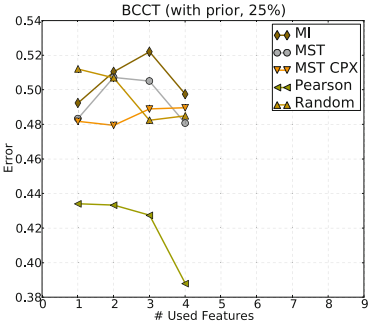
(a) Results for SWD dataset



(b) Results for Letter (A vs. H sub-problem) dataset



(c) Results for BCCT dataset (without prior)



(d) Results for BCCT dataset (with prior)

Fig. 2. Results for real datasets performed with the several FS techniques

Since a wrapper methodology was used, a best classifier would be required to assess each subset of features used. However, due to time restrictions such analysis could not be performed. Therefore, further studies could improve the assessment of this behaviour. Nevertheless, either MST or MST CPX attain very similar results when compared to MI.

5 Conclusion

This paper presents an adaptation of a recent FS method [8] by exchanging the relevance term with a MST as a complexity measurement. The goal of this FS scheme is to quantify the within variables similarity (redundancy) and the correlation between features and the class labels (relevance) through a quadratic optimisation problem. A constant α is used as a trade-off term to define the importance of the redundancy or relevance for the optimisation problem. After selecting a subset of features, they were used during the learning model

construction and their use on a test dataset was assessed using the Misclassification Error Rate (MER). For these experiments, three real (Letter AH, SWD and BCCT) datasets were used.

A preliminarily study shows that MST provides just as good results as MI. However, MST has the advantage of not assuming any data density distribution. It was also interesting to see that on the BCCT database the results show that the baseline method with prior knowledge performs better in terms of classification error when compared to the other approaches.

Regarding to future work, this work can be extended in several directions. One of them passes through the stability assessment of this FS method [11]. Another one is the analysis of the non-linearity within the features. Finally, experiments on larger datasets are required.

Acknowledgements. This work was partially funded by Fundação para a Ciência e Tecnologia (FCT) with reference PTDC/EIA/64914/2006 and SFRH/BD/43772/2008. The authors would also like to thank Sohan Seth for the mathematical clarifications.

References

1. Balagani, K.S., Phoha, V.V.: On the feature selection criterion based on an approximation of multidimensional mutual information. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1342–1343 (2010)
2. Ben-David, A., Sterling, L.: Generating rules from examples of human multiattribute decision making should be simple. Expert Systems with Applications 31(2), 390–396 (2006)
3. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artificial Intelligence in Medicine 40, 115–126 (2007)
4. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. Trans. Neur. Netw. 20, 189–201 (2009)
5. Frey, P.W., Slate, D.J.: Letter recognition using holland-style adaptive classifiers. Mach. Learn. 6, 161–182 (1991)
6. Friedman, J.H., Rafsky, L.C.: Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. Annals of Statistics 7(4), 697–717 (1979)
7. Oliveira, H.P., Magalhaes, A., Cardoso, M.J., Cardoso, J.S.: An accurate and interpretable model for bcct.core. In: Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6158–6161 (2010)
8. Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C.S.: Quadratic programming feature selection. Journal of Machine Learning Research 11, 1491–1516 (2010)
9. Seth, S., Príncipe, J.C.: Variable Selection: A Statistical Dependence Perspective. In: Proceeding of the Ninth International Conference on Machine Learning and Applications, pp. 931–936 (2010)
10. Singh, S.: Multiresolution estimates of classification complexity. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1534–1539 (2003)
11. Somol, P., Novovicova, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11), 1921–1939 (2010)

Automatic Estimation of the Number of Segmentation Groups Based on MI

Ziming Zeng[1,2], Wenhui Wang[3,4], Longzhi Yang[1], and Reyer Zwiggelaar[1]

[1] Department of Computer Science, Aberystwyth University, UK
[2] Faculty of Information and Control Engineering, Shenyang Jianzhu University, Liaoning, China
[3] Network Information Center of the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[4] Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China

Abstract. Clustering is important in medical imaging segmentation. The number of segmentation groups is often needed as an initial condition, but is often unknown. We propose a method to estimate the number of segmentation groups based on mutual information, anisotropic diffusion model and class-adaptive Gauss-Markov random fields. Initially, anisotropic diffusion is used to decrease the image noise. Subsequently, the class-adaptive Gauss-Markov modeling and mutual information are used to determine the number of segmentation groups. This general formulation enables the method to easily adapt to various kinds of medical images and the associated acquisition artifacts. Experiments on simulated, and multi-model data demonstrate the advantages of the method over the current state-of-the-art approaches.

1 Introduction

Tissue segmentation in magnetic resonance (MR) images is an important application area in medical image analysis, for which clustering algorithms have been proposed [1]. A remaining problem is the robust estimation of the optimal number of segmentation groups. The Akaike information criteria (AIC) algorithm [2] is widely used, however, the theoretical possible and actual results do not always agree. Lv et al. proposed a measure based on the difference of mutual information for determining the number of clusters [3]. Wang [4] improved the iteration conditions, but this approach does not work well when noise levels are increased. To overcome the problem of the sensitivity to the presence of noise, we propose a more robust method to estimate the number of segmentation groups.

Mutual information (MI) is a measure of statistical correlation of two random variables. It has been widely used in multi-model medical image registration [5]. Kim et al. and Rigau et al. used MI for image segmentation [6]. Lv et al. proposed a method based on the entropy difference of mutual information to determine the number of segmentation groups. Wang et al. improved the iteration condition and obtained improved results. The complete algorithm from Wang et al. is

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 532–539, 2011.

described in Fig. 1 where k denote the number of groups, $kmeans()$ stand for the K−means algorithm, $mutual\_information()$ denotes the mutual information algorithm, ∂MI denotes the entropy difference in mutual information, ε which is defined as 0.2 is the stopping criterion. Results based on synthetic and simulated data (see Section 4) indicate that this approach does not deal well with increased noise levels.

Input: *Original image.* $k = 2$
Output: *The number of segmentation groups* $= k - 1$
(1) **do**
(2) $groups\_label = kmeans(Original\_image, k)$;
(3) $MI(k) = mutual\_information(Original\_image, groups\_label)$;
(4) **if** $k = 2$
(5) $\partial MI(k) = MI(k)$; $\partial^2 MI(k) = MI(k)$;
(6) **else**
(7) $\partial MI(k) = MI(k) - MI(k-1)$;
(8) $\partial^2 MI(k) = \mathbf{abs}(\partial MI(k-1) - \partial MI(k))/\partial MI(k-1)$;
(9) **end**
(10) $k = k + 1$;
(11)**while** $(\partial^2 MI(k) < \varepsilon)$

Fig. 1. Wang et al.'s algorithm [4]

2 The Proposed Method

In order to enable the approach to handle certain level of noise, we propose a new approach to determine the number of segmentation groups by adopting Wang et al.'s approach. Noise removal, prior to estimating the number of segmentation groups, is expected to be beneficial. However, any such approach should preserve edge information. We propose an improved anisotropic diffusion approach to deal with such issues. Then, we use a class-adaptive Gauss Markov Random Field model, a clustering approach using both individual pixel and spacial information, to cluster. From this, a mutual information-based approach is utilized to judge what is the best number of iterations and segmentation groups. Our proposed method is summarized in Fig. 2.

2.1 Noise Removing

In this step, we use anisotropic diffusion to decrease noise. It is expected this can deal with additive noise and retain edges. The traditional nonlinear diffusion filtering method was proposed by Perona and Malik [7], which is described by

$$\begin{cases} \frac{\partial I}{\partial t} = div\left[c\left(\mid \nabla I \mid\right) \cdot \nabla I\right] \\ I\left(t=0\right) = I_0 \end{cases} \tag{1}$$

where ∇ is the gradient operator, $div()$ is the divergence operator,|| denotes the magnitude, $c\left(x\right)$ denotes the diffusion coefficient, and I_0 is the initial image. The diffusion coefficient which was improved in [8] is defined as

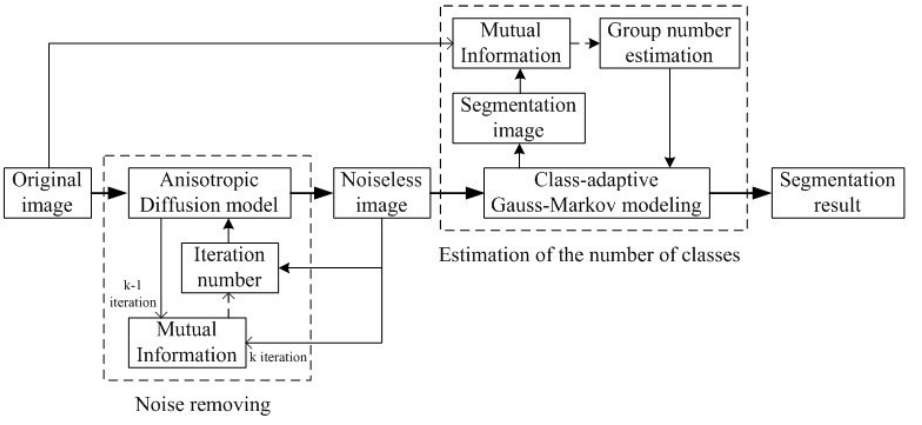


Fig. 2. Schematic representation of the developed approach

$$c(x) = \frac{1}{ln\left[e + (x/\kappa)^2\right]} \tag{2}$$

where κ is an edge magnitude parameter.

Under the control of $c(|\nabla I|)$, the model can achieve a selective smooth diffusion from the original image based on the gradient. At the edge the gradient magnitude is large and $c(|\nabla I|)$ is small and the model will perform hardly any smoothing in order to keep the edge details. When the gradient magnitude is small at a flat area, $c(|\nabla I|)$ is large and noise is removed. In addition, we have included an automatic stopping criteria in the developed approach. This is based on estimating the difference in mutual information (MI) in successive steps and stopping the diffusion when this difference (∂MI) is smaller than an empirically determined threshold value (δ). The traditional anisotropic diffusion only considers 4 directional gradient (up-down-left-right). In our work we extend the anisotropic diffusion gradient direction to eight directions to calculate the pixel gray level.

2.2 Estimation of the Number of Segmentation Groups

In the segmentation step, a class-adaptive Gaussian Markov Random Field is used. The class-adaptive penalty factor β which is anisotropic for each group is automatically calculated from the posterior probability. By incorporating both a Markov random field model and expectation-maximization into a GMRF-EM framework, we can achieve accurate and robust segmentation results.

Let L=$\{1,2,\cdots,l\}$ denote the set of group labels. Let D=$\{1,2,\cdots,d\}$ denote the grey level set. Let S=$\{1,2,\cdots,N\}$ denote the set of indexes which contain N pixels. Image segmentation is achieved by assigning a value from L to each pixel in the image. We define the label assignment X=$\{x = (x_1, \cdots, x_i, \cdots, x_N) \mid x_i \in L, i \in S\}$ to all sites as a realization of a family of random variables defined on S.

We also define an observation field Y=$\{y = (y_1, \cdots, y_i, \cdots, y_N) \mid y_i \in D, i \in S\}$ According to the Hammersley-Clifford theorem [6]: the prior probability P(x) satisfies a Gibbs distribution: $P(X) = Z^{-1} exp(-U(X))$ where Z is a normalisation factor. The potential function is defined as $\mathrm{U}(x) = \beta \sum_{c \in C}(\mathrm{V_c}(x))$ where $C = \{(i, i') \mid i' \in N, i \in S\}$,$c \in$ C. MRF multi-level logistic (MLL) is introduced as the prior probability for which

$$\mathrm{V_c}(x_i) = \begin{cases} -1 \text{ if } x_i = x_j, j \in N_i \\ 1 \text{ else} \end{cases} \tag{3}$$

where N_i denotes the eight neighborhood pixels of the i^{th} pixel. According to [9] the penalty factor β of the i^{th} pixel is defined as

$$\beta = \sqrt{\frac{\sum_{i \in S} \sum_{j \in \mathrm{N}_i^d} (\mathrm{p}_m^i - \mathrm{p}_m^j)^2}{N/k}} \tag{4}$$

where t denotes the number of iterations, m denotes the group, d denotes the number of directions (e.g. horizontal, vertical and two diagonal directions). p_m^i denotes the posterior probability of pixel i. p_m^j denotes the posterior probability of pixel j in the neighborhood i on the direction of d. Assuming that the distribution of intensities y_i follows a Gaussian distribution with parameters $\theta_l = (\mu_l, \sigma_l)$, $\mu_l \in \mu_{x_i}$, $\sigma_l \in \sigma_{x_i}$, given the group label $x_i = l$.

$$p(y_i \mid x_i) = g(y_i; \theta_l) = \frac{1}{\sqrt{2\pi\sigma_l^2}} exp\left(-\frac{(y_i - \mu_l)^2}{2\sigma_l^2}\right) \tag{5}$$

Based on the conditional independence assumption, the joint likelihood probability is given by

$$P(Y \mid X) = \Pi_{i \in S} p(y_i \mid x_i) = \Pi_{i \in S}\left[\frac{1}{\sqrt{2\pi}} exp\Big(-\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} - log(\sigma_{x_i})\Big)\right] \tag{6}$$

which can be written as $P(Y \mid X) = Z^{-1} exp\big(-U(Y \mid X)\big)$, with the likelihood energy

$$U(Y \mid X) = \sum_{i \in S} U(y_i \mid x_i) = \sum_{i \in S}\left(\frac{(y_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} + log(\sigma_{x_i})\right) \tag{7}$$

We seek a labeling $\widehat{x}$ of an image, which is an estimate of the true labeling. According to the MAP criterion: $\widehat{x} = argmax_{x \in X}\{P(Y \mid X)P(X)\}$. The MAP estimation is equivalent to minimizing the posterior energy function $\widehat{x} = argmin_{x \in X}\{U(Y \mid X) + U(X)\}$. Application of the EM algorithm to solve the GMRF model results in (see [10] for details)

$$\mu_l^{(t+1)} = \frac{\sum_{i \in S} \mathrm{P}^{(t)}(l \mid y_i)\, y_i}{\sum_{i \in S} \mathrm{P}^{(t)}(l \mid y_i)} \tag{8}$$

$$\left(\sigma_l^{(t+1)}\right)^2 = \frac{\sum_{i \in S} \mathrm{P}^{(t)}\left(l \mid y_i\right)\left(y_i - \mu_l\right)^2}{\sum_{i \in S} \mathrm{P}^{(t)}\left(l \mid y_i\right)} \tag{9}$$

$$P^{(t)}\left(l \mid y_i\right) = \frac{g^{(t)}\left(y_i; \theta_l\right) P^{(t)}\left(l \mid x_{N_i}\right)}{p\left(y_i\right)} \tag{10}$$

where t denotes the iteration number.

Let $U^{*(t)} = \sum_{i \in S}\{U^{(t)}(Y \mid \widehat{X}) + U^{(t)}(\widehat{X})\}$, where $\widehat{X}$ is an estimate of X. The termination criterion of the EM algorithm is $\mid (U^{*(t+1)} - U^{*(t)})/(U^{*(t)}) \mid \leq \psi$.

The GMRF-EM approach can be repeated depending on the number of segmentation groups/labels. For each we determine the mutual information and derive $\partial MI(k) = MI(k) - MI(k-1)$ and $\partial^2 MI(k) = \partial MI(k) - \partial MI(k-1)$. For these first and second order derivative we introduce thresholds ε_1 and ε_2 respectively, which can be used as stopping criteria and will result in an estimation for the number of segmentation groups/labels. We use the two thresholds as $\partial MI(k+1) \leq \varepsilon_1$ and $\partial^2 MI(k+2) \leq \varepsilon_2$. The final segmentation result by using the GMRF-EM approach is the segmentation after k iterations.

3 Experiment and Discussion

We set κ in Eq. 2 equal to 20, the termination criterion of the EM algorithm $\psi = 0.001$. We create a synthetic images with Gaussian noise from 1% to 20% to estimate model parameters: resulting in $\delta = 0.1$, $\varepsilon_1 = 0.07$, $\varepsilon_2 = 0.02$. We used two types of images as shown in Fig. 3. According to the termination condition ($\partial MI < \delta$), the number of iterations are calculated as 10 and 13, respectively. From 2 to 5 groups, we can see the mutual information increasing rapidly, but after 5 groups this is reduced which denotes the number of groups are equal to five. According to our termination criterions ($\partial MI(k+1) \leq \varepsilon_1$ and $\partial^2 MI(k+2) \leq \varepsilon_2$), we can easily determine that the number of segmentation groups are five for both images which is the same as the ground truth, but the numbers of groups determined by the approach of Wang et al. are 6 and 7, respectively. We did run the experiment at variable noise levels and from 1%-19% noise the developed method produced the correct number of groups/labels. With increased noise levels the method will over or under estimate the number of groups/labels.

Our method was also tested on simulated T1 and T2 MR data from Brain Web [11]. In Fig. 4 non-brain tissue regions have been removed using a standard approach [12]. According to our termination criteria, the anisotropic diffusion iteration numbers are determined as 1 and 2 when using an anisotropic diffusion model and mutual information. In the second graphs the ∂MI curves on T1 and T2 are 0.0404 and -0.0308 at five groups, respectively, the $\partial^2 MI$ curves on T1 and T2 are 0.0089 and -0.0665 at six groups, respectively. According to our termination criterion, we can automatically determine both of the MR images can be segmented into 4 groups which are the same as the ground truth. We also compared our method with the method proposed by Wang et al., for which the

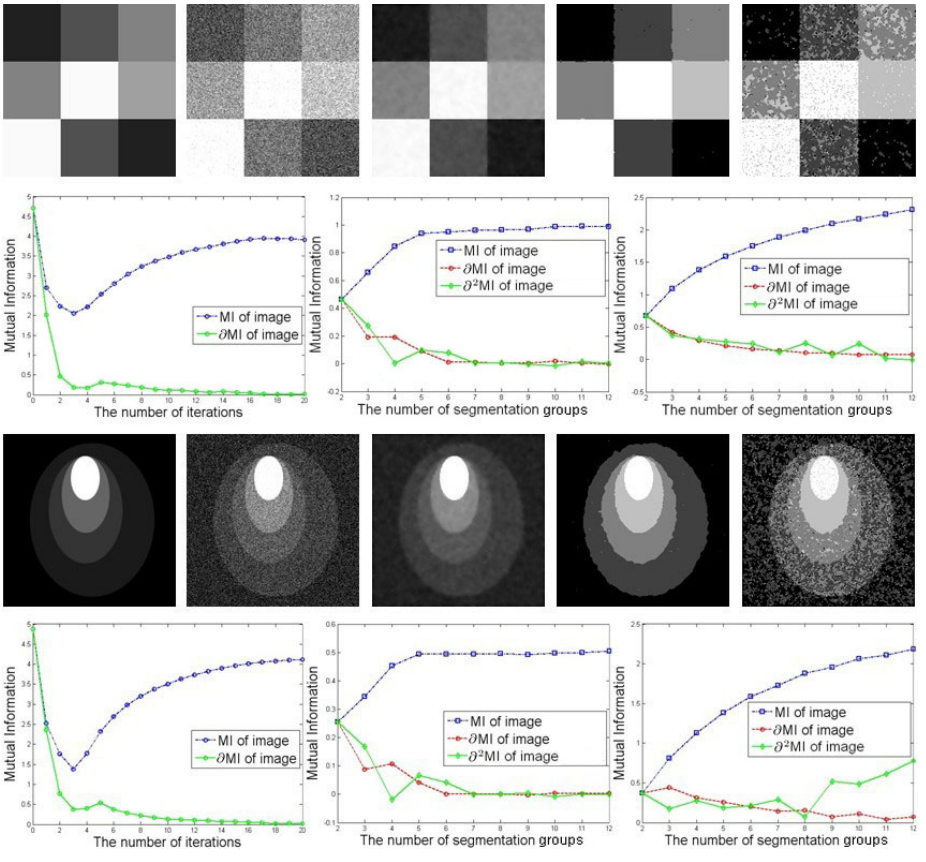


Fig. 3. Experimental results on synthetic images. The first and third rows show from left to right the original images, the original images with 15% Gaussian noise, the images after using the anisotropic diffusion model, the segmented images using our method, and segmentation results based on the GMRF-EM approach only. In the second and fourth rows, the first graphs show the mutual information (MI) curves of the filter result by using the anisotropic diffusion model for 20 interations, and the difference of mutual information curves (∂MI). The second graphs are the mutual information curves of numbers of segmentation groups and its derivative curves based on the proposed approach, while the third graphs show the equivalent based on the approach of Wang et al. [4].

results show the number of segmentation groups are determined as 6 groups for T1 and 3 groups for T2. We also segmented the images using only the GMRF-EM approach.

Note that the proposed approach has also been applied to some other images from the Brain Web and other sources, but the detailed results are not given here due to space limit. Nevertheless, the results showed that our approach can

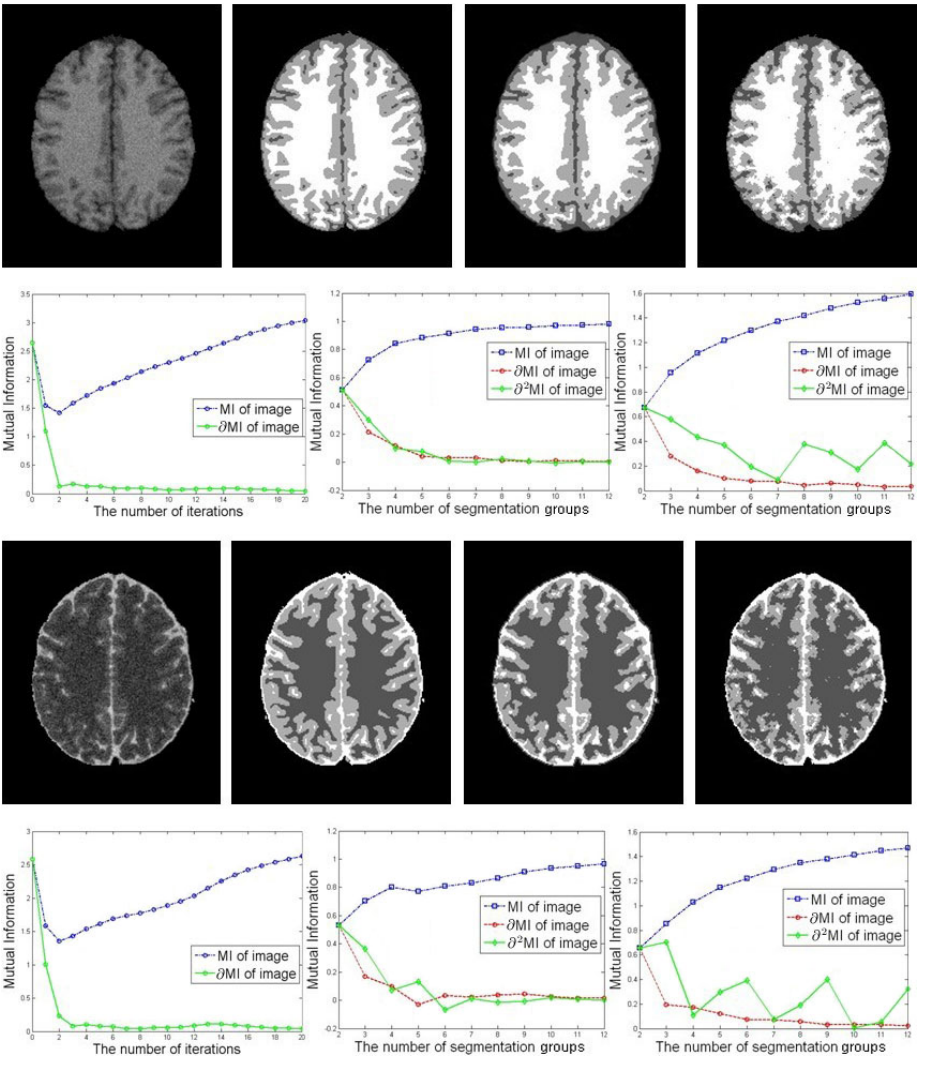


Fig. 4. Experimental results on simulated MR data. The first (T1) and third (T2) rows show from left to right the original images with 9% noise and 40% intensity inhomogeneity, the ground truth images, the segmented images using our method, and segmentation results based on the GMRF-EM approach only. The graphs in the second and fourth rows are arranged in the same manner as those in Fig. 3.

effectively determine the number of segmentation groups by comparing them with the ground truth, which are more promising than those of Wang's approach. In addition, as a side effect but importantly, the segmentation results obtained by our approach are better than those obtained by only using the GMRF-EM approach.

4 Conclusion

In this paper, we have proposed a method which automatically determines the number of segmentation groups/labels. This is achieved by initially using an anisotropic diffusion model for noise removal which automatically estimate the number of iteration by mutual information. Subsequently, mutual information and class-adaptive Gauss-Markov modeling are used to determine the number of groups/labels. We tested our method on simulated and multi-modal images. The evaluation showed that our method based on the difference of mutual information can effectively and automatically determine the number of segmentation groups. In addition, segmentation results are improved and the methodology leads to unsupervised automatic segmentation.

References

1. Bankman, I.: Handbook of Medical Image Processing and Analysis, pp. 71–258 (2008) ISBN 978-0-12-373904-9
2. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
3. Lv, Q.W., Chen, W.F.: Image segmentation based on mutual information. Chinese Journal of Computers 29(2), 296–301 (2006)
4. Wang, W.H., Feng, Q.J., Chen, W.F.: Segmentation of brain MR images based on the measurement of difference of mutual information and Gauss-Markov random field model. Journal of Computer Research and Development 46(3), 521–527 (2009)
5. Maes, F., Collignon, A.: Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging 16(2), 187–198 (1997)
6. Kim, J., Fisher, J.W., Cetin, M., Yezzi, A., Willsky, A.S.: Incorporating complex statistical information in active contour-based image segmentation. In: International Conference on Image Processing, pp. 655–658 (2003)
7. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(7), 529–539 (1990)
8. Tong, C., Wang, S.T., Man, L.H.: Improved image denoising method based on PDE. Journal of Computer Engineering and Applications 46(15), 176–179 (2010)
9. Wang, W.H., Feng, Q.J., Liu, L., Chen, W.F.: Segmentation of brain MR images through class-adaptive Gauss-Markov random field model and the EM algorithm. Journal of Image and Graphics 13(3), 488–493 (2008)
10. Markov, S.Z.L.: Random Field Modeling in Image Analysis, pp. 19–21. Springer, Heidelberg (2001)
11. Brain Web. Montreal Neurological Institute, McGill University (2006), `http://www.bic.mni.mcgill.ca/brainweb/`
12. Zhuang, A.H., Valentino, D.J., Toga, A.W.: Skull-stripping magnetic resonance brain images using a model-based level set. NeuroImage 32(1), 79–92 (2006)

Vitality Assessment of Boar Sperm Using N Concentric Squares Resized (NCSR) Texture Descriptor in Digital Images

Enrique Alegre[1], María Teresa García-Ordás[1], Víctor González-Castro[1], and S. Karthikeyan[2]

[1] Dep. of Electrical, Systems and Automatic Engineerings, Univ. of León, Spain
[2] Dep. of Electrical and Computer Engineering, Univ. of California Santa Barbara
enrique.alegre@unileon.es, karthikeyan@ece.ucsb.edu

Abstract. Two new textural descriptor, named N Concentric Squares Resized (NCSR) and N Concentric Squares Histogram (NCSH), have been proposed. These descriptors were used to classify 472 images of alive spermatozoa heads and 376 images of dead spermatozoa heads. The results obtained with these two novel descriptors have been compared with a number of classical descriptors such as Haralick, Pattern Spectrum, WSF, Zernike, Flusser and Hu. The feature vectors computed have been classified using kNN and a backpropagation Neural Network. The error rate obtained for NCSR with $N = 11$ was of 23.20% outperforms the rest of descriptors. Also, the area under the ROC curve (AUC) and the values observed in the ROC curve indicates the performance of the proposed descriptor is better than the others texture description methods.

Keywords: texture descriptors, semen assessment, classification, digital image analysis.

1 Introduction

In this work we have assessed the vitality of boar sperm cells using a new texture descriptor that allow us to classify each spermatozoon head as dead or alive. Currently, this assessment is carried out using fluorescence microscopy and stains and is not possible to perform it with phase contrast microscopes without stains. It means the assessment is a time consuming process and also requires the laboratory to have access to expensive equipment.

The sperm assessment is a very important problem for the porcine industry. In most of the countries there is a big demand of alimentary products obtained from pig's meat so there is lots of companies trying to obtain as better as possible pork flesh and, at the same time, at the lower price available. The way to do it is by selecting the semen used in artificial insemination i.e. to assess the semen of the donor boars, picking the best specimens up and use only the best ones.

For several decades the CASA (Computer-Assisted Semen Analysis) systems have been used for assessing the seminal quality. Currently these systems analyse

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 540–547, 2011.

the mobility, concentration and provide some simple geometric measures of the spermatozoa's head to characterize abnormal head shapes, obtaining an assessment of the studied sample based on that values. But there are three valuable criteria, used by veterinary experts, that theses systems are not yet able to analyse automatically, as are the integrity of the acrosomal membrane, the number and presence of proximal and distal droplets and the vitality of the sample based on the presence of dead or alive spermatozoa.

A number of works have addressed some of the problems related to the semen analysis using digital image processing. Most of them use CASA systems for evaluating the sperm motility [1] or for studying the relationship among sperm cell motility patterns, morphology and boar fertility [2,3]. Others researches have developed new methods to characterize the sperm shape by using spectral approaches [4,6], or they have been looking for subpopulations [7] using shape descriptors of the spermatozoa head. There is only little work addressing the evaluation of the membrane integrity, in this case using texture descriptors [8] and, as far as we know, there is not any work published that assess the vitality of a sample classifying the spermatozoa heads as dead or alive.

The rest of the paper is organized as follows: section 2 explains how the images have been captured, segmented and preprocessed. In section 3 the features vectors of classical texture descriptors used are detailed and the new proposed descriptors are explained. Section 4 indicates what classifiers have been used. In section 5 the results obtained with the proposed and classical descriptors are shown and, finally, the main conclusions are presented in section 6.

2 Image Dataset and Preprocessing

2.1 Image Acquisition

The images used have been captured in CENTROTEC, an Artificial Insemination Centre that is an University of Leon spin-off. The semen were obtained from boars of three different races- Piyorker, Large White and Landrace. 450 pairs of images have been obtained using a Nikon Eclipse microscope and a Baster A312f camera of progressive scan. Each pair contains an image in positive phase contrast and a fluorescent image obtained using two different stains, propidium iodide (PI) that dyes dead spermatozoa as red and dichlorofluorescein (DCF)

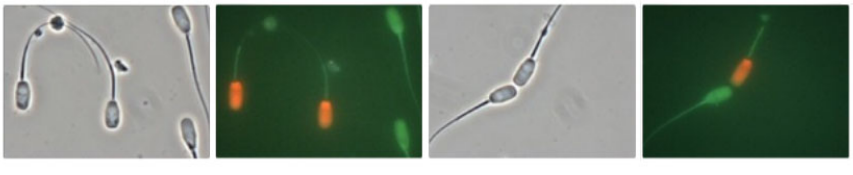


Fig. 1. Figures shown images in positive phase contrast and images with fluorescent stains respectively. Alive spermatozoa are coloured in red and dead ones in green.

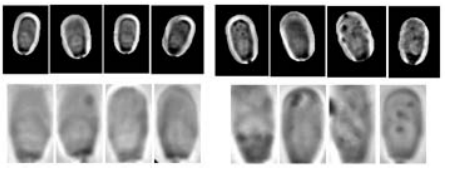


Fig. 2. Upper images represents masked images, and grey scale spermatozoa are shown in the lower images (alive in the left side and dead in the right)

for turning green the alive spermatozoa, see fig. 1. Further information about the sample preparation can been found in [10]. We have captured these pairs of images because we have used the phase contrast images for developing and testing the texture descriptors evaluated on the proposed method. For this purpose, first we need a ground truth that we have obtained using the red and green colours of the fluorescent images for labelling the grey level ones.

2.2 Segmentation

Every spermatozoa head from the phase contrast images have been automatically segmented. First, the image regions containing the heads have been detected by thresholding and later head regions are cropped. Then, the heads have been segmented using the method presented in [9]. The binary image obtained have been used for masking the region cropped previously so the images that are used in the description step have a black background. The original grey level of the spermatozoa heads and the masked images can be seen in figure 2. Later, all the spermatozoa heads have been cropped using its bounding box, then resized to 108×63 pixels, rotated placing the longest side in vertical. Using the location of the tail, the apical part have been placed at the upper side and the tail's insertion at the lower side (figure 2). Finally, 472 images of alive's heads and 376 images of dead's heads have been obtained.

3 Texture Description of the Spermatozoa Heads

3.1 Texture Oriented Descriptors

Our first approach has been to try a number of descriptors for classifying the spermatozoa as dead or alive using the different texture distributions present in their heads. We also have proposed two new texture descriptors, called NCSR (N Concentric Squares Resized) and NCSH (N Concentric Squares Histogram) that we will explain in next section. The descriptors considered have been Hu, Flusser, Zernike, Haralick, some statistical descriptors, Wavelet Statistical Features and several variations of the Pattern Spectrum.

The following feature vectors have been used. The Hu descriptors used have been the seven normalized moments defined for Hu [13]. Likewise, we have

computed the six invariant affine moments proposed by Flusser [12]. The Zernike moments vector contains the nine first Zernike moments until order four [15]. In the Haralick's feature vector we have include the value of the first thirteen values computed over the grey level co-ocurrence matrix, with distance 1 and the average of the directions 0, 45, 90 and 135 degrees. The thirteen descriptors [14] are: energy, contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, first measure of correlation and second measure of correlation. From the histogram of each cropped and masked region we have computed four statistical values: average grey level, average contrast, measure of uniformity, entropy that make up the statistical feature vector used.The WSF feature vector is the proposed by Arizazhagan and Ganesan [11]. It is a 24-D features vector containing the mean and the standard deviation of the twelve images from the three first sub-bands of the wavelet decomposition. Finally, three different Pattern Spectrum vectors have been computed using the Maragos's proposal [16], with 10 elements both without normalization and normalizing the vectors.

3.2 Proposed Descriptors: N Concentric Squares (NCS)

The N Concentric Squares (NCS) descriptor gathers the grey levels along N equidistant squares that are concentric to the bounding box of the interest image. The value assigned to N sets the number of squares obtained. As N increases, the squares will get closer. All the N squares are concatenated making up one vector whose longitude depends on the size of the image and the number of squares extracted. As we have been working with registered images, all the images are vertical and have the same size. The first part of the vector is the horizontal segment closer to the left upper corner of the image, the second part is the vertical segment of the outer NCS that is closer to the right side of the image, the third part is the horizontal segment at the bottom and the last part of the outer square is the vertical segment closer to the left side of the image. The four segments are concatenated constituting the vector that comes from the outer square. Later, the same process is followed with the rest of the inner squares

Fig. 3. Spermatozoon image is divided in N squares. In this case, N=7.

concatenating all the resulting vector into one. The grey levels gathered along this vector are the maximum value in the 3×3 neighbourhood of each position. The distances between squares are given by the equation 1.

$$NpbHs = \frac{\frac{nRows}{2}}{N+1}, N = 1, 2, 3, \ldots; NpbVs = \frac{\frac{nCols}{2}}{N+1}, N = 1, 2, 3, \ldots \quad (1)$$

where NpbHs and NpbVs are the number of pixels between Horizontal and Vertical segments, respectively; and nRows, nCols are the number of rows and columns of the image.

The first descriptor proposed, called NCS Resized (NCSR), takes the previous vector and resizes it to a new vector of 100 elements by interpolation using a method based on the Fourier transform. The second descriptor proposed, figure 4, called NCS Histograms (NCSH), computes a histogram with 20 bins for each of the concentric squares and then it concatenates all the N vectors yielding a new vector of length $20 \times N$.

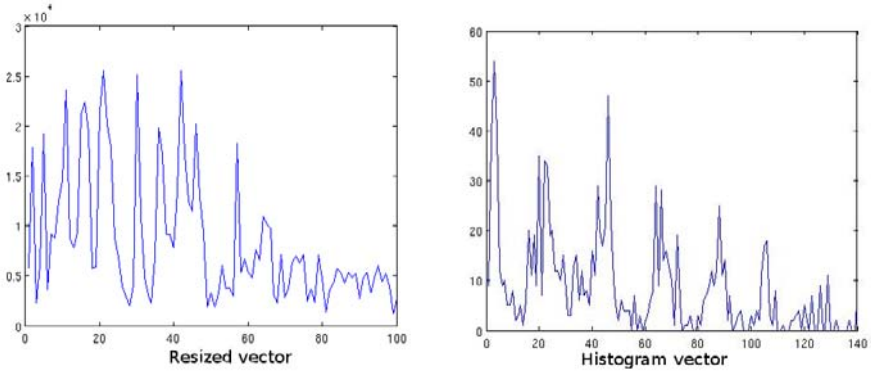


Fig. 4. Examples of NCSR (left) and NCSH (right) vectors

4 Classification

First, we have carried out a kNN classification with k values $1, 3, 5, 7, 9, 11$ and 13. In most cases the best hit rate have been obtained with $k = 9$ and $k = 11$.

Later, we have also classified the data using a Neural Network looking for robust classification. The NN used was trained using backpropagation. It has one hidden layer and a logistic sigmoid transfer function for the hidden and the output layer. Learning was carried out with a momentum and adaptive learning rate algorithm. Data were normalized with zero mean and standard deviation equal to one. Classification was carried out by means of 10-fold cross validation with several different combinations of neurons in the hidden layer and training cycles in order to find out the optimal configuration in terms of accuracy. As there are some authors [17] from the machine learning community who claim that the hit rate is not the most suitable option for illustrating the performance

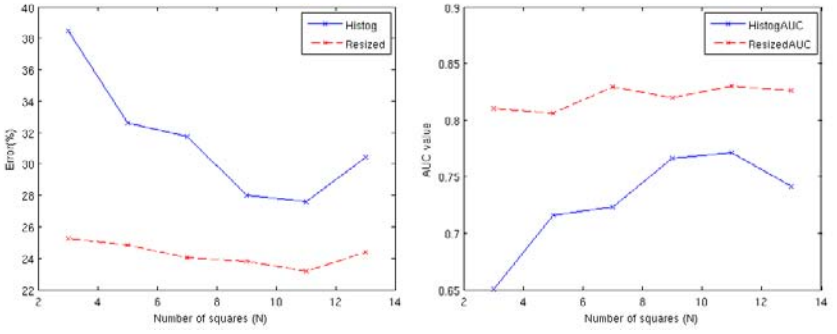


Fig. 5. Descriptor error and AUC values for different values of N

of a classifier, we have also obtained ROC curves and we present in the results section the area under that curve (AUC) as a more robust tool for comparing the computed descriptors.

5 Results

5.1 Best N Value for NCSR and NCSH

The errors obtained classifying with NCSR and NCSH when different values of N used, and classifying with NN, can be seen in figure 5. It is possible to appreciate that the error decreases with higher values of N, from $N = 1$ to $N = 11$. The behavior for the AUC is the opposite, also obtaining the best values for $N = 11$. In this figure is clear the performance of NCSR is much higher than NCSH so our choice for this problem has been NCSR with $N = 11$.

5.2 Errors and ROC Curves

Table 1 summarize the errors obtained using kNN, Neural Networks and the AUC value for each descriptor. The table is arranged, from left to right, in ascending order using the global error rate yielded by the NN. This order is almost the same, but descending, when the AUC criteria is used. It is possible to observe the errors provide by kNN do not correspond exactly with the yield by the NN classifier but, in both cases, one of the proposal descriptors, NCSR, outperforms to the rest of descriptors. The same behaviour can be seen in the

Table 1. Errors using kNN and NN classifiers. It also shows the AUC value.

	NCSR	Haral13	PS10N	PS20N	PS10	NCSH	FlussGrey	Zern	HuGrey	HuBW	FlussBW
NNError	23.20%	24.07%	24.93%	25.14	25.24%	27.63%	28.63%	30.55%	32.44%	39.50%	40.47%
AUC	0.8299	0.8297	0.8092	0.8053	0.8032	0.7707	0.7751	0.7453	0.7318	0.6096	0.5872
kNNError	25.22%	31.30%	28.09%	30.84%	28.46%	27.09%	44.57%	32.21%	31.59%	43.82%	44.44%

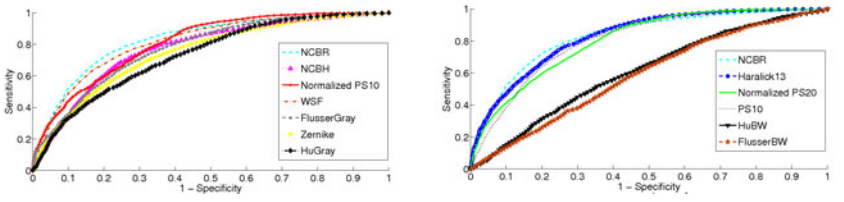


Fig. 6. ROC Curve for descriptors comparison

Receiver Operating Characteristic (ROC) curve 6. Whereas there are several descriptors with a high accuracy, as the Haralick13, the WSF, the proposed one with $N = 11$ surpasses the rest methods in the interval of specificity $[0.1, 0.4]$, been similar or slightly lower to Haralick13 in the other parts of the curve. That values, jointly with the error rates obtained, allow us to say that the proposed descriptor has a better performance than the other ones in this context.

6 Conclusions

We have proposed two new texture descriptors named NCSR (N Concentric Squares Resized) and NCSH (N Concentric Squares Histogram) for describing the texture present in the images taken of boar spermatozoa heads. A first evaluation of which one was the best value of N for this problem was carried out. Later, we have computed these descriptors and we have use them for classify the spermatozoa heads as dead or alive. The results obtained have been compared with a number of classical texture descriptors as Haralick, Pattern Spectrum, WSF, Zernike, Flusser and Hu, using both kNN and a backpropagation Neural Network. The results illustrate that one of the proposed descriptors, NSCR, outperforms the others using both the error rate criteria and the accuracy seen at the ROC curve.

Acknowledgement

This work has been supported by grants DPI2009-08424 and PR2009-0280 from the Spanish Government. Thanks to CENTROTEC, to professor Manjunath and to all the people from the Vision R. Lab (Univ. of California, Santa Barbara).

References

1. Contri, A., Valorz, C., Faustini, M., Wegher, L., Caluccio, A.: Effect of semen preparation on CASA motility results in cryopreserved bull spermatozoa. Theriogenology 74(3), 424–443 (2010)
2. Didion, B.A.: Computer-assisted semen analysis and its utility for profiling boar semen samples. Theriogenology 70(8), 1374–1376 (2008)

3. Verstegen, J., Iguer-Ouada, M., Onclin, K.: Computer assisted semen analyzers in andrology research and veterinary practice. Theriogenology 57(1), 149–179 (2002)
4. Beletti, M., Costa, L., Viana, M.: A spectral framework for sperm shape characterization. Computers in Biology and Medicine 35(6), 463–473 (2005)
5. Beletti, M.E., Costa Lda, F.: A systematic approach to multispecies sperm morphometric characterization. Anal. Quant. Cytol. Histol. 25(2), 97–107 (2003)
6. Severa, L., Máchal, L., Svábová, L., Mamica, O.: Evaluation of shape variability of stallion sperm heads by means of image analysis and Fourier descriptors. Animal Reproduction Science 119(1-2), 50–55 (2010)
7. Thurston, L., Watson, P., Mileham, A., Holt, W.: Morphologically distinct sperm subpopulations defined by Fourier shape descriptors in fresh ejaculates correlate with variation in boar semen quality following cryopreservation. Journal of Andrology 22(3), 382–394 (2001)
8. Alegre, E., Biehl, M., Petkov, N., Sánchez, L.: Automatic classification of the acrosome status of boar spermatozoa using digital image processing and LVQ. Computers in Biology and Medicine 38(4), 461–468 (2008)
9. González-Castro, V., Alegre, E., Morala-Arguello, P., Suarez, S.A.: A combined and intelligent new segmentation method for boar semen based on thresholding and Watershed Transform. International Journal of Imaging 2(S09), 70–80 (2009)
10. Sanchez, L., Petkov, N., Alegre, E.: Statistical approach to boar semen evaluation using intracellular intensity distribution of head images. Cellular and Molecular Biology 52, 38–43 (2006)
11. Arivazhagan, S., Ganesan, L.: Texture classification using wavelet transform. Pattern Recognition Letters, 1513–1521 (June 2003)
12. Flusser, J.: Moment invariants in image analysis. Proceedings of the World Academy of Science, Engineering and Technology 11, 196–201 (2006)
13. Hu, M.-K.: Visual pattern recognition by moment invariants. IRE Trans. Inform. Theory 8, 179–187 (1962)
14. Haralick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE, 45–69 (1978)
15. Liao, S., Pawlak, M.: Image analysis with Zernike moment descriptors. In: IEEE Canadian Conference on Electrical and Computer Engineering, St. Johns, Canada, vol. 2, pp. 700–703 (May 1997)
16. Maragos, P.: Pattern spectrum and multiscale shape representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(7), 701–716 (1989)
17. Provost, F.J., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the 15 th International Conference on Machine Learning, pp. 445–453 (1998)

Filled-in Document Identification Using Local Features and a Direct Voting Scheme*

Joaquim Arlandis, Vicent Castello-Fos, and Juan-Carlos Perez-Cortes

Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{arlandis,vcastello,jcperez}@iti.upv.es

Abstract. In this work, an approach combining local representations with a direct voting scheme on a k-nearest neighbors classifier to identify filled-in document images is presented. A document class is represented by a high number of local feature vectors selected from its reference image using a given criterion. In the test phase, a number of vectors are equally selected from an image and used to classify it. The experimental results show that the parameterization is not critical, and good performances in terms of error-rate and processing time can be obtained, even though the test documents contain a large proportion of filled-in regions, obviously not present in the reference images.

Keywords: Document identification, filled-in documents, local features, k-nearest neighbors.

1 Introduction

In document classification, a significant amount of effort has been traditionally devoted to develop approaches focused on clustering documents having a certain degree of semantic similarity, as belonging to the same class or category. However, in some applications, like those dealing with data digitalization and extraction, among others, the classes must be defined to represent particular types of documents. In that case, the task is commonly known as document identification, and clustering methods are therefore not adequate. In most of these applications, the identification of a document image is required as a first step, previous to any other specific process.

Also, in many applications, methods for automatic processing of filled-in documents such as invoices, forms, passports and other business, medical, or personal documents are required. In this case, large handwritten, typed or stamped regions in the documents can be found, and that can radically change the values of

* Work partially supported by the Spanish MICINN grants TIN2009-14205-C04-02 and Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and by IMPIVA and the E.U. by means of the ERDF in the context of the R+D Program for Technological Institutes of IMPIVA network for 2010 (IMIDIC-2010/191).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 548–555, 2011.

global image features that are often extracted from the images. In that case, the use of local features can be a good solution to specifically locate the pre-printed contents in a document identification process.

In this work, we face the issue of identifying document images from multiple application domains, regardless of their layout, structure, text and non-text contents as well as different amount of filled-in contents. To achieve this goal, an approach combining local representations with a direct voting scheme on a k-nearest neighbors classifier is used. The experiments carried out show the robustness of the approach, taking into account that no filled-in contents or representations are used in the training phase.

2 Related Work

Many kinds of features have been used for document image classfication. They are related to document layouts, texture primitives, string and character recognition, shape codes, frame detection, salient visual features, global image transformations and projections, or semantic block structures detection.

In the scope of Information Retrieval, when no filled-in contents exist, document identification can be seen as a duplicate detection task [3]. In this case, the approaches have to tackle with differences among document instances, like resolution, skew, distortions and image quality, speed and robustness, as well as, handling very large databases.

Most works dealing with filled-in documents are related to form identification. Many of them are based on analyzing global and local structures [4], [12], [7]. Structural features are usually limited to documents having frames, cells, lines, blocks, or similar items, and it may not help with different types of documents having similar structure. Other works rely on using character and string codes to achieve the document identification [14], as well as, on computing pixel densities from image regions [5]. Within form-type documents, specific applications are addresed to coupons [9], banking [11] or business [15] form identification.

The purpose of this work is to deal with the task of classifying documents with total flexibility of designs, layouts, sizes, and amount of filled-in contents in an efficient way. Hence, the above aproaches may not be appropiate as they use global features, or they are focussed on specific document types, or they are not able to handle filled-in contents.

3 Approach

Typically, document images are composed of white background pixels and black foreground pixels, but other combinations like gray, colored, or heterogeneous backgrounds and foregrounds can be found. The foreground is mostly composed of text (in many cases having different appearances like typed fonts, handwriting styles, case letters, bolded text, sizes, etc.), although other objects like images, graphics, logos, or frames are frequent, too. Usually, the text areas also include background patterns interleaved, and some background often covers most of the surface of a document.

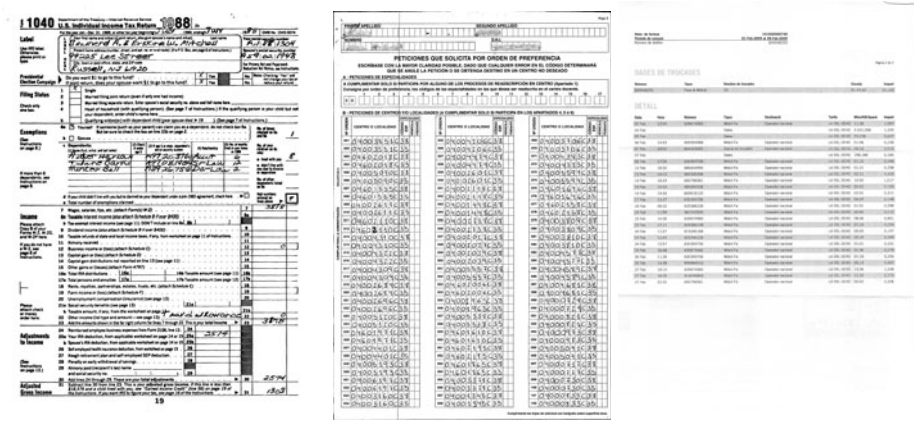


Fig. 1. Document examples. On the left, a form where the static contents cover most of the document. On the center, a form page with a large number of cells. On the right, a business document having very scarce structural patterns and static contents (located in the header), while the variable contents can be any amount of text lines.

A filled-in document can be seen as an image having static (pre-printed) and variable contents (machine printed or handwritten). Under this assumption, a type or class of document is defined as the set of images having different static content from the other classes and an approximately common intra-class static content. The variable content (filled in), however, can vary wildly for different documents within a class. In Figure 1, some filled-in document types are shown.

In this work, the voting approach used in [13] for a face recognition task has been applied to document identification. In our case, a number of known document classes are represented by a set of reference images, and the identification of a test document relies on the combination of the evidence contributed by multiple local features. One or more reference images can be used to represent each class, but only the static part of the document must appear in them.

In the training phase, a number of small sub-images from each class are selected from its reference images. The selection criterion can be made on the basis of image contrast, or variance, or on more complex operators, like corner detection or specific filters. The goal of this selection is to retain the areas with clear graphical content, as text or any other possibly discriminative pattern, avoiding uniform areas or uninformative background regions.

In the test phase, also a number of sub-images from a test image are selected, using the same filter as in the training phase. After feature extraction, each vector is classified according to the k-nearest neighbors rule, and finally, a voting scheme is applied to find the class with the largest number of votes. Adequate sub-sampling can be applied in both phases to reduce the computational burden while assuring that any selected sub-image having static contents from a test image, corresponds to some sub-image included in the training set.

3.1 Feature Extraction and Classificacion

The first operation applied to the sub-images was a conventional preprocess to normalize the local contrast. The resulting vectors were transformed by means of principal component analysis (PCA) and only the most significant components retained, resulting in a low dimension feature vector for each sub-image. In section 4, more details and the parameters used in these processes are given.

To increase the information content and, potentially, the discriminative power of the feature vectors, a controlled amount of (non-local, or *global*) geometric information can be taken into account. In this case, the coordinates of the center of each window with respect to the whole image are added as two new components before classification, and normalized to have a standard deviation related by a factor to the standard deviation of the first PCA component. This factor can be seen as a weight to tune the effect of the global features with respect to the rest of the components (local features).

As detailed in [13], the classification procedure used is related to the methods often referred to as *direct voting schemes* [8] which can be included under the more formal statistical framework of *classifier combination* [6]. Let Y be a test image. Given a prototype set representing the reference classes, Y can be optimally classified in a class $\hat{w}$ having the maximum posterior probability,

$$\hat{w} = \arg\max_{1 \leq j \leq d} P(\omega_j | Y) \tag{1}$$

Given a set of feature vectors $m_Y = \{y_1, \ldots, y_m\}$ extracted from Y, the classifier can be rewritten as a linear combination of m_Y classifiers [6] each one from every feature vector of Y,

$$\hat{w} = \arg\max_{1 \leq j \leq d} \sum_{i=1}^{m_Y} P(\omega_j | y_i), \tag{2}$$

which corresponds to the so called *sum rule* often used in practical applications. Note that feature vectors coming from sub-images having variable contents, or any other kind of noise, introduce "noisy" feature vectors to the classifier which will generally be poorly estimated with low probabilities. The sum rule provides a way of smoothing the effect of these low probabilities. Hence, the noisy vectors actually will have a beneficial effect because misclassified vectors will tend to distribute among different classes.

Several variants of the k-nearest neighbors rule can be used to estimate the posterior probabilities of the expression 2. Assuming that the number of vectors of each class in the prototype set is fixed according to the *a priori* probabilities of the classes, the following estimate can be used,

$$\hat{P}(\omega_j | y_i) = \frac{k_{ij}}{k},$$

where k_{ij} is the number of neighbors of y_i belonging to the class ω_j, and the classification rule becomes,

$$\hat{w} = \arg\max_{1 \le j \le d} \sum_{i=1}^{m_y} k_{ij}$$

That is, a class $\hat{w}$ with the largest number of votes accumulated over all vectors extracted from the test image is selected. The number of test vectors used can be adjusted empirically.

Note that the selection criterion, as proposed in section 3, is applied within a class, and, in spite of adding location components to the training feature vectors, it does not guarantee that similar sub-images from different classes could be found at the same location. However, it is expected that using a large number of vectors per class will decrease the overall probability of multiple matchings of test vectors on the same "erroneous" class (we call *casual matchings*).

Nevertheless, among very similar documents, a high number of casual matchings can occur, and this could affect the performance of the method. To solve this issue, the use of a two-level hierarchical classifier could be used: very similar classes can be grouped in the same category, and if the class with the highest posterior probability is a category, then a fine-grained method, like the one described in [1], can be applied to discriminate classes within that category.

An inherent characteristic of the method proposed is that a high number of vectors could be needed to represent a document class. Also, in practice, a high number of classes could be required to be identified by a system. As a consequence, the use of an efficient technique to find the k-nearest neighbors of a vector in very large sets is imposed. In the experiments, a *kd*-tree data structure [2] has been used, which provides fast approximate k-nearest neighbors search in $\theta(log(N))$, where N is the number of vectors of the prototype set.

4 Experiments

Experiments were carried out on a database consisting of 683 binary and gray-scale documents belonging to 69 different types, most of them DIN A4 size, and having different amount of filled-in contents. It includes the well-kown NIST SD6 database [10] (20 form faces) and documents from different real office workflows consisting of business invoices, bank documents, personal documents, and a variety of forms. In Figure 1, three examples of document types used are shown.

One reference image per class was selected for the training set, while the rest were used for test (6 to 10 images per class). The reference images have been checked and, if necessary, manually cleaned to remove filled-in contents.

Several preprocesses were applied to the whole database. Because of the scanning, some documents revealed different rotations, and a correction was applied to them. Then, to overcome the drawback of different acquisition resolutions found in different instances of a class, each image was resized to the same pixel surface (equivalent to an A4 300dpi area) preserving its original aspect ratio. Finally, a smoothing filter using a 5x5 convolution matrix, as well as, contrast normalization and thresholding were applied to help in the matching process.

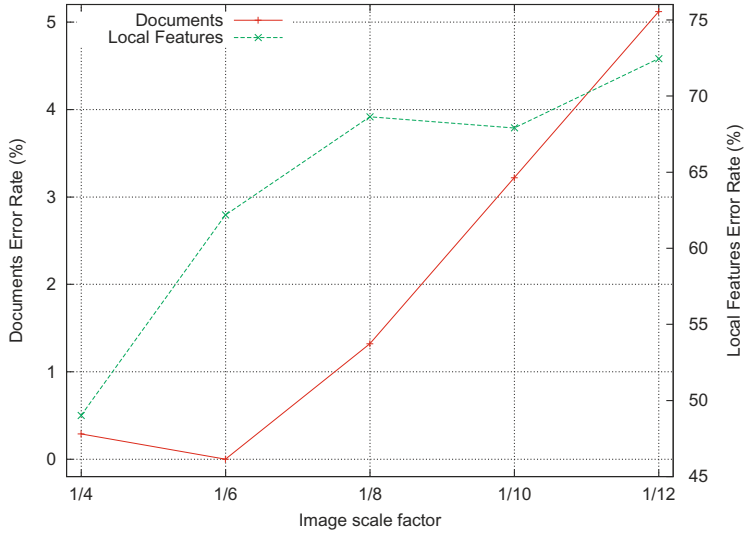


Fig. 2. Document-level (left axis) and LF-level (right axis) error rates for different scale reductions. The best combination of the remaining parameters is shown.

4.1 Parameter Optimization

Several parameters are involved in preprocessing, selection and feature extraction. To see how the system performance is affected, comprehensive tests on combining the following parameters were performed:

- *Window size.* Several window sizes, potentially enclosing sub-images having different number of text characters, or other objects, were tested. Windows around 80 pixels wide and 30 pixels high provided the best results.
- *Scale.* The images were scaled by different factors from 1/4 to 1/12, and the window size was scaled as well.
- *Number of local features (LF).* A number of non-overlapped sub-images from each class having the highest contrast (variance) are selected for training and test. Values from 100 to 1000 were tested. In practice, to reduce the number of computations in the search for candidates, subsampling was applied in the training as well as in the test phase, and additional sub-images extracted from a neighborhood window of size equal to the subsampling step around the previous selected sub-images were added to the training set, in order to assure that any selected sub-image having static contents from a test image is included in the training set.
- *Dimensionality reduction.* A PCA transformation of the local feature vectors and a dimensionality reduction to 15 components were applied.
- *Global features weight.* The coordinates of the center point of each window were added to its feature vector after the dimensionality reduction. The values of these new features were normalized to the standard deviation of

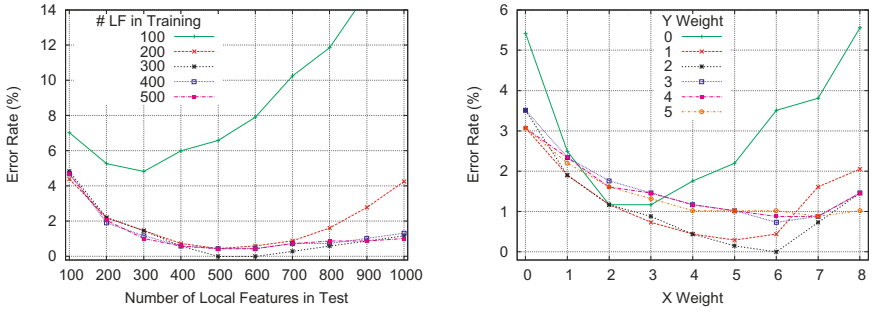


Fig. 3. Document-level error rate as a function of: (left) the number of local features used in training and test; and (right) the coordinate weight.

the first PCA component and multiplied by a weight factor to tune their effect with respect to the rest of the components. Combinations of weight factors (α_x, α_y) from 0 to 8 were tested.

A fast classifier was implemented using an approximate *kd*-tree [2] search. The nearest neighbor of each of the 17-dimensional (15+2) vectors extracted from a test image was obtained using a value of $\epsilon = 2$, and the test image was classified by means of the *sum rule* described in section 3.1.

Figure 2 shows the results for different scale factors and the best combination of the remaining parameters. A 0% error rate was achieved at document-level with a factor 1/6, with 300 vectors for training and 500 vectors for test, and $(\alpha_x, \alpha_y) = (6, 2)$. The processing time measured for this setting was 4.5 documents/s on an AMD 64-bits 4 CPU 3 GHz machine. Notice that, in spite of the error rates at LF-level are over 50% in most of the cases, very lower error rates at document-level can be achieved because the wrong votes get distributed among several classes. In most cases, misclassified vectors included filled-in contents.

The number of local features and the weight of the global features are parameters strictly related to the approach used. Hence, an analysis of the results on varying both parameters while fixing the remaining ones as in the 1/6 scale factor experiment of Figure 2 is presented. Figure 3 (left) shows the error rate as a function of the number of vectors taken for training and test. Usually, the best combinations are found using a few more test vectors than the number used in training, and no big differences are found for different settings. Figure 3 (right) shows that the window location is a very discriminant information. The weight combination $(\alpha_x, \alpha_y) = (6, 2)$ led to a 0% error rate. This means an improvement of 5% (33 documents) on the recognition rate compared to not taking into account the location as a global feature. It also should be noted that the x coordinate is significantly more discriminant than the y coordinate. This is probably because of the higher translations on the y axis present in some documents due to the mechanical tolerances of the scanning process.

5 Conclusions

An approach combining local representations with a direct voting scheme on a k-nearest neighbors classifier to identify filled-in document images has been presented. Comprehensive experiments have been carried out on a database consisting of 69 document types to determine performances as a function of the parametrization used. It included a variety of business, bank, and personal documents, as well as forms, having different amount of filled-in contents. The results show that a 0% error rate can be achieved in the data set used, while the procesing time is about 4.5 documents per second.

References

1. Arlandis, J., Perez-Cortes, J.C., Ungria, E.: Identification of very similar filled-in forms with a reject option. In: ICDAR, pp. 246–250 (2009)
2. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching. Journal of the ACM 45, 891–923 (1998)
3. Doermann, D.: The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding 70(3), 287–298 (1998)
4. Fan, K.-C., Chang, M.-L., Wang, Y.-K.: Form document identification using line structure based features. In: ICDAR, pp. 704–708 (2001)
5. Heroux, P., Diana, S., Ribert, A., Trupin, E.: Classification method study for automatic form class identification. In: Proc. 14th Int. Conf. on Pattern Recognition, ICPR 1998, pp. 926–928 (1998)
6. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20(3), 226–239 (1998)
7. Mandal, S., Chowdhury, S.P., Das, A.K., Chanda, B.: A hierarchical method for automated identification and segmentation of forms. In: International Conference on Document Analysis and Recognition, pp. 705–709 (2005)
8. Mohr, R., Picard, S., Schmid, C.: Bayesian decision versus voting for image retrieval. In: Sommer, G., Daniilidis, K., Pauli, J. (eds.) CAIP 1997. LNCS, vol. 1296, pp. 376–383. Springer, Heidelberg (1997)
9. Nagasaki, T., Marukawa, K., Kagehiro, T., Sako, H.: A coupon classification method based on adaptive image vector matching. In: 18th International Conference on Pattern Recognition, pp. 280–283 (2006)
10. Dimmick, D.L., Garris, M.D.: Structured Forms Database 2, NIST Special Database 6 Technical Report and CD-ROM, National Institute of Standards and Technology (1992)
11. Ogata, H., Watanabe, S., Imaizumi, A., Yasue, T., Furukawa, N., Sako, H., Fujisawa, H.: Form-type identification for banking applications and its implementation issues. In: DRR, pp. 208–218 (2003)
12. Ohtera, R., Horiuchi, T.: Faxed form identification using histogram of the hough-space. In: International Conference on Pattern Recognition, vol. 2, pp. 566–569 (2004)
13. Paredes, R., Pérez-Cortes, J.C., Juan, A., Vidal, E.: Local representations and a direct voting scheme for face recognition. In: PRIS, pp. 71–79 (2001)
14. Sako, H., Seki, M., Furukawa, N., Ikeda, H., Imaizumi, A.: Form reading based on form-type identification and form-data recognition. In: International Conference on Document Analysis and Recognition, vol. 2, p. 926 (2003)
15. Ting, A., Leung, M.: Business form classification using strings. In: ICPR 1996, pages II: 690–694 (1996)

Combining Growcut and Temporal Correlation for IVUS Lumen Segmentation

Simone Balocco[1,2], Carlo Gatta[1,2] , Francesco Ciompi[1,2], Oriol Pujol[1,2], Xavier Carrillo[3], Josepa Mauri[3], and Petia Radeva[1,2]

[1] Computer Vision Center, 08193 Bellaterra, Spain
[2] Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via 585, 08007 Barcelona, Spain
[3] Hospital universitari Germans Trias i Pujol Badalona

Abstract. The assessment of arterial luminal area, performed by IVUS analysis, is a clinical index used to evaluate the degree of coronary artery disease. In this paper we propose a novel approach to automatically segment the vessel lumen, which combines model-based temporal information extracted from successive frames of the sequence, with spatial classification using the Growcut algorithm. The performance of the method is evaluated by an *in vivo* experiment on 300 IVUS frames. The automatic and manual segmentation performances in general vessel and stent frames are comparable. The average segmentation error in vessel, stent and bifurcation frames are 0.17 ± 0.08 mm, 0.18 ± 0.07 mm and 0.31 ± 0.12 mm respectively.

1 Introduction

Atherosclerosis is a progressive disease affecting arterial blood vessels. The accurate assessment of luminal area is one of the most important guiding parameters during percutaneous intervention. It allows in fact to evaluate the local amount of stenosis, thus determining the degree of Coronary Artery Disease (CAD). The clinical inspection of coronary arteries is in general performed through Intravascular Ultrasound (IVUS) which is a catheter-based imaging technique providing real-time hi-resolution cross-sectional sequences. The IVUS sequence (pullback) can be represented as $I(x, y; t)$ where x and y are the spatial coordinates of the image (Figure 1), and t the frame number of a pullback.

Several automatic methods for segmentation the arterial lumen of IVUS images has been proposed so far. Recently a robust approach for lumen segmentation, based on the RF signal processing has been presented. [1]. This method requires dedicated hardware or a RF export device which is not commercially available and widespread in hospitals.

A second category of algorithms aims at segmenting the luminal contour using region growing techniques on the gray scale reconstructed images [1–4]. Those methods exploit probabilistic approaches based on active shape models. However, when blood and plaque present similar echogenicity, as in the case of Figure 1, the methods based on the local image statistics, become less robust and occasionally fail to identify the contour of the lumen. The main difficulty

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 556–563, 2011.

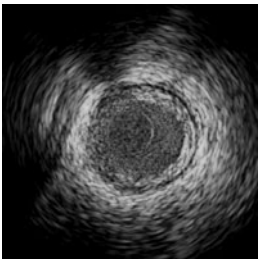


Fig. 1. IVUS image of a pathological patient artery. The plaque contour can hardly be defined on the left region because of the similar echogenicity with the blood.

lies in the impossibility to distinguish the changes in the statistical properties of the two contiguous areas. Such limitation was partially solved by several authors [5–7] who proposed the contemporaneous segmentation of successive frames of the pullback in order to improve the robustness of the method. However, as observed by [1] , the main limitation of all the active contour methods presented so far, is that the appearance of the B-mode image depends on the characteristic of the IVUS system and the parameters used for the image reconstruction. Thus, no segmentation method is guaranteed to perform correctly on IVUS images from different systems. A third strategy was explored by Kudo [8], who observed that during successive frames of the IVUS sequence, the texture inside the lumen exhibits a large variability of the speckle pattern due to the presence of blood flow, while the speckle pattern changes slowly in the tissue area. Kudo [8] introduced a model-based approach, exploiting the decorrelation generated by the blood flow. The faisibility of the approach [8] was illustrated by an *in vitro* experiment, using an acrylic tube phantom. Unfortunatelythe method cannot be straightforward extended to *in vivo* images because the model didn't account for the vessel pulsations and catheter oscillations. We decided, for the first time, to extend his approach by building a workflow applied to *in vivo* sequences.

In this paper we propose to first identify the most stable frames of the pullback, then register its contiguous frames in order to generate a parametric image discriminating the presence of steady tissues from the blood and finally segments the vessel lumen by classifying the vessel pixels. The method automatically segments the lumen border by combining temporal information obtained computing the local correlation between successive frames of the sequence with spatial classification performed using the Growcut [9] algorithm.

The main advantage of this approach is the segmentation robustness due to the combination of temporal and spatial analysis.

2 Method

The proposed segmentation method is composed of two steps: a model-based temporal correlation analysis and a spatial classification. The pipeline of the approach is sketched in Fig 2-a.

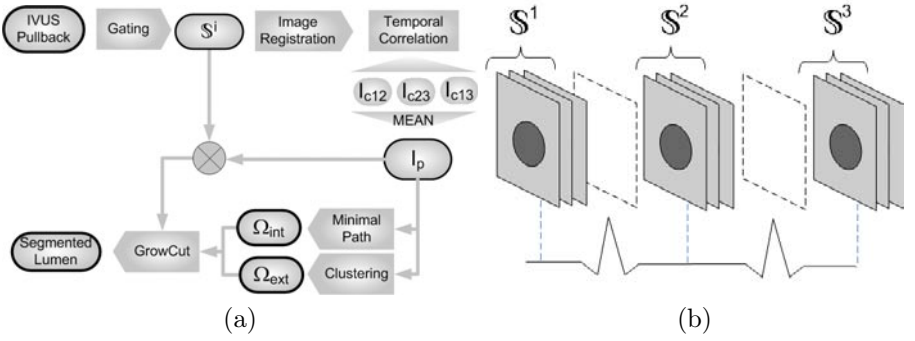


Fig. 2. Pipeline of the lumen segmentation method (a) and frames selection scheme (b)

2.1 Model-Based Temporal Analysis

Frames Selection and Motion Compensation. Heart beating causes a repetitive longitudinal oscillation of the catheter (swinging effect) along the vessel axis, resulting in multiple sampling of the same vessel positions. Additionally the best alignment between vessel tissues can be achieved when the arterial pulsation is minimal. For this reasons the the most stable frames of the pullback, commonly interpreted as belonging to the end-diastolic phase, can be robustly identified by ECG gating or, as in our case, by image-based gating [10]. Then, in order to take advantage of the information present in successive frames of the pullback, for each gated frame $I^G = I(x, y; G)$ (where G is a gated frame index), the previous and successive frames of the pullback are extracted for composing a local stack $\mathbb{S}^i = \left\{I^{G-1}, I^G, I^{G+1}\right\}$. Figure 2-b summarizes the frame selection process and the creation of local stacks. Finally, possible catheter translation and rotation, due to in-plane oscillation are compensated by computing a registration between all the stack frames. The optimal rigid registration is obtained by computing the transformation parameters that minimizes the mutual information between two successive according to the equation $X_2 = R(\theta) \cdot X_1 + T$ where $X = \begin{bmatrix} x \\ y \end{bmatrix}$, $R(\theta)$ is the rotation matrix, and T is the translation matrix.

Temporal Correlation. As proposed by [8], the correlation between successive frames of an IVUS sequence can provide useful information for the lumen border detection. In order to compute the temporal correlation of the speckle pattern as discriminative feature, the Pearson Correlation Coefficient is computed over a sliding window of size $H \times H$ along each pair of images of a stack $\mathbb{S}^i$ generating three correlation images $(I_{c12}, I_{c23}, I_{c13})$. The size of the window H is a trade off between edge over-smoothing when large windows are used and lack of precision in the correlation computation when a small number of samples are involved. The optimal size can be defined as the distance between two speckle noise peaks which can be automatically obtained from the duration of the pulse waveform [11], computed as the *full width at half maximum* of the noisy

autocorrelation coefficient. Finally a parametric image I_P is computed by averaging the correlations $(I_{c12}, I_{c23}, I_{c13})$ (Figure 3-a).

2.2 Spatial Classification

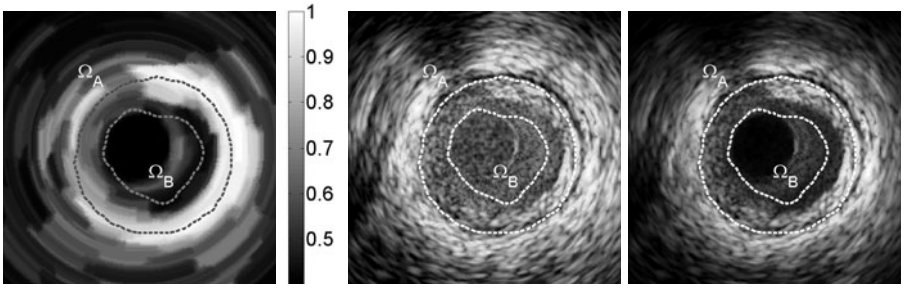
In this paper a fast and robust spatial segmentation technique has been used to classify the image pixel. The Growcut algorithm [9] is an iterative classification method assuring the spatial coherence of the segmentation. The method is able to optimally separate two regions of an image, R_A and R_B according to the pixel spatial relations and intensity. The approach relies to a cellular automaton technique, discrete in both space and time, that operates on pixels p of the image I and its neighborhood cells p. A cellular automaton system, is based on a cell state defined by a triplet $\Gamma = (L_p, \theta_p, V_p)$, where L_p is the label of the current cell, θ_p is the weight of the current cell, and V_p is the textural feature (in our case the intensity of the gray-scale image). The seeded cells are initially defined by labeling two compact subsets of the image (Ω_A and Ω_B) composed by all the pixel who likely belong to the external and internal regions R_A and R_B respectively, while not-assigned pixels Ω_C are initially set to: $L_p = 0$ and $\theta_p = 0$. The cell's state Γ_q^{t+1} at time step is defined by a rule considering the states of the neighborhood cells Γ_q^t at previous time step t. Starting from the initial seeds, at each iteration, the strongest neighbor cells q propagate its label to the cell p , and the new strengths of θ_p is computed weighting the strength of the neighbor cell and its distance:

$$\theta_p^{t+1} = g\left(\|V_p - V_q\|\right)\theta_q^{t+1} \tag{1}$$

where $\|\|$ is the euclidean distance and g is a monotonically decreasing function defined as $g(\xi) = 1 - \frac{\xi}{max(V)}$. The final goal of the segmentation is to assign a label to all the pixels.

Labels. The initial Growcut seed areas, Ω_A and Ω_B, are computed from the polar parametric image I_P, which is characterized by low correlation where the blood is flowing (lumen area) and high correlation where the arterial vessel and plaque are present. The correlation reaches its maximum close to the lumen border, and progressively decrease with the dept of the signal.

Based on these assumptions, the contour of Ω_A (Fig 3), corresponding to the surrounding tissues, is obtained by identifying the maximum intensity profile along the columns of I_P. The tracking of the ridge is done by the minimum cost path technique proposed by Cohen [12]. Such method computes the contour guaranteeing the minimal trajectory along an energy potential surface between two end points. The potential surface depends on both intensity of the image and distance from the target point. The method provides a smooth curve able to connect tissues separated by discontinuities (for instance generated by the shadow of the catheter guide, by stent wires or by calcium spots). The minimal path method [12] requires the definition of source and target coordinates. These are obtained as the image coordinate of the maximum intensity computed on the first and the last column of the image.



(a) (b) (c)

Fig. 3. Growcut seed areas $\Omega_A$and Ω_B superimposed to the parametric image I_P (a) to the initial frame I^G (b) and to the modulated frame I_P^G (c)

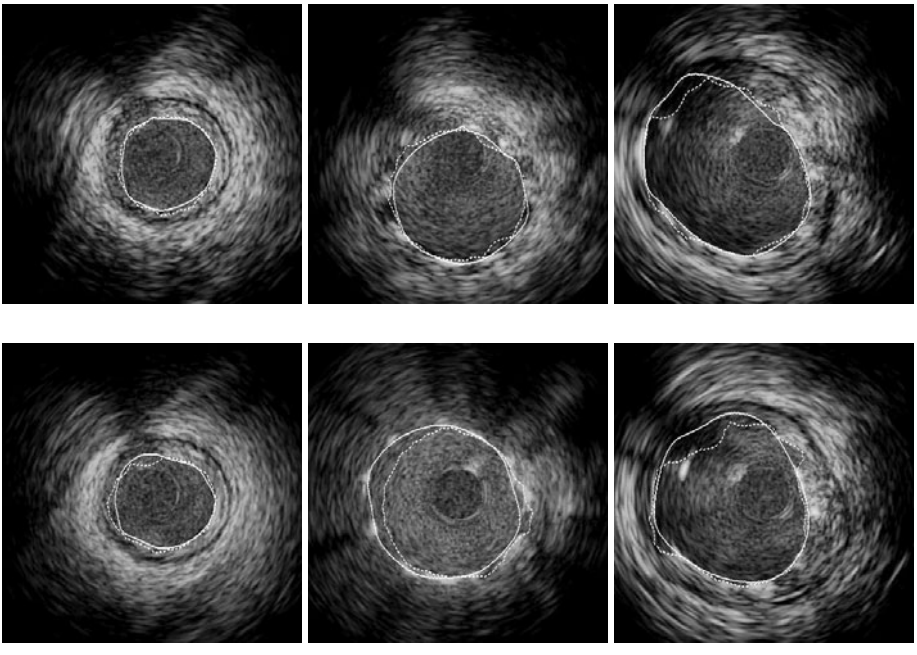
The second contour Ω_B, corresponding to the blood area is obtained by classifying the pixel of the polar parametric image I_P enclosed in Ω_A. Such samples can be either blood vessel or surrounding tissues, hence the pixels having low correlation are identifyed applying the classical Otsu threshold method [13]to the pixels enclosed in Ω_B .

Modulated Images. The Growcut classification method is applied to the modulated image I_M^G obtained multiplicating the gated frame I^G for the parametric image I_P^G . Such modulation combines the high frequencies of the IVUS image I^G, with the low frequencies of the parametric map I_P^G. Hence, it enhances the boundaries between lumen and plaque (especially in the regions where the speckle echogenicity is similar), since in I_P, the vessel border is sharp and lumen and vessel areas are uniform. Figure 3 illustrate the Growcut seed areas Ω_A and Ω_B superimposed to the parametric image I_P to the initial frame I^G and to the modulated frame I_M^G.

3 *In vivo* Experiments

The proposed segmentation method has been tested on three pullbacks, (each of them composed by about 2000 frames, for a total of 300 gated frames) of *in vivo* coronary images. Such database guarantees a representative number of different vascular structures (plaque and vessel shape, presence of stent, different lumen area and diameter) . The acquisition has been performed using an IVUS Galaxy II System with a catheter Atlantis SR Pro 40 MHz (Boston Scientific).

Ground truth areas, indicating the expected segmentation result, were manually delineated in each gated frames of the IVUS pullback by two experts. In order to obtain a fairly segmented reference data set, the validation interface enabled the expert to navigate from the previous/successive frame to provide insights about the temporal evolution of the pullback. The operator performing



d

Fig. 4. Examples of segmented images (vessel (a-d), stent (b-e) and bifurcation (c-f) frames). The continuous and dotted lines represent respectively the manual and the automatic contours.

the analysis was blinded to the automatic and to the other manual segmentation results. Figure 4 illustrates some segmentation examples of general vessel, stent and bifurcation frames, respectively first, second and third column obtained using a window size of $0.56 \times 0.56\ mm^2$. In the first row three successfully segmented frames are shown. In particular Figure 4-a illustrates how the segmentation of a challenging frame, in which only half of the plaque contour is visible (see Figure 1), has been achieved. The second row of Figure 4 present cases in which the result is not optimal. In Figure 4-d and -f the algorithm is fooled by the catheter guide, while in Figure 4-e the segmentation under-estimates the temporal correlation lumen area.

4 Discussion

In this paper we presented a lumen segmentation method which combines temporal information extracted from successive frames of the sequence with spatial classification issues from the spatial analysis. Encouraging results has been obtained, since the automatic and manual segmentation error are similar in most

Table 1. Segmentation errors in [mm] computed as the distance between automatic and manual observers contours

	Automatic		Inter-observer	
	Max	Mean	Max	Mean
Vessel	0.57 ± 0.24 [mm]	0.17 ± 0.08 [mm]	0.59 ± 0.73 [mm]	0.16 ± 0.06 [mm]
Stent	0.62 ± 0.19 [mm]	0.18 ± 0.07 [mm]	0.54 ± 0.31 [mm]	0.15 ± 0.06 [mm]
Bifurcation	1.12 ± 0.44 [mm]	0.31 ± 0.12 [mm]	1.80 ± 2.57 [mm]	0.23 ± 0.11 [mm]

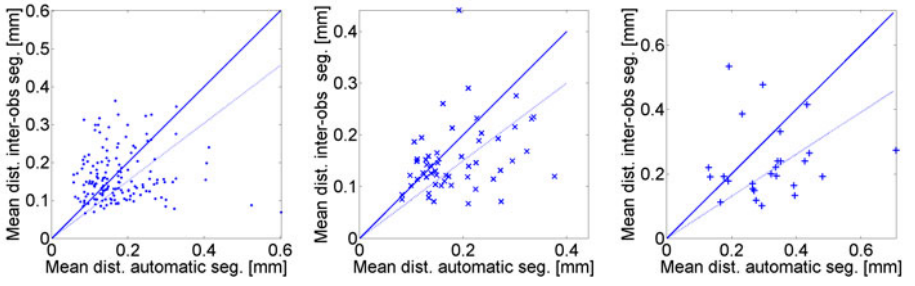


Fig. 5. Scatter plots comparing the automatic with the inter-observer segmentation errors. The plots report the average error measurements in the case of a vessel (a), stent (b), and bifurcation (c) frames. The solid and the dotted lines represent the unitary slope line and the linear regression curve, respectively.

of the cases. The performance of the approach is particularly good in vessel and stent frames showing an average segmentation error of 0.17 ± 0.08 mm and 0.18 ± 0.07 mm respectively. This approach is fully automatic and additionally it doesn't require parameter tuning or training making it potentially suitable for segmenting images from different echographs.

Future work will be addressed towards the validation of the technique comparing the performance on pullbacks belonging to different echographs brands, models and probes (different central frequency). The proposed framework will be evaluated using alternative spatial segmentation methods such as GraphCut algorithm. Finally, the applicability of the method to the whole pullback sequence is in progress.

Acknowledgments

This paper has been partially supported by projects TIN2009-14404-C02, La Marató de TV3 082131 and CONSOLIDER-INGENIO CSD 2007-00018. E-mail: balocco.simone@gmail.com.

References

1. Mendizabal-Ruiz, E.G., Biros, G., Kakadiaris, I.A.: An inverse scattering algorithm for the segmentation of the luminal border on intravascular ultrasound data. Med. Image Comput. Comput. Assist. Interv. 12(Pt2), 885–892 (2009)
2. Unal, G., Bucher, S., Carlier, S., Slabaugh, G., Fang, T., Tanaka, K.: Shape-driven segmentation of the arterial wall in intravascular ultrasound images. IEEE Trans. Inf. Technol. Biomed. 12(3), 335–347 (2008)
3. Gil, D., Radeva, P., Saludes, J., Mauri, J.: Automatic segmentation of artery wall in coronary ivus images: a probabilistic approach. Comput. Cardiol., 687–690 (2000)
4. Brusseau, E., Korte, C.D., Mastik, F., Schaar, J., Steen, A.V.D.: Fully automatic contour detection in intravascular ultrasound imaging. IEEE Trans. Med. Imag. 5(27), 108–118 (2004)
5. Klingensmith, J.D., Shekhar, R., Vince, D.G.: Evaluation of three-dimensional segmentation algorithms for the identification of luminal and medial-adventitial borders in intravascular ultrasound images. IEEE Trans. Med. Imaging 19(10), 996–1011 (2000)
6. Jianming, H., Xiheng, H.: An approach to automatic segmentation of 3d intravascular ultrasound images. In: Nuclear Science Symposium and Medical Imaging Conference, vol. 3, pp. 1461–1464 (1994)
7. Sonka, M., Liang, W., Zhang, X., DeJong, S., Collins, S.M., McKay, C.R.: Three-dimensional automated segmentation of coronary wall and plaque from intravascular ultrasound pullback sequences. In: Computers in Cardiology 1995, pp. 637–640 (1995)
8. Kudo, N., Kanenari, T., Zhang, X., Yamamoto, K.: In vitro study on arterial lumen detection using a correlation technique in ivus, vol. 2, pp. 830–831 (1998)
9. Vezhnevets, V., Konouchine, V.: Grow-cut - interactive multi-label n-d image segmentation, pp. 150–156 (2005)
10. Gatta, C., Balocco, S., Ciompi, F., Hemetsberger, R., Leor, O.R., Radeva, P.: Real-time gating of ivus sequences based on motion blur analysis: Method and quantitative validation. Med. Image Comput. Comput. Assist. Interv. 13, 59–67 (2010)
11. Smith, S.F., Wagner, R.F.: Ultrasound speckle size and lesion signal to noise ratio: verification of theory. Ultrason Imaging 6(2), 174–180 (1984)
12. Cohen, L., Kimmel, R.: Global minimum for active contour models: A minimal path approach. International Journal of Computer Vision 24, 57–78 (1997)
13. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9(1), 62–66 (1979)

Topographic EEG Brain Mapping before, during and after Obstructive Sleep Apnea Episodes*

David Belo[1], Ana Luísa Coito[1], Teresa Paiva[2], and João Miguel Sanches[1]

[1] Institute for Systems and Robotics / Instituto Superior Técnico
[2] Centro de Electroencefalografia e Neurofisiologia Clínica / Faculdade de Medicina da Universidade de Lisboa
djbelo@isr.ist.utl.pt, anacoito20@gmail.com, tpaiva@ispa.pt, jmrs@ist.utl.pt

Abstract. Obstructive Sleep Apnea Syndrome (OSAS) is a very common sleep disorder that is associated with several neurocognitive impairments. The present study aims to assess the electroencephalographic (EEG) power before, during and after obstructive apnea episodes, in four frequency bands: delta (δ), theta (θ), alpha (α) and beta (β). For that purpose, continuous wavelet transform was applied to the EEG signals obtained with polysomnography, and topographic EEG brain mapping (EBM) to visualize the power differences across the whole brain. The results demonstrate that there is a significant decrease in the EEG δ power during OSAS that does not totally recover immediately after the episode. Furthermore, a power decrease in a specific brain region was noticed for all EEG frequency ranges.

Keywords: Obstructive Sleep Apnea, Electroencephalogram, Spectral Analysis,Continuous Wavelet transform, Brain Mapping.

Introduction

Obstructive Sleep Apnea Syndrome (OSAS) is a very common sleep disorder affecting 4% of men and 2% of women [1] and is sometimes undiagnosed. It is is characterized by recurrent apneas during sleep, which are caused by the partial or complete collapse of the upper airway, resulting in repetitive hypoxemic and hypercapnic episodes, and interruptions of the normal sleep pattern.

OSAS contribute to the development of not only respiratory and cardiovascular disorders but also neurocognitive impairments. Indeed, neuropsychological investigations of patients with OSAS have shown impairments in functions as memory, attention and executive control [2]. The pathophysiological mechanisms underlying the morbidity of OSAS are not completely understood, which make the research on the OSAS an important issue.

* This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 564–571, 2011.

It is known that intermittent hypoxia, as it occurs in OSAS, is associated with cortical neuronal cell death (gray matter loss) in cognitively relevant brain regions and consequent cortico-hippocampal damage [3]. Moreover, it was found out that during an apnea episode there is a decrease of oxyhemoglobin, increase of deoxyhemoglobin and an increase of cerebral blood flow, however the latter cannot compensate for reduced arterial oxygen saturation and cerebral tissue hypoxia may occur during OSA [4].

Neurophysiological assessment through the electroencephalographic (EEG) signal provides an objective method for detecting changes in cortical activity. The EEG signal shows patterns of electrical activity, each one characterized by a typical frequency band and amplitude. The normal human EEG shows activity over the range of 1-30 Hz with amplitudes in the range of 20-100 μV [5]. The lowest amplitude waves and highest frequency, 18-30 Hz, are named beta (β) rhythm. Alpha (α) rhythm lies between 8-12 Hz with amplitude of 50 μV. Larger regular waves of frequency range 4-7 Hz called theta (θ) rhythm have been recognized along with a slow wave of less than 4 Hz called the delta (δ) rhythm [5]. The EEG spectral analysis was found to be a very useful tool to assess the EEG power in the four stated EEG frequency ranges [5].

In the current study, obstructive sleep apnea (OSA) episodes were carefully selected and segmented in three parts, the OSA event (*dur*), a certain period immediately preceding (*pre*) and a time interval after (*post*) the event in order to assess the dynamic EEG power changes. The analysis of the EEG signal in the four bands is performed by using the continuous wavelet transform (CWT) and topographic EEG brain mapping (EBM) for visualization of the power in the whole brain.

As far as the authors know this is the first study where data segmentation in *pre*, *dur* and *post* was performed in adult OSAS patients to study the EEG power changes. The studies that analyzed the EEG power during apnea are usually focused in the detection of non-visible arousals (related to autonomic activation). Furthermore, EBM is introduced as a powerful tool to visualize spectral changes during OSA episodes across the brain in a local basis.

This is the second study applying EBM to assess spectral power changes during OSA. The first one dates back to 1993 and had a poor resolution [6]. We consider a promising method for studying neurophysiological aspects of brain function during OSA since it allows the assess of the local power distribution across the whole brain and, thus, to visualize which specific region is affected during an apneic event. The relevance of this paper is held in the correlation of some known neurocognitive impairments occurring in OSAS patients with their affected brain areas, establishing a new door to a better understanding of the effects of this sleep disorder in the human being.

1 Methods

A total of 15 male individuals with OSAS, mean aged 55 $\pm$ 6.10 (mean $\pm$ standard deviation, SD) and with a mean body mass index (BMI) 28.74 $\pm$ 5.26

$Kg.m^{-2}$, participated in this study. They underwent overnight polysomnographic (PSG) assessment through a computerized PSG system (*Somnologica 5.0.1*, *Embla*) during approximately 8 hours.

EEG electrodes were positioned according to the International 10-20 System and 21 recordings were acquired, at a sampling frequency of 100 Hz, from the following leads: Fp1, Fp2, Fpz, F3, F4, F7, F8, Fz, C3, C4, Cz, P3, P4, Pz, O1, O2, Oz, T3, T4, T5 and T6, in reference to linked ears (A1 and A2). Electrocardiogram, thorax and abdominal efforts, airflow, oxygen saturation (SpO_2) and electromyographic channels were also recorded.

Each recording was visually examined and sleep stages were scored manually at 30 seconds intervals, according to the criteria of the *American Academy of Sleep Medicine* [7].

An oxygen desaturation event was detected when the oxygen saturation fell by at least 4%.

1.1 Dataset

A sleep apnea event was detected when a 10 second interval of the airflow signal dropped below 20% of the reference amplitude. Only episodes that obeyed to some criteria were considered: obstructive apnea-type events occurring in NREM-2 sleep stage, lasting up to 60 seconds and be preceded and followed by at least 30 seconds of continuous breathing, since it was intended to analyze data not only during an OSA (*dur*) but also before (*pre*) and after (*post*) the event. Each *pre* and *post* have a duration of 30 seconds.

The final dataset included 171 isolated OSA episodes, without artifacts, extracted from the 15 mentioned patients. In total, 10773 epochs (171 episodes×3 periods×21 channels) were analyzed. The mean duration of these episodes is 14.54 ± 6.43 s.

1.2 Signal Processing

The recordings were exported to European Data Format (EDF) files in order to be analyzed in Matlab 7.5.0, in which all the signal processing was performed.

First of all, a noise reduction step was taking into account. The moving average of each EEG signal was removed using time windows of 4s. A median filter of order $n = 10$ was also applied to the signals.

In order to obtain the power in δ, θ, α and β frequency bands, the continuous wavelet transform (CWT) method was then applied to each EEG epoch.

The wavelet transform has been considered over the recent years as a powerful time-frequency analysis for the manipulation of complex nonstationary signals, such as physiological signals [8]. This technique decomposes a signal into a set of basic functions called wavelets, which are obtained by dilations, contractions and shifts of a unique function: the mother wavelet [8], that in the case of the present study was the Morlet function, defined as:

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}} \left(e^{iw_0 t} - e^{-\frac{w_0^2}{2}} \right) e^{-\frac{t^2}{2}} \tag{1}$$

where w_0 is the central frequency of the mother wavelet (frequency at the center of a Gaussian curve), the term in brackets is known as the correction term (it corrects the nonzero mean of the complex sinusoid of the first term)

For a practical implementation, CWT is computed over a discretized time-frequency grid, which involves an approximation of the transform integral. [8].

1.3 Power Calculation

After processing the EEG signal (x), the mean energy (E_p) for each apnea epoch ($p = pre, dur, post$), episode ($i = 1, ..., M$), channel (c), frequency band (ω_b), and was calculated by the following equation:

$$E_p(c, i, \omega_b) = \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{s=1}^{S} |x(c, n) * W(n, \omega_s)|^2 \right], \forall\, \omega_s \in \omega_b \tag{2}$$

where n represents the sample ($n = 1, ..., N$) and s the wavelet scale. W(n,ω_s) represents the scaled version of the Morlet wavelet with the central frequency of ω_s. The integral of the power of the frequency band ω_b is the combination of S equal spaced wavelets in the frequency domain.

Finally, the relative energy of each EEG channel and frequency band is calculated by the following equation:

$$e_p(c, \omega_b) = \frac{1}{M} \sum_{i=1}^{M} \left[\frac{E_p(c, i, \omega_b)}{E_{pre}(c, i, \omega_b)} \right] - 1 \tag{3}$$

1.4 Brain Mapping and Statistical Analysis

The EBM were made by the aproximation of the head to a semi-sphere [9]. This simplification allows spherical interpolation of the vector $B_p^{\omega_b} = [e_p(1, \omega_b), e_p(2, \omega_b), ..., e_p(21, \omega_b)]^T$ for the mean values. The same processing was made for the standard deviation. The EBM was computed using EEGLAB's function topoplot() with $B_p^{\omega_b}$ as the input vector. EEGLAB is a software toolbox for Matlab (more information is freely available from http://www.sccn.ucsd.edu/eeglab/) [10].

Powers corresponding to *dur* and *post* for each EEG frequency domain were compared by a tailed two-sample *t-test* against *pre* segments. This test considers as null hypothesis the independency of two samples from normal distributions and that the mean of one is higher than the other. A p-value < 0.05 was considered statistically significant.

All the episodes were tested for E_{dur} and E_{post} against E_{pre} for each frequecy band (ω_b) and channel (c). The binary vector $S_p^{\omega_b}(c)$ was the result of the validation (1 for pass, 0 otherwise). New brain maps were made using topoplot() with each $S_p^{\omega_b}$ as the input vectors. The interpolated values were considered true if they were in the interval $]0.5, 1]$ and false for $[0, 0.5]$. These maps were used as a mask to hide the points that weren't statisticaly significant of the respective mean power maps.

2 Results

2.1 Demographic, Respiratory and Polysomnographic Variables

The resume of the polysomnographic characteristics of the 15 male patients considered in this study are shown in Table 1.

Table 1. Polysomnographic characteristics in 15 OSAS patients

Parameter	Mean ± SD
AHI (hours^{-1})	30.87 ± 13.04
*TST (min)	411.87 ± 109.46
Sleep efficiency (%)	80.07 ± 16.39
TST in NREM1 (% of TST)	16.24 ± 7.18
TST in NREM2 (% of TST)	58.91 ± 9.03
TST in NREM3 (% of TST)	12.63 ± 6.87
TST in REM (% of TST)	12.23 ± 5.88
Number of arousals	105.67 ± 78.23
SpO_2 baseline (%)	94.19 ± 1.32
Nadir SpO_2 (%)	78.67 ± 8.79
Number of desaturations	82.71 ± 64.34

* *TST - total sleep time*

As it is shown, these OSAS patients had low percentages of NREM-3 and REM sleep, higher percentages of stage NREM-1 and NREM-2 sleep, and a high number of arousals, which show a clear disturbed sleep pattern. In terms of hypoxemia, the patients included in the study were severely affected: they were characterized by a mean SpO_2 nadir of less than 80% and a high number of dessaturations.

2.2 EEG Analysis

The results (Figures 1 and 2) show a statistically significant generalized δ power decrease during OSA, which is not fully recovered after the episode for all the brain spectra.

For θ waves, there is a statistically significant power decrease only in frontal (F3 and F4) and temporal (T3 and T4) regions during OSA, and in part of the occipital (O1 and O2), temporal (T3 and T4), central (C3 and C4) and frontal (F3, F4, Fp1 and Fpz) regions in *post*.

The EBM for the α and β frequency bands during OSA show a statistically significant power decrease in occipital (O1 and O2), temporal (T3 and T4), central (C3 and C4) and frontal (F3 and F4) regions of the brain, and an increase in all the parietal are for the β band. In *post*, there is an overall slight power increase, which is stronger in the frontal and occipital brain regions. However, statistically, only channels F3, T5 and P4 are significant for β band, and Fp1, P4 and Oz for α band. The β and α increases after the OSA episode, specially

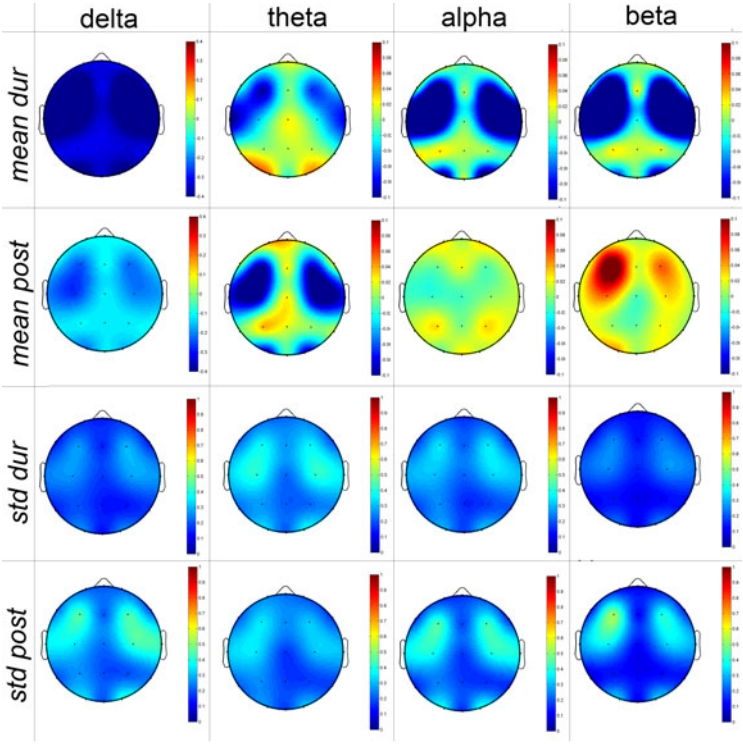


Fig. 1. EBM of the normalized power with *pre* apnea as reference. The sections of the first row represent the mean power values of *dur* apnea for the four frequency bands, and the third row the respective standard deviations. The second row represent the power mean of *post* apnea for the four different frequency bands and the forth row the respective standard deviations. Each brain map has a gauge with a color range from dark red, the highest energy content, to dark blue, the lowest energy content, according to the respective scale.

in the frontal (motor area) and occipital (visual area) regions, are probably due to an arousal mechanism that often accompanies the termination of an apneic event, which are responsible for sleep fragmentation [11].

The SD maps validate the results above described. Note that areas with a higher SD for all frequency bands are those representing pre-motor, motor and visual areas of the brain.

Decreases in δ power preceding arousal and termination of apneic events in both REM and NREM sleep are reported [6,12]. However, the opposite conclusions were also addressed [13,14]. Perhaps the difference between their results and the ones obtained in the current study are due to the fact that δ power was assessed only near or even at the apnea termination, so subcortical arousals, including K-complexes and δ bursts, might occur and, thus, contribute to the

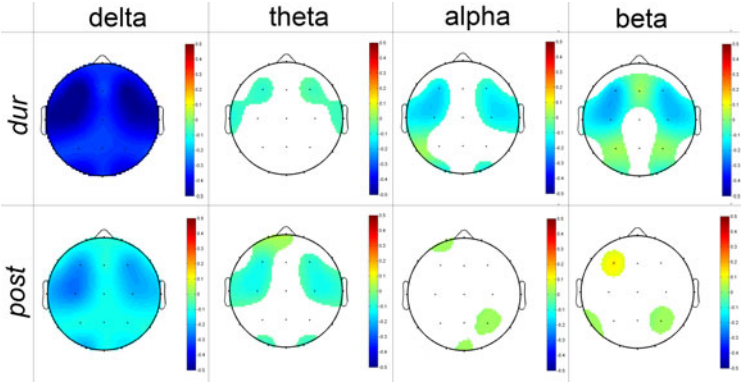


Fig. 2. Statistical significance map. The displayed maps represent the normalized power points that passed the tailed two-sample t-test for *post* and *dur* against *pre* apnea episodes. Each brain map has a gauge with a color range from dark red, the highest energy content, to dark blue, the lowest energy content, according to the respective scale.

reported δ increases. In this work, all OSA episodes were visually examined and δ bursts were removed at the end of OSA episodes, so that they did not influence the mean δ power of *post*.

Please note that there is a power decrease during OSA also in θ, α and β in specific regions - frontal (F3 and F4), temporal (T3 and T4) and central (C3 and C4) regions - which shows a clear decreased EEG activity in these regions that might evidence that this is the most affected brain region during OSA episodes. Since working memory and executive tasks performance are localized at the frontal cortex [2,11], it is possible that memory impairments reported in OSAS patients are due to a decrease of the brain's activity in this referred specific brain region.

It was suggested for various authors that there is a correlation between δ power changes and the severity of hypoxemia and hypercapnia during the OSA [6,15]. Moderate hypoxemia has been shown to elicit a depression of absolute power in the EEG δ band [6]. So, it is possible that the detected δ fluctuations may be due to hypoxemia and/or hypercapnia.

3 Conclusion

A new approach was carried out for assessing EEG changes during an OSA episode: EBM. This technique was proved to be very useful, since it allowed to draw new conclusions about the visualization of the brain has a whole. It is a reliable tool for the assessment of EEG spectral power changes, of each region, resulted from an obstructive event.

The present study confirms that the majority of OSA, which can be or not terminated by visually scored arousals, are associated with significant spectral

power changes, mainly in δ frequency band, where there is a clear decrease in δ power during OSA. Moreover, it was notice that a power decrease in a specific brain region occurred for all EEG frequency ranges. This sugests that

Future studies include the performance of memory tests to the OSAS patients assessed by the presented analysis in order to correlate the memory impairments with the spectral EEG power changes observed during OSA episodes.

References

1. Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S.: The occurrence of sleep-disordered breathing among middle-aged adults. N. Engl. J. Med. 328, 1230–1235 (1993)
2. Naegel, B., Launois, S., Mazza, S., Feuerstein, C., Ppin, J., Lvy, P.: Which memory processes are affected in patients with obstructive sleep apnea? An evaluation of 3 types of memory. Sleep 29(4), 533–544 (2006)
3. Xu, W., Chi, L., Row, B., Xu, R., Ke, Y., Xu, B., Luo, C., Kheirandish, L., Gozal, D., Liu, R.: Increased oxidative stress is associated with chronic intermittent hypoxia-mediated brain cortical neuronal cell apoptosis in a mouse model of sleep apnea. Neuroscience 126, 313–323 (2004)
4. Hayakawa, T., Terashima, M., Kayukawa, Y., Ohta, T., Okada, T.: Changes in cerebral oxygenation and hemodynamics during obstructive sleep apneas. Chest 109, 916–921 (1996)
5. Muthuswamy, J., Thakor, N.: Spectral analysis methods for neurological signals. J. Neurosci. Methods 83, 1–14 (1998)
6. Walsleben, J., O'Malley, E., Bonnet, K., Norman, R., Rapoport, D.: The utility of topographic EEG mapping in obstructive sleep apnea syndrome. Sleep 16, 76–78 (1993)
7. Iber, C., Ancoli-Israel, S., Chesson, A., Quan, S.: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. American Academy of Sleep Medicine, Westchester (2007)
8. Addison, P., Walker, J., Guido, R.: Time-Frequency Analysis of Biosignals: a wavelet transform overview. IEEE EMB Magazine 28(5), 14–29 (2009)
9. Ferree, T.: Spherical Splines and Average Referencing in Scalp Electroencephalography. Brain Topography 19, 43–52 (2006)
10. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods 134, 9–21 (2004)
11. Verstraeten, E.: Neurocognitive Effects of Obstructive Sleep Apnea Syndrome. Current Neurology and Neuroscience Reports 7, 161–166 (2007)
12. Bandla, H., Gozal, D.: Dynamic Changes in EEG Spectra During Obstructive Apnea in Children. Pediatric Pulmonology 29, 359–365 (2000)
13. Svanborg, E., Guilleminault, C.: EEG Frequency changes during sleep apneas. Sleep 19, 248–254 (1996)
14. Berry, R., Asyali, M., McNellis, M., Khoo, M.: Within-night variation in respiratory effort preceding apnea termination and EEG delta power in sleep apnea. Journal of Applied Physiology 85, 1434–1441 (1998)
15. Morisson, F., Lavigne, G., Petit, D., Nielsen, T., Malo, J., Montplaisir, J.: Spectral analysis of wakefulness and REM sleep EEG in patients with sleep apnoea syndrome. Eur. Respir. J. 11, 1135–1140 (1998)

Classifying Melodies Using Tree Grammars

José Francisco Bernabeu, Jorge Calera-Rubio, and José Manuel Iñesta

Dept. Lenguajes y Sistemas Informáticos, Universidad de Alicante, Spain
{jfbernabeu,calera,inesta}@dlsi.ua.es

Abstract. Similarity computation is a difficult issue in music information retrieval, because it tries to emulate the special ability that humans show for pattern recognition in general, and particularly in the presence of noisy data. A number of works have addressed the problem of what is the best representation for symbolic music in this context. The tree representation, using rhythm for defining the tree structure and pitch information for leaf and node labeling has proven to be effective in melodic similarity computation. In this paper we propose a solution when we have melodies represented by trees for the training but the duration information is not available for the input data. For that, we infer a probabilistic context-free grammar using the information in the trees (duration and pitch) and classify new melodies represented by strings using only the pitch. The case study in this paper is to identify a snippet query among a set of songs stored in symbolic format. For it, the utilized method must be able to deal with inexact queries and efficient for scalability issues.

Keywords: Music Modeling & Analysis, Stochastic Methods, Learning with Structured Data, Music Similarity, Classification.

1 Introduction

One of the main concerns in music information retrieval (MIR) tasks is how to assess melodic similarity in a similar way to how humans do. We are very good at recognizing previously known patterns, even if our perceived inputs are distorted, they are presented just partially, or in the presence of noisy data. This happens in music comparison in a number of situations, for example, when comparing different cover versions of a given melody or when searching in databases using a query that will be, by its own nature, partial and can be distorted or even wrong. Two issues are concerned to this problem: the similarity computation and the representation structure.

In this paper, the problem of comparing symbolically encoded (any format of digital scores) musical works is addressed. For it, we use probabilistic tree grammars [12]. These grammars can be obtained from probabilistic k-testable tree models [9]. The duration information implicit in the tree representation is captured by the grammar and this is used for classifying the new melodies represented by strings that only have the pitch information. This is a solution when duration information is not available or unreliable for the input data. The results are compared with those obtained by Bernabeu [1] where input and training data are trees.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 572–579, 2011.

2 Melody Tree Representation

For representing the note pitches in a monophonic melody s as a string, symbols σ from a pitch representation alphabet Σ_p are used: $s \in \Sigma_p^*, s = \sigma_1\sigma_2...\sigma_{|s|}$. In this paper, the interval from the tonic of the song modulo 12 is utilized as pitch descriptor and the symbol '$-$' to represent rests. Thus the alphabet used is: $\Sigma_p = \{p \in \mathcal{N} \mid 0 \leq p \leq 11\} \cup \{\text{'}-\text{'}\}$. This way, in 'G Major', any pitch 'G' is mapped to 0 and the other pitches are represented by the number of semitones mod 12 from 'G'. This alphabet permits a transposition invariant representation and keeps cardinality low ($|\Sigma_p| = 13$).

In the tree approach, each melody bar is represented by a tree, $t \in T_{\Sigma_p}$. Bars are coded by separated trees and then they are linked to a common root. The level of a node in the tree determines the duration it represents. Each tree root (level 1) represents the duration of the whole bar. For a binary meter, the two nodes in level 2 represent the duration of the two halves of a bar, etc. In general, nodes in level i represent duration of a $1/2^{i-1}$ of a bar for a binary meters ($1/3^{i-1}$ for ternary). Therefore, during the tree construction, nodes are created top-down when needed and guided by the meter, to reach the appropriate leaf level to represent a note duration. In that moment, the corresponding leaf node is labeled with the pitch representation symbol, $\sigma \in \Sigma_p$ (see [10] for details).

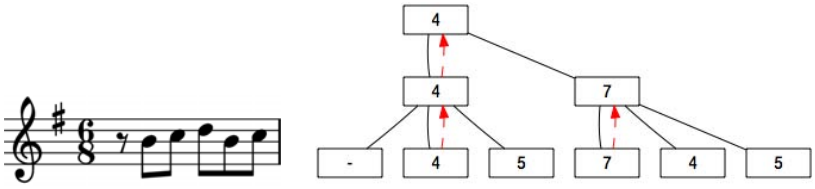


Fig. 1. Tree representation of a one-bar melody in a ternary meter with an example of how pith labels are propagated

Once the tree has been built, a bottom-up propagation of the pitch labels is performed to label all the internal nodes. The rules for this propagation are based on a melodic analysis [6]. All the notes are tagged either as *harmonic tones*, for those belonging to the current harmony at each time, or as *non-harmonic tones* for ornamental notes. Harmonic notes have always priority for propagation and when two harmonic notes share a common father node, propagation is decided according to the metrical strength of the note (the stronger the more priority), depending on its position in the bar and the particular meter of the melody. Notes have always higher priority than rests (Fig. 1 shows an example). Eventually, when all the internal nodes are labeled, all bar trees are linked to a common forest root, labeled with the root of the first bar tree.

3 Stochastic *k*-Testable Tree Models

Stochastic models based on k-grams predict the probability of the next symbol in a sequence depending on the $k-1$ previous symbols. They have been

extensively used in natural language modeling and also in some music tasks [4]. k-gram models can be regarded as a probabilistic extension of locally testable languages [13]. Informally, a string language $\mathcal{L}$ is locally testable if every string w can be recognized as a string in $\mathcal{L}$ just by looking at all the substrings in w of length at most k, together with prefixes and suffixes of length strictly smaller than k to check near the string boundaries. These models are easy to learn and can be efficiently processed.

Locally testable languages, in the case of trees, were described by Knuutila [7]. The concept of k-fork, f_k, plays the role of the substrings, and the k-root, r_k, and k-subtrees, s_k, play the role of prefixes and suffixes. For any $k > 0$, every k-fork contains a node and all its descendants lying at a depth smaller that k. The k-root of a tree is its shallowest k-fork and the k-subtrees are all the subtrees whose depth is smaller than k.

These kind of probabilistic tree languages can be defined using the formalism of deterministic tree automata (DTA). The procedure to infer this kind of automata from a training sample, can be done easily (see [9] for details). This learning procedure can be extended to the case where the sample Ω is stochastically generated, incorporating probabilities to the DTA.

As shown in [9], a probabilistic DTA (PDTA) incorporates a probability, $p_m(\sigma, t_1, ..., t_m)$, for every transition in the automaton, with the normalization that the probabilities of the transitions leading to the same state $q \in Q$ must add up to one. For this purpose, one should note that, in this kind of deterministic models, the likelihood of the training sample is maximized if the stochastic model assigns to every tree t in the sample a probability equal to its relative frequency in Ω [8]. So, these probabilities must be calculated as the ratio of the number of occurrences of a transition to the number of occurrences of the state to which this transition leads. PDTA also incorporate a probability $\rho(q)$ for every accepting state, $q \in F$ ($F \subseteq Q$). These probabilities are calculated as the ratio between the number of occurrences of an accepting state and the number of trees in the sample, $|\Omega|$. It is useful to store the above probabilities as the quotient of two terms. This way, if a new tree (or subtree) t is provided, the automaton can be easily updated to account for the additional information. For this incremental update, it suffices to increment each term with the partial sums obtained for parsing t. Finally, the probability of the tree t is computed as the product of the transitions utilized in the parsing of the tree (see [9] for details).

4 Grammars

At this point, we can classify a new melody in a particular class. For this purpose, we need to infer a PDTA for each class, C_j, from well classified melodies. Once the PDTAs for the different classes have been inferred and the probabilities estimated (see [9] for details), a melody M can be classified in the class $\hat{C}$ that maximizes the likelihood (see [1]).

In order to do this, both training and new melodies must be represented by trees. However, what happens if the new melodies are only represented by

strings? Moreover, what happens if a new melody string has the pitches but the duration information is not available? This situation appears when a melody query is given using only note pitches or when durations are not reliable. In other words, we have a set of melodies represented by trees to train the system but the target data are melodies represented by pitch strings. Therefore, we need to transform the k-testable tree automata in context-free grammars [12] in order to use them for parsing the input melody strings.

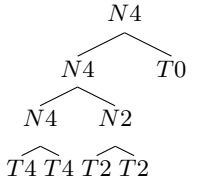
Context-free grammars may be considered to be the customary way of representing syntactical structure in natural language sentences. In many natural language processing applications, to obtain the correct syntactical structure for a sentence is an important intermediate step before assigning an interpretation to it. In our case, we use these grammars to obtain the correct structure for a given melody represented by a string.

Treebank grammars - which are explained in detail in [12]- are probabilistic context-free grammars in which the probability that a particular nonterminal is expanded according to a given rule is estimated as the relative frequency of that expansion by simply counting the number of times it appears in a manually-parsed corpus.

Before transforming our k-testable tree automata in context-free grammars we need introduce some changes in the melody tree representation. These changes are necessary because if we use the alphabet described in section 2 then a transition in the automaton could be transformed in different grammar rules. This happens because we can not distinguish between symbols that are terminals or nonterminals. In order to solve these ambiguities we need to label tree nodes adding the symbol 'T' to that of Σ_p if it is a leaf node (terminal) and the symbol 'N' if it is an inner node (nonterminal).

Therefore, if $\Omega = t_1, t_2, \ldots, t_{|\Omega|}$ is a treebank, that is, a stochastic sample of parse trees, the alphabet Σ can be safely partitioned into the subset $s_1(\hat{\Omega})$ of labels that may only appear at leaves ($\Sigma_{pT} =$ 'T'Σ_p) and its complementary subset $\Sigma - s_1(\hat{\Omega})$ ($\Sigma_{pN} =$'N'Σ_p) of labels at internal nodes (propagated labels in the trees).

Then, we can define a probabilistic k-testable grammar as $G^{[k]} = (V^{[k]}, T, I, R^{[k]}, p^{[k]})$, where $V^{[k]} = I \cup r_{k-1}(f_k(\Omega) \cup s_{k-1}(\hat{\Omega})) - s_1(\hat{\Omega})$ is the set of nonterminals, $T = s_1(\hat{\Omega})$ is the set of terminals, I is the start symbol, $R^{[k]}$ is the set of production rules, and $p^{[k]}$ a probability function (see [12] for details).



(1/1) $I \rightarrow N4$
(1/3) $N4 \rightarrow N4 \;\; T0$
(1/3) $N4 \rightarrow N4 \;\; N2$
(1/3) $N4 \rightarrow T4 \;\; T4$
(1/1) $N2 \rightarrow T2 \;\; T2$

Fig. 2. Example of a probabilistic tree grammar for $k = 2$

Figure 2 shows the corresponding probabilistic context-free grammars (PCFG) (right) according to the tree in left. For this tree we have the sets (roots) $r_1(t) = N4$, (forks) $f_2(t) = \{N4(N4\,T0), N4(N4\,N2), N4(T4\,T4), N2(T2\,T2)\}$, (subtrees) $s_1(t) = \{T0, T2, T4\}$. Therefore, we can obtain k-testable grammars with different values for k.

4.1 Smoothing

In general, k-testable grammars with larger values of k contain more specialized rules and, therefore, are less ambiguous and allow for faster parsing. In contrast, typical treebank grammars have 100 percent coverage (as remarked in [3]) unlike with higher order grammars where sentences without a valid parse tree are common. Therefore, the use of smoothing techniques becomes necessary if one wants to use these models for parsing. Two classical techniques of this type are linear interpolation and backing-off [8].

Smoothing through linear interpolation is performed by computing the probability of events as a weighted average of the probabilities given by different models. This approach has a problem when the higher order models return a zero probability due to unseen labels. Then, a new melody is classified only with the more general and more ambiguous model discarding the entire more specific model.

In contrast, backing-off allows to avoid lower-order parsing when possible. In other words, backing-off tries to parse with the higher-order grammar unless no parse tree is provided by this grammar. Only in such a case the lower-order model is called, so backing-off is faster than linear interpolation. However, the lack of a single rule in the sample can force the parser to use the lower-order model, loosing all the higher-order information for a whole sentence.

Here, we use an alternative approach: the rule-based backing-off [12]. Using this rule-based backing-off requires the implementation of specific parsers since building the whole grammar is unfeasible due to the large number of implicit rules (even if only those assigning a strictly positive probability in the last case of [5] are considered). An alternative scheme that requires only minor modifications is to use a quasi-equivalent grammar G' built as in [12].

However, we need to define a universal grammar because some labels in Σ_p for the leaves of the trees do not appear in the training data for the grammars. Then, if a particular new melody contains an unseen label, the parser will return a zero probability. For solving this problem we only have to introduce the rules of the form $N\sigma_1 \rightarrow T\sigma_2$ (where $\sigma_1, \sigma_2 \in \Sigma_p$) (updating the corresponding counters) if the rule did not appear during training. Introducing these rules the grammar assigns a non zero probability for each string even if the label does not appear in the training data.

4.2 Classification

As explained before, we want to study if the proposed approach can be used to classify new melodies represented by strings. After a grammar G_j is inferred for

each class C_j we need an algorithm for obtaining the probability that a given string s is generated by a grammar G_j. For this purpose, we have used the Stolcke algorithm [11] and the CYK+ algorithm [2] for string parsing. These parsing algorithms are able to give the probability $p(s|G)$ that a string s is generated by a probabilistic Context-free grammar G without requiring conversions to Chomsky Normal Form (CNF). Then, a melody M is classified in the class $\hat{C}$ that maximizes the likelihood

$$\hat{C} = \arg\max_j \ l(M|C_j) \tag{1}$$

We can calculate this likelihood two ways: splitting the melody (SplitBars) in bars or computing the whole melody (Whole).

In SplitBars, the melody string is split in $|M|$ (number of bars in a melody M) bar strings $s_1, \ldots, s_{|M|}$. Therefore we are able to compute the probability of each bar string to belong to a particular class (grammar). Suppose we have a finite number of classes and we have computed the membership probability of each bar string s_i to each of these grammars, $p(s_i|G_j)$. These probabilities can be combined to give a decision for the whole song. For the combination of bars the geometric mean has been used. The geometric mean is less sensitive to outliers than the arithmetic one. For our purposes is enough to multiply all bar strings probabilities of the whole melody. Calculating the $|M|$-th root of the resulting product is not needed for classification because, given a particular melody M to classify, $|M|$ is the same for all classes. Therefore,

$$l(M|C_j) = \prod_{i=1}^{|M|} p(s_i|G_j) \tag{2}$$

On the other hand, in the Whole strategy we only need the probability of the melody string (now $M = s$). Then

$$l(M|C_j) = p(s|G_j) \tag{3}$$

For calculating this probability we need to introduce the start rules $S \to IS$ and $S \to I$ which define a melody recursively (melody is formed by a bar and a melody) (I is the initial symbol for a bar as explain in section 4). Therefore, the grammars allow to recognize a whole melody instead of melodies split in bars.

5 Results

In our experiments, we try to identify a problem melody using a set of different variations played by musicians for training. For that, we use a corpus consisting of a set of 420 monophonic 8-12 bar incipits of 20 worldwide well known tunes of different musical genres[1]. For each song, a canonic version was created using a score editor and synthesized. The audio files were given to three amateur and

[1] The MIDI data set is available upon request to the authors.

two professional musicians who listened to the songs (to identify the part of the tune) and played them on MIDI controllers (real-time sequencing them) 20 times with different embellishments and capturing performance errors. This way, for each of the 20 original scores, 21 different variations have been built.

A 3-fold cross-validation scheme was carried out to perform the experiments, obtaining average success rates and dispersions $((max - min)/4)$.

Table 1. Success rates with the different approaches used

Approach	Success rate
PDTA	87.3 ± 0.7
StringBars	92.4 ± 1.1
Whole	86.9 ± 1.3

Table 1 shows the results of classification using the approaches explain in section 4.2. These results are compared with the results using the approach of probabilistic deterministic tree automata used in [1] but using the notation change described in section 4.

Note that PDTA uses the duration information implicit in tree representation, however the grammar approaches use less information (only pitch) for classifying. From the results, it is observed that the new approach using the strings through the StringsBars method improves significantly the PDTA results. The Whole approach did not improve the results because it is more sensitive to variations in the data than the geometric mean of the bar probabilities.

6 Conclusions

In this paper, we applied probabilistic tree grammars constructed from stochastic k-testable tree-models showing that this approach can be used for classifying new melodies represented by strings using the information captured in the grammar rules. This approach allows avoiding the duration information in the input data (strings with pitch only), making easier querying a music database. Our goal was to identify a melody from a set of different variations. The results overcame those previously obtained using probabilistic deterministic tree automata for the same corpus. According to the results, we can say that the classification is improved splitting the melody in bars. Also the results keep in good performance taking the string of the whole melody, which is important since not always the bar information is available. We are persuaded that these promising results can be improved by defining a more complex universal grammar for unseen labels and removing some rules that make the grammars more ambiguous.

Acknowledgements. This work is supported by the Spanish Ministry project TIN2009-14247-C02-02, TIN2009-14205-C04-C1, and the program Consolider Ingenio 2010 (CSD2007-00018).

References

1. Bernabeu, J.F., Calera-Rubio, J., Iñesta, J.M., Rizo, D.: A probabilistic approach to melodic similarity. In: Proceedings of MML 2009, pp. 48–53 (2009)
2. Chappelier, J.-C., Rajman, M.: A generalized cyk algorithm for parsing stochastic cfg. In: TAPD, pp. 133–137 (1998)
3. Charniak, E.: Tree-bank grammars. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 1031–1036 (1996)
4. Stephen Downie, J.: Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic n-grams as Text. PhD thesis, University of Western Ontario (1999)
5. Frazier, L., Rayner, K.: Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology 14(2), 178–210 (1982)
6. Illescas, P.R., Rizo, D., Iñesta, J.M.: Harmonic, melodic, and functional automatic analysis. In: Proc. of the 2007 International Computer Music Conference, vol. I, pp. 165–168 (2007)
7. Knuutila, T.: Inference of k-testable tree languages. In: Bunke, H. (ed.) Advances in Structural and Syntactic Pattern Recognition (Proc. of the S+SSPR 1992). World Scientific, Singapore (1993)
8. Ney, H., Essen, U., Kneser, R.: On the estimation of small probabilities by leaving-one-out. IEEE Trans. Pattern Anal. Mach. Intell. 17(12), 1202–1212 (1995)
9. Rico-Juan, J.R., Calera-Rubio, J., Carrasco, R.C.: Smoothing and compression with stochastic k-testable tree languages. Pattern Recognition 38(9), 1420–1430 (2005)
10. Rizo, D., Lemström, K., Iñesta, J.M.: Tree representation in combined polyphonic music comparison. In: Ystad, S., Kronland-Martinet, R., Jensen, K. (eds.) CMMR 2008. LNCS, vol. 5493, pp. 177–195. Springer, Heidelberg (2009)
11. Stolcke, A.: An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. Computational Linguistics 21, 165–201 (1995)
12. Verdu-Mas, J.L., Carrasco, R.C., Calera-Rubio, J.: Parsing with probabilistic strictly locally testable tree languages. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(7), 1040–1050 (2005)
13. Zalcstein, Y.: Locally testable languages. J. Comput. Syst. Sci. 6(2), 151–167 (1972)

A Tree Classifier for Automatic Breast Tissue Classification Based on BIRADS Categories

Noelia Vállez[1,2], Gloria Bueno[1], Oscar Déniz-Suárez[1], José A. Seone[2], Julián Dorado[2], and Alejandro Pazos[2]

[1] VISILAB, E.T.S.I.I, Universidad de Castilla-La Mancha, Spain
[2] RNASA-IMEDIR, Universidade a Coruña, Spain

Abstract. Breast tissue density is an important risk factor in the detection of breast cancer. It is also known that interpretation of mammogram lesions is more difficult in dense tissues. Therefore, getting a preliminary tissue classification may aid in the subsequent process of breast lesion detection and analysis. This article reviews several classification techniques for two datasets, both digitized screen-film (SFM) and full-field digital (FFDM) mammography, classified according to BIRADS categories. It concludes with a tree classification procedure based on the combination of two classifiers on texture features. Statistical analysis to test the normality and homoscedasticity of the features was carried. Thus, just features that are significant influenced by the tissue type were considered. The results obtained on 322 mammograms of the SFM dataset and on 1137 mammograms of the FFDM dataset demonstrate that up to 80% of samples were correctly classified using using 10-fold cross-validation to train and test the classifiers.

1 Introduction

Breast cancer continues to be an important health problem. Early detection can potentially improve breast cancer prognosis and significantly reduce female mortality. CAD systems can be of tremendous help to radiologists in the detection and classification of breast lesions. The development of reliable CAD systems is an important and challenging task. The automated interpretation of mammogram lesions remains a difficult task. The presence of dense breast tissue is one of the potential problems. Dense tissue may cause suspicious areas to be almost invisible and may be easily misinterpreted as calcification [1], [2]. Since the discovery by Wolfe [3] of the relation between mammographic parenchymal patterns and the risk of developing breast cancer in 1976, there has been a heightened interest in investigating breast tissue. A good review of the work done on breast tissue classification can be found in [4], [5].

This research has been prompted by this need to classify breast tissue and drive the development of CAD algorithms for automatic analysis of breast lesions. In our study several classification methods have been compared, and a hierarchical classification procedure combined with principal component analysis (PCA) on texture features is proposed as the best solution. Statistical analysis

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 580–587, 2011.

to test the normality and homoscedasticity of the data was carried out for features selection. Thus, just features that are significant influenced by the tissue type were considered. Experimental results have been given on different mammograms with various densities and abnormalities.

Section 2 describes the methods and material used for this work. This include the feature extraction procedure applied to the classifiers, the tested classifiers, the data training and testing carried on, as well as the experimental database used. Section 3 describes the results obtained with the proposed method and finally, in Section 5 the main conclusions are drawn.

2 Methods and Materials

There are several classification techniques to classify datasets according mammographic breast density [6]. We use the American College of Radiology BIRADS that has been used in a number of studies and is the most common technique used in the USA [7]. In this classification datasets have been classified according to 4 categories. These are: T.I) fatty, T.II) fatty-glandular or fibroglandular, T.III) heterogeneously dense and T.IV) extremely dense.

To deal with breast tissue classification problem several studies have been described in the literature. These studies are based on: a) the use of grey-level histograms and b) texture information extracted from different regions. Our proposal is to apply texture analysis on the whole breast tissue. Thus, all mammograms have been previously preprocessed to identify the breast region and remove the background and possible labels. This is illustrated in Figures 1 and 2 together with the tisue type classification.

2.1 Feature Extraction

Most studies on texture classification are based on morphological and texture features obtained from the image [8],[4] but without apply significancy statistical analysis. Here we analyze 241 features, both 1^{st} and 2^{nd} order texture statistics, drawn from the preprocessed image histogram and the co-occurrence matrix based features. The latter are the Haralicks coefficients, and they are calculated for a distance parameter $d = 1$, 3 and 5 at 0°, 45°, 90° and 135°. Once all texture features are calculated an statistical analysis of them is done to test the normality, the homoscedasticity and the influence of the tissue type over the obtained features respectively. However, although some calculated features have been obviated the total number still remains high (237 and 192 for the SFM and the FFDM databases respectively). Large numbers of features could make the accuracy and the computational time of the classifiers worse. Then, to reduce and select the feature space a PCA was applied.

Different tests were performed by varying the number of components from the space reduced by the PCA. This number of components varies between 10 and 190 at intervals of 10. The average errors for all classifiers were measured at each interval. These average errors range between 42% and 45%. The minimun error corresponds to the reduction of the space to 20 components.

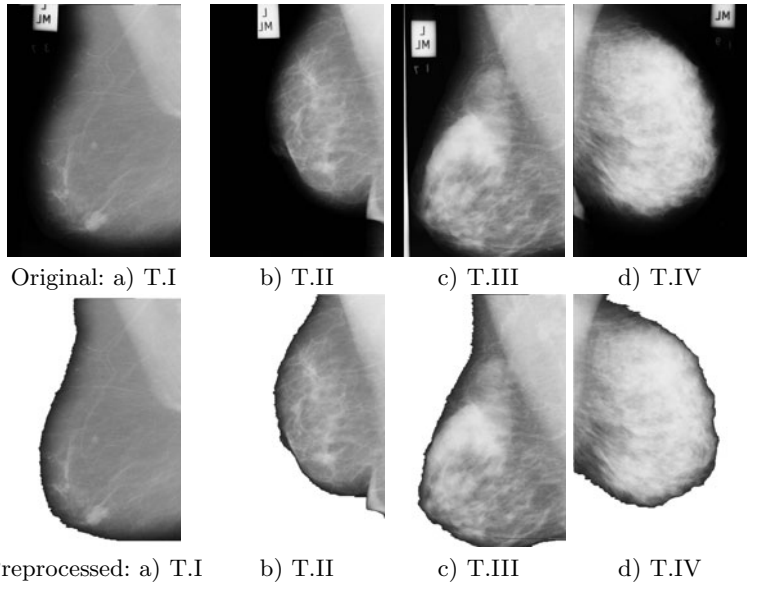


Fig. 1. BIRADS tissue classification and preprocessed images from the SFM dataset

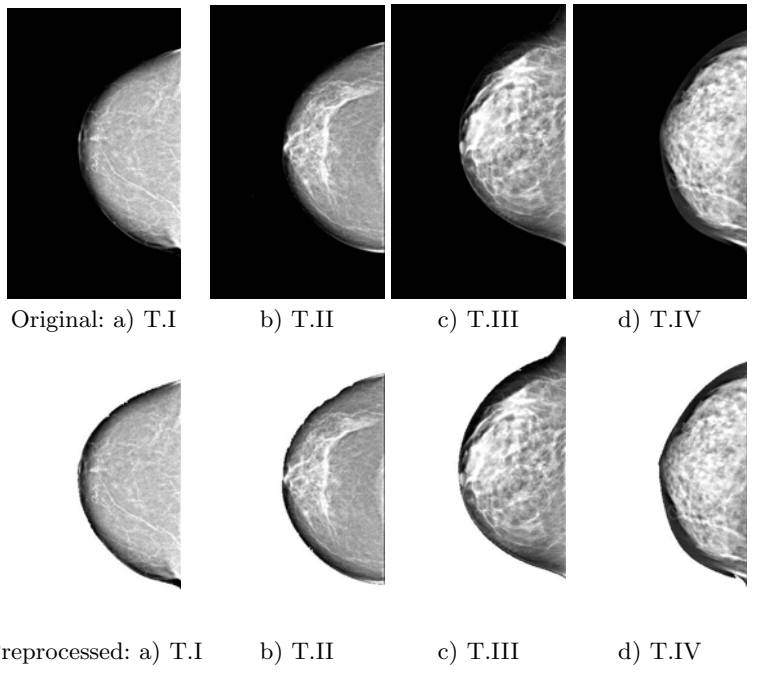


Fig. 2. BIRADS tissue classification and preprocessed images from the FFDM dataset

Table 1. Feature Reduction SFM Dataset

Feature	PCA Selection	Intra/Inter-cluster Distance Selection
1^{er} cuartil	$percentil25$	
3^{er} cuartil		$percentil75$
Contrast	$d = 3$ with $a = 0°, 45°, 135°$	$d = 5$ with $a = 0°, 45°, 135°$
	$d = 5$ with $a = 0°, 45°, 135°$	
Variance		$d = 3$ with $a = 45°$
		$d = 5$ with $a = 0°, 45°, 90°, 135°$
Sum Variance		$d = 3$ with $a = 0°$
		$d = 5$ with $a = 45°, 135°$
Difference Variance	$d = 3$ with $a = 45°$	$d = 5$ with $a = 0°, 45°, 135°$
Difference Entropy	$d = 3$ with $a = 0°$	
	$d = 5$ with $a = 0°, 90°, 135°$	
Correlation Inf. 2	$d = 1$ with $a = 0°, 45°, 90°, 135°$	
Homogeneity 1	$d = 1$ with $a = 90°$	
Homogeneity 2	$d = 1$ with $a = 90°$	
Cluster Shade		$d = 1$ with $a = 0°, 90°$
Autocorrelation		$d = 5$ with $a = 0°, 45°, 135°$
Dissimilarity	$d = 5$ with $a = 0°, 135°$	

Table 2. Feature Reduction FFDM Dataset

Feature	PCA Selection	Intra/Inter-cluster Distance Selection
Range		range
Maximum		max
Minimum		min
Asimmetry		asimmetry
Variance	variance	
Intercuartile Range		int. range
Correlation Inf. 1	$d = 1$ with $a = 45°, 135°$	$d = 1$ with $a = 0°$
		$d = 5$ with $a = 45°$
Correlation Inf. 2		$d = 1$ with $a = 0°$
		$d = 5$ with $45°$
Variance		$d = 3$ with $a = 90°$
		$d = 5$ with $a = 90°$
Correlation		$d = 5$ with $a = 90°, 135°$
Autocorrelation		$d = 5$ with $a = 0°, 135°$
Contrast	$d = 1$ with $a = 0°, 45°, 90°, 135°$	$d = 1$ with $a = 90°$
		$d = 3$ with $135°$
Difference Variance	$d = 1$ with $a = 0°$	$d = 5$ with $a = 0°$
	$d = 5$ with $45°$	
Sum Variance		$d = 1$ with $a = 45°$
Cluster Prominence	$d = 1$ with $a = 0°, 45°, 90°, 135°$	
	$d = 3$ with $a = 0°, 45°, 90°, 135°$	
	$d = 5$ with $a = 0°, 45°, 90°, 135°$	
Cluster Shade		$d = 3$ with $a = 135°$

The discriminatory power of these features was also analyzed using a feature ranking of individual performance for each classification method. This evaluation is based on the results from intra/inter-cluster distances between the 4

tissue types. These distances measure the variability within and between different classes. This identifies the features that maximize inter-cluster distance and minimize the intra-cluster distance. The 20 most significant features for the PCA and for the feature ranking are indicated in Tables 1 and 2.

2.2 Classifiers, Data Training and Testing

Different classification methods were tested on the selected features. These methods were: support vector machine (SVM) with polynomial, minkowski distance, exponential, radial_basis and sigmoid kernels, neural networks (feedforward, backpropagation, perceptron and radial basis) (NN), k-NN with k = 1, linear bayes normal (LBN), quadratic (QD) and tree-classifier with two layers. The best results obtained for the SVM were with a polynomial kernel and for the NN were with the backpropagation (BPNN), and these are shown here.

To train and test classifiers a non-biased evaluation following a woman-based cross-validation based on a k-fold cross-validation with $k = 10$ is used. This method consists of randomly dividing the data into k different groups containing approximately the same number of samples. One of these groups is selected to train the classifier and tests are performed on the rest of the groups. The process is then repeated with the other $k - 1$ groups of the dataset. The performance of these classifiers are shown and discussed in section 3.

2.3 Experimental Database

Two datasets were considered. One composed of 322 screen-film mammography (SFM) obtained from the MIAS public database [9] and another one composed of 1137 full-field digital mammography (FFDM) provided by local Hospitals. We focus our attention on the use of a FFDM dataset. However, the SFM dataset was used to compare our results with those of other authors. Both datasets were labeled according to the BIRADS categories by expert clinicians from the General Hospital of Ciudad Real. The image sizes are 3328×4084 and 1024×1024 respectively for the FFDM and SFM datasets.

The MIAS database contains images from medial-lateral projections (RMLO, LMLO) of 161 different cases while the FFDM contain in most cases the four projections, medial-lateral as well as cranial-caudal (RCC, LCC).

3 Results

Tables 3 and 4 show the results of the classifiers for the SFM and the FFDM datasets using a 10-fold cross-validation method to train and test the classifiers. The results are given with the complete dataset, with 20 features selected by the inter/intra cluster criteria and with 20 features of the PCA. We selected 20 features since the best results were obtained with this number of features.

On average, and weighted with respect to the number of mammograms of each type, the classification with PCA provides better results. The best classifier using 10-fold cross-validation are LBN with up to 69% agreement for SFM dataset and

Table 3. % of SFM mammograms correctly classified using 10-fold cross-validation

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	61%	43%	81%	80%	72%	58%	79%
T.II	57%	55%	73%	67%	55%	48%	68%
T.III	41%	38%	60%	53%	47%	37%	69%
T.IV	20%	18%	60%	67%	16%	39%	43%

(a) All Features

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	82%	86%	76%	86%	67%	53%	72%
T.II	68%	74%	70%	77%	75%	43%	66%
T.III	51%	51%	51%	43%	60%	42%	54%
T.IV	43%	50%	50%	59%	59%	36%	59%

(b) with Inter/Intra Cluster Selection

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	84%	83%	80%	82%	69%	85%	72%
T.II	71%	71%	72%	66%	56%	76%	66%
T.III	52%	46%	61%	54%	67%	51%	54%
T.IV	45%	52%	57%	66%	57%	41%	60%

(c) with PCA

FISH with up to 72% for the FFDM dataset. Analysing by tissue type, for the SFM dataset the best classifiers are the SVM with inter/intra cluster distance selected criteria for T.I and T.II, the BPNN with all the selected features for T.III and SVM with PCA for T.IV. For the FFDM dataset the best classifiers for each tissue type are FISH for T.I and T.IV, LBN for T.II and BPNN for T.III, where FISH and LBN are applied with inter/intra cluster distance selection criteria and BPNN is applied over all the selected features.

A 2-layer tree classifier was also tested using the best classifiers obtained previously. That is, the 1^{st} layer is composed of the {T.I ∪ T.II, T.III ∪ T.IV} and the 2^{nd} one of {{T.I, T.II}}; {{T.IIII, T.IV}. The results for the SFM dataset improved upon the previous ones, obtaining up to up to 90% in the 1^{st} layer with LBN, and 75% in the 2^{nd} layer with SVM. For the FFDM dataset we obtain up to 87% in the 1^{st} layer with FISH, and 75% in the 2^{nd} layer with BPNN by means of 10-fold cross-validation and both with PCA (see Table 5). For the SFM dataset we obtain an average of 76% TPD for T.I, 80% for T.II, 78% for T.III and 57% for T.IV, and 6.7% FPD for T.I, 10% for T.II, 10.8% for T.III and 5.7% for T.IV, with 10-fold cross-validation for testing and training. For the FFDM dataset we obtain an average of 79% TPD for T.I, 70% for T.II, 70% for T.III and 80% for T.IV, and 3.5% FPD for T.I, 7.3% for T.II, 1.3% for T.III and 9.9% for T.IV.

The difference between the two datasets is due to the use of 4 different projections in the FFDM dataset. While the two CC projections do not contain

Table 4. % of FFDM mammograms correctly classified using 10-fold cross-validation

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	88%	73%	78%	80%	43%	48%	83%
T.II	51%	55%	42%	35%	0%	33%	39%
T.III	61%	42%	52%	49%	1%	41%	69%
T.IV	83%	54%	65%	61%	63%	41%	67%

(a) All Features

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	90%	84%	86%	87%	80%	64%	84%
T.II	47%	54%	59%	52%	38%	48%	45%
T.III	65%	60%	65%	62%	49%	42%	67%
T.IV	87%	74%	77%	79%	81%	54%	68%

(b) with Inter/Intra Cluster Selection

Types	FISH	LOG	LBN	SVM	QD	k-NN	BPNN
T.I	89%	84%	78%	80%	72%	49%	88%
T.II	52%	58%	49%	38%	36%	33%	46%
T.III	65%	64%	65%	50%	58%	42%	64%
T.IV	83%	79%	61%	62%	60%	41%	76%

(c) with PCA

the pectoral muscle tissue, the MLO projections do contain this muscle. This muscle is usually denser than the rest of the breast tissue and influences the classification because half of the images do not contain it.

Table 5. Tree classifier with PCA (a) SFM (b) FFDM

Types	1^{st} Layer	2^{nd} Layer
T.I	91%	76%
T.II		89%
T.III	86%	78%
T.IV		57%

(a) LBN+SVM

Types	1^{st} Layer	2^{nd} Layer
T.I	82%	79%
T.II		70%
T.III	92%	70%
T.IV		80%

(b) FISH+BPNN

In order to improve the results and test the performance and accuracy of the classifiers, we are doing different ongoing tests. These include to combine classifiers by using bagging and rule combination (mean, median and voting).

4 Discussion and Conclusions

In this work a hierarchical procedure based on statistically selected texture features has been proposed for breast tissue classification. There are just three

works in the literature presenting breast tissue classification according to BI-RADS categories on SFM. Their overall correct classification is about 71% [8], 76% [10] without tissue segmentation and 82% [4] with it. None of them present results in FFDM dataset. Our approach reflect up to 89% of samples correctly classified for the SFM dataset and 80% for the FFDM one.

Acknowledgements

The authors acknowledge partial financial support from the Spanish Research Ministry through project RETIC COMBIOMED.

References

1. Ursin, G., Hovanessian-Larsen, L., Parisky, Y.R., et al.: Greatly increased occurrence of breast cancers in areas of mammographically dense tissue. Breast Cancer Research 7(5), 605–608 (2005)
2. Brem, R., Hoffmeister, J., Rapelyea, J., et al.: Impact of breast density on CAD for breast cancer. American Journal of Roentgenology 184, 439–444 (2005)
3. Wolfe, J.N.: Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 37, 2486–2492 (1976)
4. Oliver, A., Freixenet, J., Martí, R., et al.: A novel breast tissue density classification methodology. IEEE Trans. on Inform. Techn. in Biomed. 12, 55–65 (2008)
5. Parthalain, N.M., Jensen, R., Shen, Q., Zwiggelaar, R.: Fuzzy-rough approaches for mammographic risk analysis. Intelligent Data Analysis 14, 225–244 (2010)
6. Yafee, M., Boyd, N.: Mammographic breast density and cancer risk: The radiological view. Gynecological Endocrinology 21(S1), 6–11 (2005)
7. Harvey, J., Bovbjerg, V.: Quantitative Assessment of Mammographic Breast Density: Relationship with Breast Cancer Risk. Radiology 230(1), 29–41 (2004)
8. Bovis, K., Singh, S.: Classification of mammographic breast density using a combined classifier paradigm. In: 4th IWDM, pp. 177–180 (2002)
9. Suckling, J., Partner, J., Dance, D.R., et al.: The mammographic image analysis society digital mammogram database. In: IWDM, pp. 211–221 (1994)
10. Petroudi, S., Kadir, T., Brady, M.: Automatic classification of mammographic parenchymal patterns: A statistical approach. In: Proc. IEEE Conf. Eng. Med. Biol. Soc., vol. 1, pp. 798–801 (2003)

Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option

Ajalmar R. da Rocha Neto[1], Ricardo Sousa[2],
Guilherme de A. Barreto[1], and Jaime S. Cardoso[2]

[1] Depto. Engenharia de Teleinformática, Universidade Federal do Ceará (UFC)
{ajalmar,guilherme}@deti.ufc.br,
http://www.deti.ufc.br/~{ajalmar,guilherme}
[2] INESC Porto, Faculdade de Engenharia da Universidade do Porto, Portugal
{rsousa,jaime.cardoso}@inescporto.pt,
http://www.inescporto.pt/~{rsousa,jsc}

Abstract. Computer aided diagnosis systems with the capability of automatically decide if a patient has or not a pathology and to hold the decision on the dificult cases, are becoming more frequent. The latter are afterwards reviewed by an expert reducing therefore time consuption on behalf of the expert. The number of cases to review depends on the cost of erring the diagnosis. In this work we analyse the incorporation of the option to hold a decision on the diagnostic of pathologies on the vertebral column. A comparison with several state of the art techniques is performed. We conclude by showing that the use of the reject option techniques is an asset in line with the current view of the research community.

Keywords: Computer Aided Diagnosis System, Pattern Recognition, Support Vector Machines, Reject Option.

1 Introduction

Over the last decade, we have been assisting to an increasing number of Machine Learning (ML) techniques, such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN), applied to several medical fields. One reason for this behaviour relies on the capacity of human diagnostic, which is significantly worse than the neural system's diagnostic under adverse conditions, as stress, fatigue and little technical knowledge [2].

In the literature there are some reviews on the application of machine learning techniques in medicine. Some studies regarding to this problem describe several applications over different fields, namely: cardiology, ECG analysis, gastroenterology, pulmonology, oncology, neurology, EEG analysis, Otolaryngology, gynecology and obstetrics, ophthalmology, radiology, pathology, cytology, genetics, biochemistry among others. In another study, related to the use of ANN in the medical and biological areas, it is shown the distribution of a significant number of articles (more than 800) produced in the years 2000 and 2001 in more than 40 countries [7].

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 588–595, 2011.

Moreover, recent studies show the use of ANNs in the recovery of images of the vertebral column based in contents [8] , in the modality classification of medical images [5] and in the classification of individuals with normal or abnormal osteophyte pathology [7].

Although the use of ML techniques is already widespread in Medicine Diagnosis, in general the application of these techniques in Traumatic Orthopedics is rather sparse in the literature. This fact is due to the absence of numerical attributes that quantitatively describe the pathologies of interest to the field of orthopedics, to generate a suitable database for the design of classifiers [7].

As already stated, the incorporation of ML techniques on medical decision aiding procedures is becoming more frequent. Due to this common acceptance, Dreiseitl et al. [4] studied the credibility that physicians give to decision making support systems. By analysing the physicians' reaction when the system contradicts to their diagnostic, they noticed that specialists are very susceptible to accept the recommendations from these kinds of systems.

However, on complex cases with very similar attributes, these kind of systems can become unreliable. As consequence, the automation of decisions in these situations lead invariably to many wrong predictions. On the other hand, and although items in the historical data are labelled *only* as 'good' or 'bad', 'normal' or 'abnormal' pathology, the deployment of a decision support system in many environments has the opportunity to label critical items for manual revision, instead of trying to automatically classify every and each item. The system automates only those decisions which can be reliably predicted, labelling the critical ones for a human expert to analyse. Therefore, the development of classifiers with a third output class, the reject class, in-between the good and bad classes, is attractive.

This work intends to present an auxiliary system to medical decision aiding. The framework of study to the incorporation of the reject option will be the Intelligent System for Diagnosis of Pathologies of Vertebral Column (SINPATCO) [8]. This framework is composed by three subsystems. Namely, graphical interface, pathologies classification and knowledge extraction.

1.1 Pathologies of the Vertebral Column

The vertebral column is a system composed by a group of vertebras, invertebrate discs, nerves, muscles, medulla and joints. The main functions of the vertebral column are as follow: (i) human body support axle; (ii) osseous protector of the spine medulla and nervous roots; and (iii) body's movement axles, making movement possible in three levels: frontal, sagittal and transversal.

This complex system can suffer dysfunctions that cause backaches with very different intensities. Disc hernia and spondylolisthesis are examples of pathologies of the vertebral column that cause intense pain. They result of small or several traumas in the column that gradually injures the structure of the inter-vertebral disc.

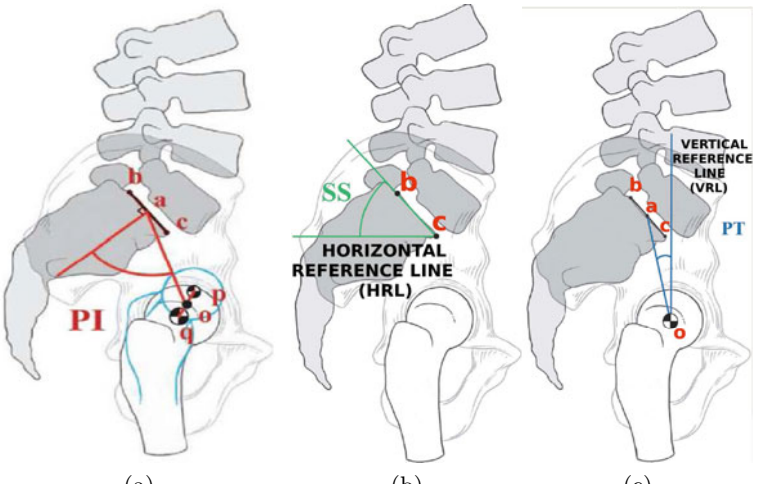
Disc hernia appears when the core of the inter-vertebral disc migrates from its place (from the center to the periphery of the disc). Once heading towards

the medullary channel or to the spaces where the nervous roots lie, this leads inevitability to their compression. Spondylolisthesis occurs when one of the 33 vertebras of the vertebral column slips in relation to the others. This slipping occurs generally towards the base of the spine in the lumbar region, causing pain or symptomatology irritation of the nervous roots. In the following section we will briefly describe characteristics (attributes) that are used to quantitatively describe each patient.

Biomechanical Attributes. The database applied in this work was kindly supplied by Dr. Henrique da Mota, who collected it during a medical residence in spine surgery at the *Centre Médico-Chirurgical de Réadaptation des Massues*, placed in Lyon, France. This database contains data about 310 patients obtained from sagittal panoramic radiographies of the spine. From this, 100 patients are volunteers that do not have any pathology in their spines (normal patients). The remaining data are from the patients operated due to disc hernia (60 patients) or spondylolisthesis (150 patients). Therefore, the database is composed of 210 abnormal patients.

Each patient in this database is represented as a vector (or pattern) with six biomechanical attributes, which correspond to the following parameters of the spino-pelvic system: angle of pelvic incidence, angle of pelvic tilt, lordosis angle, sacral slope, pelvic radius and grade of slipping. The correlation between the vertebral column pathologies and this attributes was originally proposed in reference [1].

Pelvic incidence (PI) is defined as an angle subtended by line $\overline{oa}$, which is drawn from the center of the femoral head to the midpoint of the sacral endplate and a line perpendicular to the center of the sacral endplate in Fig. 1a. The sacral



(a) (b) (c)

Fig. 1. Spino-pelvic system

endplate is defined by the line segment $\overline{bc}$ constructed between the posterior superior corner of the sacrum and the anterior tip of the $S1$ endplate at the sacral promontory. For the case when the femoral heads are not superimposed, the center of each femoral head is marked, and a connecting line segment will connect the centers of the femoral heads. Pelvic radius (RP) $\overline{ao}$ will be drawn from the center of this line to the center of the sacral endplate (Fig. 1a).

Lordosis angle is the bigger sagittal angle between the sacrum superior plate and the lumbar vertebra superior plate or thoracic limit. Sacral Slope (SS) is defined as the angle between the sacral endplate ($\overline{bc}$) and the horizontal reference line (HRL), in Fig. 1b, while Pelvic Tilt (PT) is defined as the angle between the vertical reference line (VRL) and the line joining the middle of the sacral endplate and the axis of the femoral heads in Fig. 1c. It is positive when the hip axis lies in front of the middle of the sacral endplate. Finally, the level of slipping is the percentage level of slipping between the inferior plate of the fifth lumbar vertebra and the sacrum.

The occurrence of pathologies in the vertebral column is conditioned to the morphological types of the pelvis-spine system. The pelvic incidence, being in an elevated level, is conditioned to a higher sacral slope, that generates increasing shear by the increase of the support plan inclination for lumbar lordosis, besides facilitating the conflict of posterior structures, leading to the appearing of a fracture of fatigue in the arc that supports the vertebra and generating a slope called Spondylitics. The low pelvic incidences lead to the contrary effect, with the occurrence of an increasing pressure in the intervertebral disc and facilitate the occurrence of degeneration and disc hernias. The incidence angle determines a normal condition.

The design of automatic classifiers based in biomechanical attributes of real clinical cases allows that linear and/or non-linear relations, as well as their influences in the diagnostic, are captured in a transparent way for the orthopedist, in a way to help him in the decision making. Following, we briefly describe the machine learning models evaluated in this work.

2 Problem Statement and Standard Solutions

Predictive modelling tries to find good rules (models) for guessing (predicting) the values of one or more variables (target) in a dataset from the values of other variables. Our target can assume only two values, represented by 'normal' and 'pathological' classes. When in possession of a "complex" dataset, a simple separator is bound to misclassify some points. The design of classifiers with reject option can be systematised in three different approaches:

- the design of two, *independent*, classifiers. A first classifier is trained to output $\mathcal{C}_{-1}$ only when the probability of $\mathcal{C}_{-1}$ is high and a second classifier trained to output $\mathcal{C}_{+1}$ only when the probability of $\mathcal{C}_{+1}$ is high.
- the design of a single, standard binary classifier (SBC). If the classifier provides some approximation to the a posterior class probabilities, then a pattern is rejected if the maximum of the two posterior probabilities is lower

than a given threshold. If the classifier does not provide probabilistic outputs, then a rejection threshold targeted to the particular classifier is used.
– the design of a single classifier with embedded reject option. This approach has consisted in the design of algorithms specifically adapted for this kind of problems [6, 3, 9].

3 An Ordinal Data Approach for Detecting Reject Regions

The rejection method to be presented—rejoSVM—is an extension of a method already proposed in the literature but for the classification of ordinal data. For more detail about this method, the reader should consult [3,9]. For completeness, we summarise here the rejoSVM model.

3.1 The Data Replication Method for Detecting Reject Regions

The scenario of designing a classifier with reject option shares many characteristics with the classification of ordinal data. It is also reasonable to assume for the reject option scenario that the three output classes are naturally ordered as $\mathcal{C}_1, \mathcal{C}_{reject}, \mathcal{C}_2$. In the scenario of designing a classifier with reject option, we are interested on finding two boundaries: a boundary discriminating $\mathcal{C}_1$ from $\{\mathcal{C}_{reject}, \mathcal{C}_2\}$ and a boundary discriminating $\{\mathcal{C}_1, \mathcal{C}_{reject}\}$ from $\mathcal{C}_2$.

We proceed exactly as in the data replication method for ordinal data [9]. We start by transforming the data from the initial space to an extended space, replicating the data, according to the rule:

$$\mathbf{x} \in \mathbb{R}^d \begin{cases} \nearrow [\begin{smallmatrix}\mathbf{x}\\ h\end{smallmatrix}] \in \mathbb{R}^{d+1} \\ \searrow [\begin{smallmatrix}\mathbf{x}\\ 0\end{smallmatrix}] \in \mathbb{R}^{d+1} \end{cases}, \text{ where } h = \text{const} \in \mathbb{R}^+$$

If we design a binary classifier on the extended training data, without further considerations, one would obtain the same classification boundary in both data replicas. Therefore, we modify the misclassification cost of the observations according to the data replica they belong to. In the first replica (the extension feature assumes the value zero), we will discriminate $\mathcal{C}_1$ from $\{\mathcal{C}_{reject}, \mathcal{C}_2\}$; therefore we give higher costs to observations belonging to class $\mathcal{C}_2$ than to observations belonging to class $\mathcal{C}_1$. This will bias the boundary towards the minimisation of errors in $\mathcal{C}_2$. Similar approach is conducted for the second replica.

3.2 Selecting the Misclassification Costs

The typical adoption of the same cost for erring and rejecting on the two classes is as follows: assign C_{low} cost when classifying a class as reject and assign C_{high} cost when misclassifying. Therefore, $C_{reject} = \frac{C_{low}}{C_{high}} = w_r$ is the cost of rejecting (normalised by the cost of erring). The data replication method with reject option tries to minimises the empirical risk $w_r R + E$, where R accounts for the rejection rate and E for the misclassification rate.

4 Experiments

The aim of our experimental study is to evaluate the interest of embedded reject options methodologies for aiding the diagnostic of pathology on the Vertebral Column. This dataset is thoroughly described in Section 1.1. For this study, we transformed our three class problem into a binary one. We aggregated the disc hernya and spondylolisthesis pathologies classes into a just one pathology class. The normal class remained unchanged.

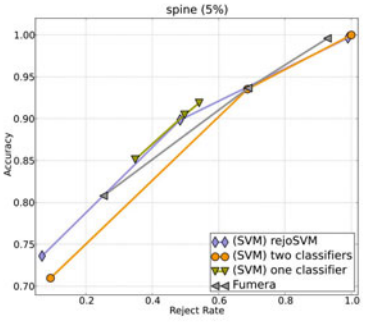
The training dataset was composed in different experiments with 5%, 40% and 80% of the data. The splitting of the data into training and test sets was repeated 100 times in order to obtain more stable results for accuracy by averaging and also to assess the variability of this measure. The best parametrisation of each model was found by 'grid-search', based on a 5-fold cross validation scheme conducted on the training set. In our experiments we used a linear kernel on all methods performing the 'grid-search' over the C parameter spanning values between $2^{-5}, \ldots, 2^{5}$. The C value is a penalty factor for each point misclassified. Finally, the error of the model was estimated on the test set.

The performance of a classifier with reject option can be represented by the classification accuracy achieved for any value of the reject rate (the so-called Accuracy-Reject curve). The trade-off between errors and rejections depends on the cost of a rejection w_r. This implies that different points of the A-R curve correspond to different values of w_r. We considered values of w_r less than 0.5, as above this value it is preferable to just try to guess randomly.

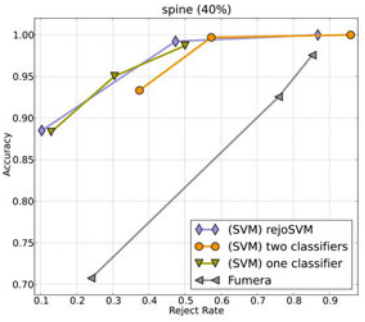
In Fig. 2 we present the results obtained for all used methods. We can first identify that as training size increases, both methods attain similar results with exception with Fumera technique (Fig. 2a-2c). This behaviour can be justified by the fact that the optimisation function is not convex and therefore not attaining the global optimum.

In order to emphasise the benefits of the incorporation of the reject option, we trained five standard learning algorithms. The training and testing evaluation was performed exactly as before. We have selected a SVM (with a linear kernel) with the C parameter encompassing the same range values (baseline learning method for the one and two classifiers). We have also used a SVM with the KMOD kernel [7]. Finally, we also conducted our experiments with a GRNN (General Regression Neural Network) and a MLP [7].

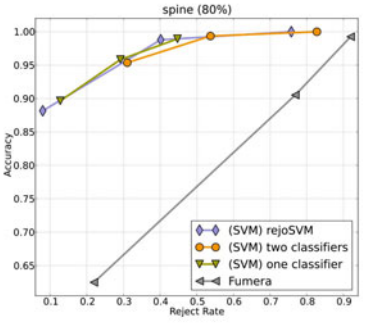
Considering further the performance evaluation function of the embedded reject option learning method, we can fairly compare all techniques. Table 1 shows that standard learning methods provide the baseline results for some of the reject option schemes. Within the SINPATCO context, the incorporation of a reject option can be an asset. Moreover, tools like SINPATCO are designed as decision aiding system which could be used on healthcare offices located on remote areas with limited access to modern resources and funding. In this way, systems with high rates of True Positive (sensitivity) and True Negatives (specificity) are required. Such techniques besides imposing high accuracies rates and a higher confidence on the diagnosis, they also avoid misclassifications. In doing so, there will not be any influence by SINPATCO on the expert to take wrong decisions



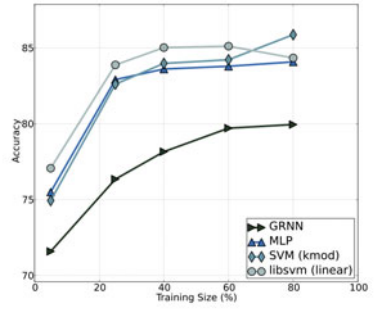
(a) A-R performance curves (Models trained with 5% of the full dataset)



(b) A-R performance curves (Models trained with 40% of the full dataset)



(c) A-R performance curves (Models trained with 80% of the full dataset)



(d) Several learning algorithms for the Spine dataset with the reject option (oSVM technique is detailed in [3])

Fig. 2. The A-R curves for the Spine dataset with different training sizes

Table 1. Performance rules according to measure $P = w_r R + E$ in p.p

Training Size	Method \wr	0.04	0.24	0.48	Method	Accuracy
40%	rejoSVM	96.5	87.9	83.5		
	one classifier	96.7	87.7	82.1	SVM (linear)	85.0
	two classifier	96.2	86.0	76.5	SVM (KMOD)	83.9
80%	rejoSVM	96.9	89.1	85.2		
	one classifier	97.1	88.8	84.4	SVM (linear)	84.3
	two classifier	96.6	86.3	82.3	SVM (KMOD)	85.9

which could lead to some intervations (being invasive or not). For instance, on the Fig. 2c, only 50% of the cases would have to be reviewed by the expert while the remaining would be correctly classified. As a final remark, we could verify that rejoSVM and two classifiers do not outperform the other. However, and as a feature of this work, rejoSVM benefits of simplicity and interpretability which could aid the medical expert in future evaluations.

5 Conclusion

Here we incorporate the reject technique proposed in [9] to the diagnosis of pathologies on the Vertebral Column. This incorporation of the reject option provided to be an asset obtained better results than traditional learning techniques, even when rejecting few instances. Finally, further studies can be developed on the analysis with the incorporation of the reject option on the original problem. Furthermore, a comparison with the ensemble learning technique as proposed in [7] should also be assessed.

Acknowledgments. This work has been partially supported by Fundação para a Ciência e a Tecnologia (FCT) - Portugal through project PTDC/EIA/64914/2006 and by CNPq - Brazil through Programa CNPq/Universidade do Porto/590008/2009-9.

References

1. Berthonnaud, E., Dimnet, J., Roussouly, P., Labelle, H.: Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. Journal of Spinal Disorders & Techniques 18(1), 40–47 (2005)
2. Brause, R.: Revolutionieren neuronale netze unsere vorhersagefigkeiten. Zentralblatt fr Chirurgie, 692–698 (1999)
3. Cardoso, J.S., da Costa, J.F.P.: Learning to classify ordinal data: the data replication method. Journal of Machine Learning Research 8, 1393–1429 (2007)
4. Dreiseitl, S., Binder, M.: Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artificial Intelligence in Medicine 33(1), 25–30 (2005)
5. Florea, F., Rogozan, A., Bensrhair, A., Darmoni, S.J.: Comparison of feature-selection and classification techniques for medical images modality categorization. Rapport Interne PSI No. 01/FIF (2004)
6. Fumera, G., Roli, F.: Support Vector Machines with Embedded Reject Option. In: Lee, S.-W., Verri, A. (eds.) SVM 2002. LNCS, vol. 2388, pp. 68–82. Springer, Heidelberg (2002)
7. Neto, A.R.R., Barreto, G.A.: On the application of ensembles of classifiers to the diagnosis of pathologies of the vertebral column: A comparative analysis. IEEE Transactions on Latin America 7(4), 487–496 (2009)
8. Neto, A.R.R., Barreto, G.A., Cortez, P.C., da Mota, H.: SINPATCO: Sistema inteligente para diagnóstico de patologias da coluna vertebral. In: Congresso Brasileiro de Automática, pp. 929–934 (2006)
9. Sousa, R., Mora, B., Cardoso, J.S.: An ordinal data method for the classification with reject option. In: Proceedings of The Eighth International Conference on Machine Learning and Applications, ICMLA 2009 (2009)

Language Identification for Interactive Handwriting Transcription of Multilingual Documents

Miguel A. del Agua, Nicolás Serrano, and Alfons Juan

DSIC/ITI, Universitat Politècnica de València
{mdelagua,nserrano,ajuan}@dsic.upv.es

Abstract. An effective approach to handwriting transcription of (old) documents is to follow a sequential, line-by-line transcription of the whole document, in which a continuously retrained system interacts with the user. In the case of multilingual documents, however, a minor yet important issue for this interactive approach is to first identify the language of the current text line image to be transcribed. In this paper, we propose a probabilistic framework and three techniques for this purpose. Empirical results are reported on an entire 764-page multilingual document for which previous empirical tests were limited to its first 180 pages, written only in Spanish.

Keywords: Language Identification, Interactive Handwriting Transcription, Multilingual Documents.

1 Introduction

Transcription of handwritten text in (old) documents is an important, time-consuming task for digital libraries. However, automated techniques for document image analysis and handwriting recognition are still far from perfect [4], and thus post-editing automatically generated output is not clearly better than simply ignoring it. Instead, a more effective approach is to follow a sequential, line-by-line transcription of the whole document, in which a continuously retrained system interacts with the user. In this way, the main task of the user is to (partially) supervise and correct, if needed, each new line transcription hypothesis of the system. This interactive handwriting transcription approach, also extended to interactive layout analysis and line detection, has been implemented in an open source tool called *Gimp-based Interactive transcription of old text DOCuments* (GIDOC) [8]. This tool was used to develop different techniques, such as, to better adapt models from partially supervised transcriptions [6], to properly balance error and supervision effort [7], as well as different active learning strategies to interact with the user on each new system hypothesis [5].

In the case of multilingual documents, however, a minor yet important issue for interactive transcription of a text line image (or an image block) is to first

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 596–603, 2011.

identify its corresponding language. A good example of multilingual document is the GERMANA database [3]. GERMANA is the result of digitizing and annotating a 764-page, single-author Spanish manuscript from 1891, solely written in Spanish up to page 180, but then also written in five other languages, especially Catalan and Latin. For simplicity, to avoid dealing with multiple languages, we limited ourselves to the first 180 pages of GERMANA in the empirical tests of the studies cited above.

To our knowledge, however, language identification for interactive transcription of multilingual documents remains unexplored. Indeed, conventional (non-interactive) script and language identification in handwritten documents is still in its early stage of research [2]. In this paper, after a brief review of GIDOC, we propose a probabilistic framework and three techniques for language identification in interactive transcription of multilingual documents (Section 3). In Section 4, empirical results are reported on the whole GERMANA manuscript. Conclusions drawn and future work are summarized in Section 5.

2 GIDOC Overview

GIDOC is a first attempt to provide user-friendly, integrated support for interactive-predictive page layout analysis, text line detection and handwritten text transcription [8]. It is built as a set of plug-ins for the well-known GNU Image Manipulation Program (GIMP), which has many image processing features already incorporated and, what is more important, a high-end user interface for image manipulation. To run GIDOC, we must first run GIMP and open a document image. GIMP will come up with its high-end user interface, which is often configured to only show the main toolbox and an image window. GIDOC can be accessed from the menubar of the image window (Fig. 1).

As shown in Fig. 1, the GIDOC menu includes six entries, though here only the last one, *Transcription,* is briefly described (see [8] for a more detailed description). The *Transcription* entry opens an interactive transcription dialog (also shown in Fig. 1), which consists of two main sections: the image section, in the middle part, and the transcription section, in the bottom part. A number of text line images are displayed in the image section together with their transcriptions, if available, in separate editable text boxes within the transcription section. The *current* line to be transcribed is selected by placing the edit cursor in the appropriate editable box. Its corresponding baseline is emphasized (in blue color) and, whenever possible, GIDOC shifts line images and their transcriptions so as to display the current line in the central part of both the image and transcription sections. It is assumed that the user transcribes or supervises text lines, from top to bottom, by entering text and moving the edit cursor with the arrow keys or the mouse.

Note that each editable text box has a button attached to its left, which is labeled with its corresponding line number. By clicking on it, its associated line image is extracted, preprocessed, transformed into a sequence of feature vectors, and Viterbi-decoded using character HMMs and a language model previously

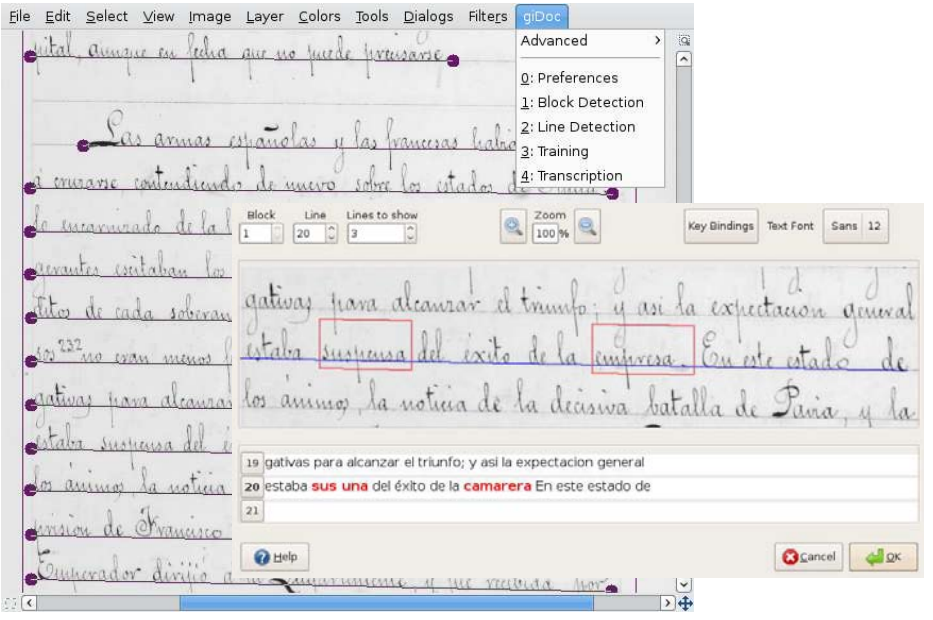


Fig. 1. Interactive Transcription dialog over an image window showing GIDOC menu

trained (see [8] for details on preprocessing, feature extraction, HMM-based image modeling and language modeling in GIDOC). As shown in Fig. 1, words in the current line for which the system is not highly confident are emphasized (in red) in both the image and transcription sections.

3 Probabilistic Framework

Let l be the number of the current text line image to be transcribed, and let x_l be its corresponding sequence of feature vectors. The task of our system is to predict first its (unknown) language identification label, c_l, and then its transcription, w_l. We assume that all preceding lines have been already annotated in terms of language labels, c_1^{l-1}, and transcriptions, w_1^{l-1}. By application of the Bayes decision rule, the minimum-error system prediction for c_l is:

$$c_l^*(x_l, c_1^{l-1}) = \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid x_l, c_1^{l-1}) \tag{1}$$

$$= \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1})\, p(x_l \mid \tilde{c}_l) \tag{2}$$

$$= \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1}) \sum_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)\, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{3}$$

$$\approx \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1}) \max_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)\, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{4}$$

where: in Eq. (2), it is assumed that x_l is conditionally independent of all preceding line language labels, c_1^{l-1}, given the current line language label, $\tilde{c}_l$; in Eq. (3), we consider all possible transcriptions for the language $\tilde{c}_l$, $W(\tilde{c}_l)$; and, in Eq. (4), the Viterbi (maximum) approximation to the sum in Eq. (3) is applied to only consider the most likely transcription.

The decision rule (4) requires a *language identification model* for $p(\tilde{c}_l \mid c_1^{l-1})$ and, for each possible language $\tilde{c}_l$, a $\tilde{c}_l$-dependent *language (transcription) model* for $p(\tilde{w}_l \mid \tilde{c}_l)$ and a $\tilde{c}_l$-dependent *image model* for $p(x_l \mid \tilde{c}_l, \tilde{w}_l)$. As done in language modeling for monolingual documents, the language models in the multilingual case, both for identification and transcription, can be implemented in terms of *n-gram language models* [8]. Those for language-dependent transcription can be implemented as usual in the monolingual case though, in our case, each language $\tilde{c}_l$ will have its own n-gram language model, trained only from available transcriptions labeled with $\tilde{c}_l$. Regarding the n-gram language identification model, $p(\tilde{c}_l \mid c_1^{l-1})$, in this paper we propose and compare three rather simple techniques:

1. A *bigram* model estimated by relative frequency counts:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \frac{N(c_{l-1}\tilde{c}_l)}{N(c_{l-1})} \tag{5}$$

2. A *unigram* model also estimated by relative frequency counts:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \frac{N(\tilde{c}_l)}{l-1} \tag{6}$$

3. And a *"copy the preceding label" (CPL)* bigram model:

$$\hat{p}(\tilde{c}_l \mid c_{l-1}) = \begin{cases} 1 & \tilde{c}_l = c_{l-1} \\ 0 & \tilde{c}_l \neq c_{l-1} \end{cases} \tag{7}$$

where $N(\cdot)$ denotes the number of occurrences of a given event in the preceding lines, such as the bigram $c_{l-1}\tilde{c}_l$ or the unigram $\tilde{c}_l$. Note that (5) and, especially (7), assume that consecutive text lines are usually written in the same language. This is not necessarily true though, in the kind of manuscripts (applications) we have in mind (e.g GERMANA), it is a reasonable assumption.

Also as in the monolingual case, the *image models* for the different languages can be implemented in terms of *character HMMs* [8]. Moreover, if only a single script is used for all the languages considered (e.g. Latin), then a single, shared image model for all of them might produce better recognition results than a separate, independent model for each language. Clearly, this can be particularly true for infrequent languages.

Finally, it is often useful in practice to introduce scaling parameters in the decision rule so as to empirically adjust the contribution of the different models involved. In our case, the decision rule given in Eq. (4) can be rewritten as

$$c_l^*(x_l, c_1^{l-1}) \approx \underset{\tilde{c}_l}{\operatorname{argmax}}\ p(\tilde{c}_l \mid c_1^{l-1})^{\beta} \max_{\tilde{w}_l \in W(\tilde{c}_l)} p(\tilde{w}_l \mid \tilde{c}_l)^{\alpha_{\tilde{c}_l}}\, p(x_l \mid \tilde{c}_l, \tilde{w}_l) \tag{8}$$

where we have introduced an *Identification Scale Factor (ISF)* β and, for each language $\tilde{c}_l$, a language-dependent *Grammar Scale Factor (GSF)* $\alpha_{\tilde{c}_l}$. Obviously, Eq. (8) does not differ from Eq. (4) when all these scaling parameters are simply set to unity. In the experiments reported below, these parameters will be adapted from a validation set.

4 Experiments

As indicated in the introduction, the experiments reported here were carried out on a multilingual, single-author manuscript from 1891 known as GERMANA database [3]. Our main goal is to empirically compare the three language identification techniques described in the preceding section. Moreover, we provide recognition results on the complete manuscript, which is also worth noting since previous results on GERMANA have been limited to its first 180 pages, solely written in Spanish. The complete manuscript, which comprises a total of 764 pages, includes five other languages, most notably Catalan and Latin.

Some basic yet precise statistics of GERMANA are given in Table 1. In terms of running words, Spanish comprises about 81% of the document, followed by Catalan (12%) and Latin (4%), while the other three languages only account for less than a 3%. Similar percentages also apply for the number of lines. In terms of lexicons, it is worth noting that Spanish and, to a lesser extent, Catalan and Latin, have lexicons comparable in size to standard databases [3]. Also note that the sum of individual lexicon sizes (29.9K) is larger than the size of the global lexicon (27.1K). This is due to presence of words common to different languages, such as common words in Spanish and Catalan. On the other hand, singletons, that is, words occurring only once, account for most words in each lexicon ($55\% - 71\%$). It goes without saying that, as usual, language modelling is a difficult task. To be more precise, in Table 1 we have included the global perplexity and the perplexity of each language, as given by a bigram model on a 10-fold cross-validation experiment.

Table 1. Basic statistics of GERMANA

Language	Lines	Words	Lexicon	Singletons	Perplexity
All	20000	217K	27.1K	57.4%	289.8±17.0
Spanish	80.9%	81.4%	19.9K	55.6%	238.1±27.7
Catalan	11.8%	12.4%	4.6K	63.2%	112.9±61.6
Latin	4.6%	3.8%	3.4K	69.2%	211.1±51.3
French	1.3%	1.4%	1.1K	71.1%	88.3±21.0
German	1.1%	0.7%	0.6K	52.7%	92.1±29.2
Italian	0.3%	0.3%	0.3K	67.3%	63.3±14.4

We divided GERMANA into 40 blocks of 500 lines each. The first block was fully transcribed and an initial system was trained from it. Then, from block 2 to 40, each new block was recognized by the system trained from all preceding

blocks, with the last block being also used as a validation set for parameter adaptation. It is worth noting that, after recognition of each block, the user supervises and, if needed, corrects both, language identification labels and transcriptions.

As a baseline, we first tried a conventional, *monolingual* system, that is, a system assuming that all lines come from a single language, and thus only requiring one language (transcription) model and one image model. On the other hand, we tried four *multilingual* systems which only differ in the way they identify the language of the current line: *supervised* (manually given), *bigram* (using Eq. (5)), *unigram* (using Eq. (6)) and *CPL* (using Eq. (7)). Clearly, in all these four systems, a different language (transcription) model was required for each of the 6 languages in GERMANA. However, as suggested at the end of the preceding section, a single, shared image model was used instead of a separate, independent image model for each language in GERMANA. The results are plotted in Fig. 2, in terms of *Word Error Rate* (WER) of the recognized text up to the current line.

As expected, the multilingual systems achieve better results than the monolingual system. Also as expected, the correct language identification label (supervised) produces better results than an automatic, error-prone technique such as CPL. Surprisingly, however, the unigram and, to a lesser extent, the bigram identification techniques achieve better results than manual supervision. In other words, it is sometimes preferable not to use the correct, but probably poorly-trained language (transcription) model, and use instead a well-trained model for

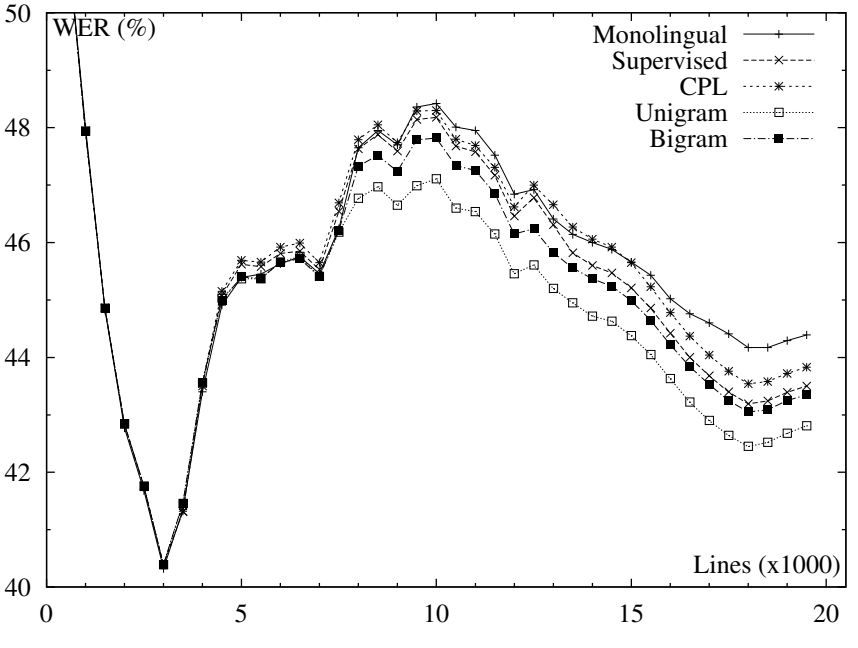


Fig. 2. WER in GERMANA as a function of the number of recognized lines

a different yet close language (e.g. Catalan and Spanish). On the other hand, it can be also observed that there are certain blocks at which the WER curve abruptly changes from a (smooth) decreasing tendency to a rapid increase. This was studied carefully in [1] by decomposing the (total) WER curve into its corresponding language-dependent WER curves. It was found that these abrupt changes are due to the occurrence of text from previously unseen languages, most notably Catalan (from line 3500) and Latin (from line 4000).

Although optimal (supervised) language identification does not necessarily lead to better recognition results than those obtained with suboptimal (imperfect) identification techniques, it is still important to have an identification technique of minimal error, maybe to just minimize user effort while correcting identification errors. Table 2 shows the Identification Error Rate (IER) of the proposed techniques for all and each language in GERMANA and both, in absolute and relative terms.

Table 2. Identification Error Rate (IER) on GERMANA for the techniques proposed

		IER (absolute)			IER (%)		
Language	Lines	2-gram	1-gram	CPL	2-gram	1-gram	CPL
All	19500	1290	2183	488	6.6	11.2	2.5
Spanish	15725	243	312	224	1.5	2.0	1.4
Catalan	2414	534	1136	181	22.1	47.1	7.5
Latin	951	255	409	49	26.8	43.0	5.2
French	266	116	182	31	43.6	68.4	11.7
German	76	74	76	2	97.4	100.0	2.6
Italian	68	68	68	1	100.0	100.0	1.5

From the results in Table 2, it becomes clear that the simplest technique, CPL, is also the most accurate. It achieves an IER of 2.5%, that is, on average, only 3 identified labels out of 100 need to be corrected by the user. In contrast, the 1-gram and 2-gram techniques clearly fail in identifying languages other than Spanish. This might be due to the fact that scaling parameters were adapted to minimize the WER instead of the IER and, indeed, these techniques provided better results than CPL in terms of WER.

5 Conclusions

We have proposed a probabilistic framework and three techniques for language identification in interactive transcription of multilingual documents. These techniques are called the bigram, unigram and CPL-bigram models. They have been empirically compared on the whole GERMANA database, a 764-page, single-author manuscript from 1891 written in six different Latin languages, mainly Spanish. According to the empirical results, the simplest technique, that is, the "copy the preceding label" (CPL) bigram model is also the most accurate.

Acknowledgements. Work supported by the EC (FEDER, FSE), the Spanish Government (MICINN, MITyC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010", MITTRAL TIN2009- 14633-C03-01 and FPU AP2007-02867), the Generalitat Valenciana (grant Prometeo/2009/014 and ACOMP/2010/051) and the UPV (grant 20080033).

References

1. del Agua, M.A.: Multilingualidad en el reconocimiento de texto manuscrito. Final Degree Project (2010)
2. Ghosh, D., Dube, T., Shivaprasad, P.: Script Recognition: A Review. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 32(12), 2142–2161 (2010)
3. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos-Terrades, O., Juan, A.: The GERMANA database. In: Proc. of the 10th Int. Conf. on Document Analysis and Recognition (ICDAR 2009), Barcelona, Spain, pp. 301–305 (2009)
4. Plötz, T., Fink, G.: Markov models for offline handwriting recognition: a survey. Int. J. on Document Analysis and Recognition (IJDAR) 12(4), 269–298 (2009)
5. Serrano, N., Giménez, A., Sanchis, A., Juan, A.: Active learning strategies in handwritten text recognition. In: Proc. of the 12th Int. Conf. on Multimodal Interfaces and the 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing (China), vol. 86 (November 2010)
6. Serrano, N., Pérez, D., Sanchis, A., Juan, A.: Adaptation from Partially Supervised Handwritten Text Transcriptions. In: Proc. of the 11th Int. Conf. on Multimodal Interfaces and the 6th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2009), Cambridge, MA (USA), pp. 289–292 (2009)
7. Serrano, N., Sanchis, A., Juan, A.: Balancing error and supervision effort in interactive-predictive handwriting recognition. In: Proc. of the Int. Conf. on Intelligent User Interfaces (IUI 2010), Hong Kong (China), pp. 373–376 (2010)
8. Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., Juan, A.: The GIDOC prototype. In: Proc. of the 10th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2010), Funchal (Portugal), pp. 82–89 (2010)

vManager, Developing a Complete CBVR System

Andrés Caro, Pablo G. Rodríguez, Rubén Morcillo, and Manuel Barrena

University of Extremadura, Computer Science Dept.
Escuela Politécnica, Av. Universidad s/n, 10003 Cáceres, Spain
{andresc,pablogr,rmorcillo,barrena}@unex.es
http://gim.unex.es

Abstract. Content-Based Video Retrieval (CBVR) is a research area that has drawn a good deal of attention in recent years. The ability to retrieve videos similar to a given one in terms of implicit features (mainly pictorial features) and/or explicit characteristics (eg. semantic context) are the cornerstones of this growing interest. In this paper we present the results obtained within the project *vManager*, a CBVR system based on local color and motion signatures, our own video representation and different metrics of similarity among videos. The results for real videos point to promising advances, not only as regards the effectiveness of the system, but also in terms of its efficiency.

Keywords: CBVR, retrieval, local color, video.

1 Introduction to CBVR Systems

In these past years, digital video processing has become a growing area of interest. The ability to retrieve videos efficiently and effectively can become a challenge from multiple points of view. A Content-Based Video Retrieval (CBVR) system can obtain efficient processes of indexing, search, and frame and scene (groups of frames) retrieval. In addition, these processes are automatic and objective.

In general, digital video management involves the reduction of the total set of frames of a video to a smaller subset, by detecting scene breaks [1][2]. Some authors use histogram-based techniques for the detection of scenes [3][4][5]. Others use techniques based on the MPEG standard [6]. The movement and algorithms based on hierarchical clustering are also used in [7][8][9]. A representative keyframe is selected for every scene of the video. The first frame of each scene [4][6], the central frame [7], a mixture of first, last and central [10] or all the frames of the scene [11] can be used as keyframes. Thus, content-based feature vectors are obtained, to represent these keyframes in indexes and retrievals. This development greatly relieves the video processing tasks.

Each keyframe is represented by a feature vector. Some authors have used criteria based on color [5] or color and motion [3][6][12]. These methods are simple and efficient, but may not provide optimal coverage of the scenes. Other authors

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 604–611, 2011.

use feature vectors based on color and texture [10] or color, shapes and textures [4][8][11]. Methods that combine spatial and temporal information (changes in intensity on consecutive frames) have also been proposed [13]. In this paper we extract color-based features to create the feature vectors, in conjunction with values that determine the spatial distribution of these colors (centroids) and the degree of dispersion of data regarding the average value (standard deviations). This combination is commonly known as "local color"[14]. In addition, to complement them, we extract features that determine the intensity and direction of motion in the scenes.

The similarities between videos is usually measured by distances [9][10]. In [15] convolution-based measures are used for the motion features, distances for ordinal characteristics, and histogram intersections measures for the color histogram-based features. In [16], predefined thresholds and measures calculated from an expression of dissimilarity are used. In any case, most methods represent a video by a list of keyframes. A different representation can be found in [17], in what its authors called Bounded Coordinate System (BCS). In this system, each keyframe of a video is located in an N-dimensional space, and the complete video is represented by the smaller "box" of N dimensions which contains their keyframes. In our paper we use a similar video representation system, where similarity metrics are based on aspects such as the distance between centroids of those boxes, or the transformation distance from one box to another.

This paper describes the development of the *vManager* project. The design of content-based strategies for video storage and search is the general objective. We propose a model of feature vector, which focuses on the extraction of color-based and motion-based features. *vManager* detects scene breaks, selects keyframes, uses a proper video representation model, and proposes various measures of similarity. The approach is tested in a practical application with an acceptable success ratio. The proposal presented may be reproduced in many other environments and CBVR systems.

2 Methods

The consecutive frames of a digital video tend to have similar characteristics. This aspect implies the appearance of high redundancy. Instead of extracting feature vectors for all the frames and performing the comparison of two videos based on these vectors, it is more efficient to group consecutive frames in clusters or similar scenes, representing these scenes by using a single keyframe. The comparison process among videos is limited to these subsets of frames (keyframes) instead of considering all the frames of the video.

To perform localization, indexing and the subsequent retrieval of videos, three main tasks may be described:

- Scene detection
- Feature extraction
- Measures of similarity

2.1 Scene Detection and Keyframe Selection

The frames of the same scene are very similar, but in the limit of scenes, the difference between two consecutive frames is great. The scene change can be abrupt, facilitating scene break detection, or can be gradual, because of dissolves, wipes, merge of frames, etc., complicating scene detection. The proper motion of the camera should also be considered to avoid false positives in scene detection. Eventually, motion compensation techniques can be applied.

Our automatic scene breaks detection process is inspired by [18] and consists in detecting the number of pixels that change from one frame to the next, basically comparing the pixels that disappear and appear from one frame to another. Thus, when processing the full video, a pixel change function is obtained, and the peaks of this function indicate where the scene breaks have been detected.

After such a detection, a representative keyframe of each scene must be selected. In *vManager*, the central frame of the scene has been selected as a keyframe.

2.2 Feature Extraction

Once the scenes have been detected and keyframes selected, the next step consists in extracting a feature vector for each of the keyframes. *vManager* uses local color features, in conjunction with motion-based features. Thus, the feature vector is composed of 63 features, divided into three groups. Position 0 in this vector indicates the number of frames in the scene, in case that short scenes should be considered, with little relevance by comparison with the rest of the video. On the other hand, there is a group made up of local color-based features (positions 1-44), and motion-based features (positions 45-62). Figure 1 shows the structure of our feature vector.

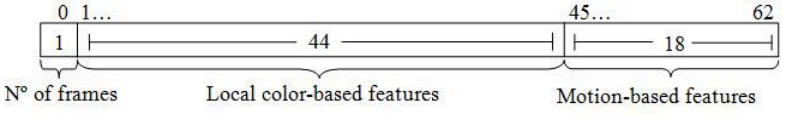


Fig. 1. Structure of the feature vector (63 features)

The local color-based features are derived from a discretization of the keyframe in the scene. This discretization consists in transforming an RGB image, in which there are 2^{24} possible colors, to an image with only 11 colors, by using the HSV color model, as Figure 2a) shows. The spatial distribution of these colors (centroids) and the degree of data dispersion regarding the average value (standard deviations) are also considered in the feature vector:

- The percentage of each of the 11 colors in the frame (positions 1 to 11).
- x coordinate of the centroid of each color in the frame (positions 12 to 22).
- y coordinate of the centroid of each color in the frame (positions 23 to 33).
- Standard deviation of each color in the frame (positions 34 to 44).

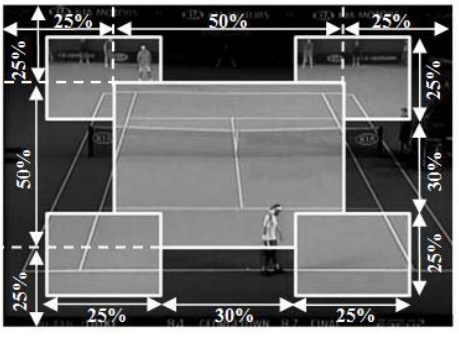
With regard to the motion characteristics of a scene, a division of the frames in five areas is performed first, as shown in Figure 2b). Thus, descriptors for the central and peripheral zones are obtained. Hence, on the one hand we have global characteristics of movement and, on the other, local features focusing on each of the five areas:

- Camera movement (position 45): based on the descriptor Camera Motion of MPEG-7, indicates whether there is camera movement in the scene or lack of it.
- Global motion intensity (position 46): based on the Motion Activity descriptor of MPEG-7, indicates the amount of movement in the scene.
- Global motion direction (position 47): also based on the Motion Activity descriptor of MPEG-7, indicates the direction of the global movement detected.
- Local motion intensity (positions 53 to 57): Indicates the amount of movement in each of the five areas.
- Local motion direction (positions 58 to 62): Indicates the direction of the local motion detected in each of the five areas.
- Order of motion intensity (positions 48 to 52): Indicates the cardinal order of the five areas according to the intensity of each movement.

To calculate the displacement (movement) in each of the areas, non-consecutive frames are compared. When comparing consecutive frames, the detected movement would be minimal. For each of the five areas of the initial frame, we find the equivalent position in the final frame. Thus, we have a displacement vector that indicates the direction and modulus (intensity of motion) for each of the considered areas.

Colors	H	S	V
White	Irrelevant	0 – 0'3	0,7 – 1
Black	Irrelevant	Irrelevant	0 – 0'1
		0 – 0'8	0'1 – 0'3
Gray	Irrelevant	0 – 0'2	0'3 – 0'7
Red	0 – 15	Irrelevant	
	330 – 360		
Orange	15 – 45		
Yellow	45 – 90		
Green	90 – 150		
Cyan	150 – 210		
Blue	210 – 255		
Purple	255 – 285		
Magenta	285 – 330		

a)



b)

Fig. 2. Feature vector. a) Local color-based features: values for the H, S and V channels for the reduced palette of 11 colors. b) Motion-based features: the five areas of the frame.

2.3 Representation and Similarity Measures

Our final representation of a video is inspired by [17]. Thus, a video is represented in an N-dimensional space, where N is the number of features of the keyframes (62 in our case). Position 0 in our feature vector indicates the number of frames of the scene. We use this feature to filter short scenes (scenes with a little number of frames and little relevance in the video). Each key-frame represents a point in this N-dimensional space. Hence, the videos are represented as the smaller N-dimensional *box* that involves their representative points (keyframes). The center of this box is determined by the centroid of its points, and the length of the box in each dimension is equal to two times the standard deviation of the points in that dimension. Through this representation of boxes, two videos can be compared to measure the similarity between them. This comparison is done by calculating two similarity distances:

Distance between centers: Euclidean distance between the centroids of the boxes to compare. The smaller the distance, the greater the similarity.

Transformation Distance: Measures the cost of transforming one box into another. It is calculated as the sum of the Euclidean distance between the minimal extreme points of the boxes plus the Euclidean distance between the maximal extreme points.

3 Practical Application and Results

In our experiments, we have worked with a set of videos of five different sports categories: basketball, handball, football, motorcycling and tennis. We considered six videos for tennis, and eight for the other four categories.

In the tests we chose a video as a query (query by example), and we retrieved videos from the database. From the retrieved videos, we checked if they belong to the same category of the query video. To quantify the numerical results we used the typical measures of *precision* and *recall*. In fact, the results shown in this paper focus on the *f-measure*, which relates the *precision* and *recall* measures for each of the similarity distances discussed.

Figure 3 shows the results of the experiments. Factors have been considered based only on local color features (44 features) and only on motion features (18 features). The color and its spatial distribution in the frame are crucial in the human visual system to identify similar scenes. Although the motion characteristics are not discriminating enough in themselves, they could help in discriminating videos, in certain circumstances. Therefore, we have also considered feature vectors based on local color and motion features (62 features), with equal weight to all features. Finally, as color characteristics are more discriminating for the human visual system than motion features, we defined a weighting function in which the characteristics of local color have a weight of 75% and the motion features of 25% in the overall feature vector, to calculate the distances among videos.

In our experience, color-based features should have three times more weight than motion-based features. To achieve this, according to the formula of the

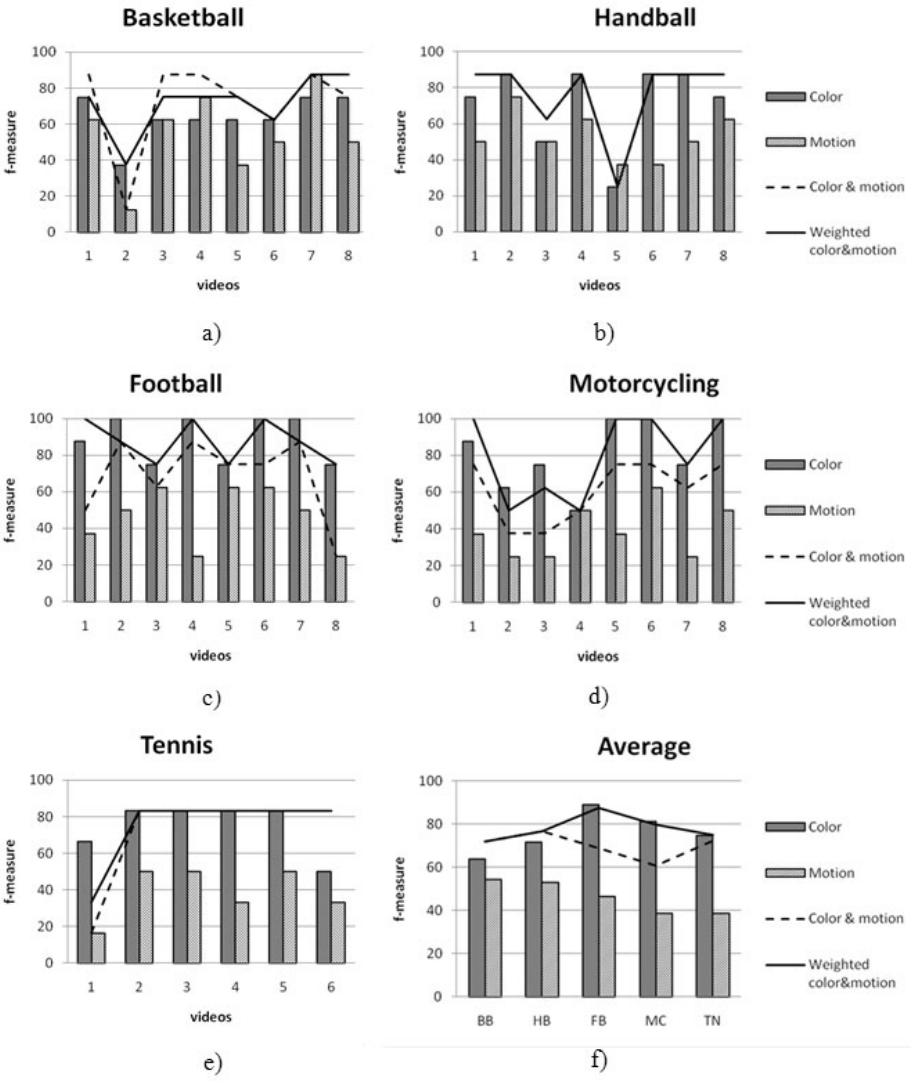


Fig. 3. Results of *f-measure* for color-based feature vectors, motion-based feature vectors, color- and motion-based feature vectors, and weighted color- and motion-based feature vectors, for the five categories (a-e). f) Average results for basketball (BB), handball (HB), football (FB), motorcycling (MC) and tennis (TN).

weighted Euclidean distance, we associate a weight w_i to each feature based on color, with three times more value than those associated with motion-based features. In addition, the sum of all weights is 1. Since the feature vector consists of 44 values based on color and 18 on the movement, the weights for color-based features are 0.2, and the weights for motion-based features are 0.0067 (i.e. $w_i = 0.2$ for $1 <= i <= 44$, and $w_i = 0.0067$ for $45 <= i <= 62$).

Figure 3 shows the results of the tests obtained by the distance between centroids. The two similarity measures were used in our tests, and the results were very similar. Figures 3a)-3e) show the results for each of the five categories. In each graph, the results of *f-measure* are shown, for color-based feature vectors, motion-based feature vectors, color- and motion-based feature vectors, and weighted color- and motion-based feature vectors. Figure 3f shows the average for each category.

As seen in Figure 3, the local color-based features perform better recovery rates than those based exclusively on movement. Therefore, the vectors of motion-based features are not discriminating. Considering vectors of local color-based and motion-based features (dashed line in Figure 3), the video retrieval is similar or improves those results obtained by using local color-based features, for basketball, handball and tennis. For the other categories, the results worsen. Pondering the motion-based features (solid line in Figure 3), on average, the results improve those obtained based only on local color features.

4 Conclusions

This paper has described the developed CBVR system. We propose a model of feature vector, which focuses on the extraction of color-based and motion-based features. These vectors are obtained after a process of scene detection, which is also included in *vManager*. Our system uses a proper video representation model, and proposes various measures of similarity. From the results obtained in the experiments we can conclude that the local color-based features provide good results in the recovery of videos in general. The vectors of motion-based features are not discriminating, but they do help as complement to those based on color, especially if they are weighted so that the color-based features are three times greater than those of motion-based features. Finally, our proposal can be implemented and replicated in many other environments and CBVR systems.

Acknowledgments. This work is funded by the Ministerio de Educación y Ciencia, and the European Union (FEDER funds) under the research projects #TIN2005-05939 and #TIN2008-003063, and the Junta de Extremadura research project #PDT08A021.

References

1. Hampapur, A., Hyun, K.H., Bolle, R., Ferman, A.M., Tekalp, A.M., Mehrotra, R.: Robust Color Histogram Descriptors for Video Segment Retrieval and Identification. IEEE Transactions on Image Processing 11(5), 497–508 (2002)
2. Lee, H.-C., Kim, S.-D.: Rate-Driven Key Frame Selection Using Temporal Variation of Visual Content. Electronics Letters 38(5), 217–218 (2002)
3. Zhai, Y., Liu, J., Cao, X.: Video understanding and content-based retrieval. School of Computer Science (2005)

4. Shiitani, S., Baba, T., Endo, S., Uehara, Y., Masumoto, D., Nagata, S.: Efficient video retrieval system using virtual 3D space. In: 6th IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 206–210 (2004)
5. Kim, J., Lim, H., Kang, D.: An implementation of the video retrieval system by video segmentation. In: 14th Asia-Pacific Conference on Communications, pp. 1–5 (2008)
6. Nguyen, D.T., Gillespie, W.: A video retrieval system based on compressed data from MPEG files. In: Conference on Convergent Technologies for Asia-Pacific Region, pp. 555–560 (2003)
7. Vakkalanka, S., Palanivel, S., Yegnanarayana, B.: NVIBRS - news video indexing, browsing and retrieval system. Intelligent Sensing and Information Processing, 181–186 (2005)
8. Camara-Chavez, G., Precioso, F., Cord, M., Phillip-Foliguet, S., de Araujo, A.: An interactive video content-based retrieval system. In: 15th International Conference on Systems, Signals and Image Processing, pp. 133–136 (2008)
9. Zhou, X., Zhou, X., Shen, H.T.: Efficient similiarity search by summarization in large video database. In: School of Information Technology and Electrical Engineering (2007)
10. Zagorac, S., Llorente, A., Little, S., Liu, H., Rueger, S.: Automated Content Based Video Retrieval, TREC Video Retrieval Evaluation Notebook Papers (2009)
11. Hua, X., Yin, P., Wang, H., Chen, J., Mingjing, L., Li, M., Zhang, H.: MSR-Asia TREC-11 video track. In: Proceedings of the Text Retrieval Conference (2002)
12. Zampoglou, M., Papadimitriou, T., Diamantaras, K.I.: Integrating motion and color for content based video classification. In: Proceedings of IEEE Intl. Conf. on Image Processing (2008)
13. Kim, C., Vasudev, B.: Spatiotemporal sequence matching for efficient video copy detection. IEEE Transactions on Circuits and Systems for Video Technology 15(1), 127–132 (2005)
14. Cinque, L., Levialdi, S., Olsen, K.A., Pellinaco, A.: Color-Based Image Retrieval Using Spatial-Chromatic Histograms. Image and Vision Computing 19, 979–986 (2001)
15. Dimitrovski, I., Loskovska, S., Kalasevski, G., Chorbev, I.: Video Contet-Based Retrieval System. In: EUROCON, The International Conference on "Computer as a Tool", pp. 978–983 (2007)
16. Huang, C.-L., Liao, B.-Y.: A Robust Scene-Change Detection Method for Video Segmentation. IEEE Transactions on Circuits and Systems for Video Technology 11(12), 1281–1288 (2001)
17. Huang, Z., Shen, H.T., Shao, J., Zhou, X., Cui, B.: Bounded Coordinate System Indexing for Real-Time Video Clip Search. ACM Transactions on Information Systems 27(3), 17.1–17.33 (2009)
18. Zabih, R., Miller, J., Mai, K.: A Feature-Based Algorithm for Detecting and Classifying Scene Breaks. In: Third ACM International Conference on Multimedia, pp. 189–200 (1995)

On the Use of Dot Scoring for Speaker Diarization*

Mireia Diez**, Mikel Penagarikano, Amparo Varona, Luis Javier Rodriguez-Fuentes, and German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
mireia_diez@ehu.es

Abstract. In this paper, an alternative dot scoring based agglomerative hierarchical clustering approach for speaker diarization is presented. Dot-scoring is a simple and fast technique used in speaker verification that makes use of a linearized procedure to score test segments against target models. In our speaker diarization approach speech segments are represented by MAP-adapted GMM zero and first order statistics, dot scoring is applied to compute a similarity measure between segments (or clusters) and finally an agglomerative clustering algorithm is applied until no pair of clusters exceeds a similarity threshold. This diarization system was developed for the Albayzin 2010 Speaker Diarization Evaluation on broadcast news. Results show that the lowest error rate that the clustering algorithm could attain for the evaluation set was around 20% and that over-segmentation was the main source of degradation, due to the lack of robustness in the estimation of statistics for short segments.

Index Terms: Speaker Diarization, Dot Scoring, Sufficient Statistics.

1 Introduction

Speaker Diarization consists of determining who spoke when in an input audio stream. It involves two main steps: determining the boundaries between speaker turns and clustering segments according to the speaker identity [1]. In recent years, speaker diarization has gained importance as a mean of indexing different types of data such as meetings, broadcast news or telephone conversations.

Most speaker diarization systems apply agglomerative hierarchical clustering with a BIC-based stopping criterion [1]. In this paper, an alternative dot scoring-based agglomerative hierarchical clustering approach is presented. Dot scoring is commonly used as a fast scoring technique in speaker verification. Applying

* This work has been supported by the University of the Basque Country under Grant GIU10/18 and the Government of the Basque Country, under program SAIOTEK (project S-PE10UN87), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

** Supported by a research fellowship from the Department of Education, Universities and Research of the Basque Country Government.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 612–619, 2011.

dot scoring to diarization has the advantage of a low computational cost for re-training speaker-cluster models.

The dot-scoring speaker diarization system developed for the Albayzin 2010 Speaker Diarization Evaluation (SDE) is based on three subsystems: the audio classifier developed for the Albayzin 2010 Audio Segmentation Evaluation [2], the acoustic change detector module developed for the system submitted to the Albayzin 2006 Speaker Tracking Evaluation [3], and the speaker verification system developed for the NIST 2010 Speaker Recognition Evaluation [4].

The paper is organized as follows: in section 2 all the stages of the proposed diarization system are described: speech/non speech segmentation, acoustic change detection, dot scoring and the clustering algorithm. In section 3 we present the experimental setup used to develop and evaluate the system. Results, as well as the processing time required, are presented in section 4. Finally, conclusions are outlined in section 5.

2 Speaker Diarization System

Speech/non-speech detection, acoustic change detection and clustering are performed separately in this approach.

The speech/non-speech detector developed for this task is based on a 5-class ergodic Continuous Hidden Markov Model including 3 speech sub-classes and 2 non-speech sub-classes. More details can be found in [2]. A simple approach, which uses a XBIC-based measure to detect any change of speaker, background or channel conditions, was applied as defined in [3]. Though it oversegments the audio stream, the set of change points includes almost all the speaker changes. The optimal configuration of the three subsystems was heuristically determined on development data (see section 3.3 for details).

2.1 Dot Scoring

Dot scoring agglomerative hierarchical clustering was performed as follows:

Universal Background Model. A gender independent GMM (Universal Background Model, UBM) was trained. The Sautrela toolkit [5] was used to estimate GMM parameters, applying binary mixture splitting, orphan mixture discarding and variance flooring.

Sufficient Statistics. Let $\lambda \equiv \{\omega_k, \mu_k, \Sigma_k | k = 1..K\}$ be a GMM composed by K Gaussians of dimension F with diagonal covariance matrices Σ_k. Let f_t be the feature vector at time t. Let $\gamma_k(t)$ be the posterior probability of Gaussian k at time t. We define:

$$n_k = \sum_t \gamma_k(t) \tag{1}$$

$$x_k = \sum_t \gamma_k(t) \Sigma_k^{-\frac{1}{2}} (f_t - \mu_k) \tag{2}$$

The sets of parameters vectors $\nu = [\nu_{ij}]$, where $\nu_{ij} = n_i$, $i \in [1..K]$, $j \in [1..F]$, and $x = [x_1, \ldots, x_K]$ (each x_i being a F-dimensional vector) are known as the zero and first order sufficient statistics, respectively. Given a dataset c, the one-iteration relevance-MAP adapted and normalized mean vectors $m = \Sigma^{-\frac{1}{2}}(\mu_c - \mu_{ubm})$ can be computed according to the following expression[1] [6,4]:

$$m = (\tau \mathbf{I} + diag(\nu))^{-1} \cdot x \tag{3}$$

Dot Scoring Similarity Measure. Dot-scoring is a simple and fast technique used in speaker verification that makes use of a linearized procedure to score test segments against target models [6]. Given a feature stream f (the target signal) and a speaker model λ_s, the first-order Taylor-series approximation to the GMM log-likelihood is:

$$\log P(f|\lambda_s) \approx \log P(f|\lambda_{ubm}) + m_s^t \cdot \nabla P(f|\lambda_{ubm}) \tag{4}$$

where m_s denotes the vector of normalized means corresponding to speaker s, ∇ denotes the gradient vector w.r.t the standard-deviation-normalized means of the UBM, and $\nabla P(f|\lambda_{ubm}) = x_f$ is the vector of first order statistics corresponding to the target signal f. The log-likelihood ratio between the target model and the UBM used for scoring can be approximated as follows:

$$score(f, s) = \log \frac{P(f|\lambda_s)}{P(f|\lambda_{ubm})} \approx m_s^t \cdot x_f \tag{5}$$

For the diarization task, the similarity $sim(a, b)$ between two segments a and b was defined as:

$$sim(a, b) = min\{score(f_a, b), score(f_b, a)\} = min\left\{m_b^t \cdot x_a, m_a^t \cdot x_b\right\} \tag{6}$$

Score Normalization. TZ normalization was applied to dot-scores. Two independent sets of development data were used for the estimation of T-norm (normalization w.r.t. the test utterance) and Z-norm (normalization w.r.t. the speaker cluster) parameters. Taking into account score normalization, the similarity measure was redefined as:

$$sim(a, b) = min\{score_{TZ}(f_a, b), score_{TZ}(f_b, a)\} \tag{7}$$

2.2 The Clustering Algorithm

The similarity measure defined above was used to perform agglomerative hierarchical clustering. Given two segments (or two clusters of segments), if they are clustered together, computation of sufficient statistics for the joint cluster is straightforward:

$$\begin{aligned} x_{a+b} &= x_a + x_b \\ n_{a+b} &= n_a + n_b \end{aligned} \tag{8}$$

[1] $diag(\nu)$ stands for a square matrix with the elements of ν in the diagonal.

This leads to a very simple clustering algorithm:

1. **Find** $s_{max} = \max_{\forall(a,b)} \{sim(a,b)\}$
   $(a*, b*) = \operatorname{argmax}_{\forall(a,b)} \{sim(a,b)\}$
2. **If** $s_{max} < \Theta$ **then** STOP
3. **Set** $x_{a*} = x_{a*} + x_{b*}$
   $n_{a*} = n_{a*} + n_{b*}$
4. **Remove** cluster $b*$
5. **Jump to** 1

3 Experimental Setup

3.1 Databases

We decided to keep independence between training and development data, therefore the Albayzin 2010 SDE database was used for development and the KALAKA database [7] for training the GMMs.

The Albayzin 2010 SDE consists of 24 sessions of TV broadcast news in Catalan, most sessions being 4 hours long (some of them being shorter). The database, recorded from the 3/24 TV channel, includes around 87 hours of audio, split into 2 sets: train/development (16 sessions, 2/3 of the total amount of data) and test (8 sessions, the remaining 1/3). Even though 3/24 TV mostly contains speech in Catalan, around 1/6 of the speech segments are spoken in Spanish. The database contains male, female and overlapped speech and the number of speakers per recording varies from 30 to 250. The distribution of background conditions within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3% [8].

KALAKA materials were also extracted from (wide-band) TV shows. The database, which was designed to build language recognition systems, contains speech in four target languages: Basque, Catalan, Galician and Spanish, all of them official languages in Spain. KALAKA contains more than 12 hours of speech per target language.

3.2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The MFCC set, comprising 13 coefficients, including the zero (energy) coefficient, was computed in frames of 32 ms at intervals of 10 ms for the two first modules (speech/non-speech detection and acoustic change detection). In the clustering approach, the MFCC set was computed in frames of 20 ms at intervals of 10 ms and augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector. Also, an energy based voice activity detector (VAD) was applied to remove those fragments

(short silences) with an energy level 30 dB (or more) under the maximum. All the speech processing computations were done by means of the Sautrela toolkit [5].

3.3 Parameter Optimization

Speech/Non-Speech Detector. A 5 state ergodic Continuous Hidden Markov Model was estimated using the Sautrela toolkit [5], under the Layered Markov Models framework. Preliminary experiments on a subset of the development data revealed that best audio segmentation performance was achieved when the number of mixtures was 512. The emission distributions were independently estimated for each state, applying the Baum-Welch algorithm on the corresponding sets of segments extracted from the reference segmentations of 12 development sessions. The number of mixtures per state and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the remaining 4 development sessions. Considering a 2-class speech/non-speech classification setup, the false alarm and the miss error rates were around 1% for the speech class (including the three sub-classes mentioned in 3.1). Note that, since we are mistaking around 2% of the speech frames, our speaker diarization error will be, at best, of that order.

Gaussian Mixture Models. Although the evaluation was limited to Catalan TV speech, in order to increase the speaker variability, TV broadcast speech in Spanish, Catalan, Galician and Basque, taken from the Kalaka database [7], was used to train gender independent GMMs (Universal Background Model, UBM) consisting of 256, 512 and 1024 mixture components. Again, the Sautrela toolkit was used to estimate GMM parameters, applying binary mixture splitting, orphan mixture discarding and variance flooring.

Threshold Selection. Threshold optimization was performed on the development set. Figure 1 shows the performance measured on development sessions 3-16 for each threshold value using a 256-mixture GMM system. Based on the average system performance, similarity thresholds were set to the values shown in Table 1.

Table 1. Threshold value selection for each GMM system based on average system performance

	#mixtures		
	256	**512**	**1024**
Threshold	3.80	3.74	3.98

3.4 Performance Criteria

The diarization error rate (DER) defined by NIST [9] was the primary metric used in the evaluation, applying a scoring "forgiveness collar" of 250 ms around each reference segment boundary [8].

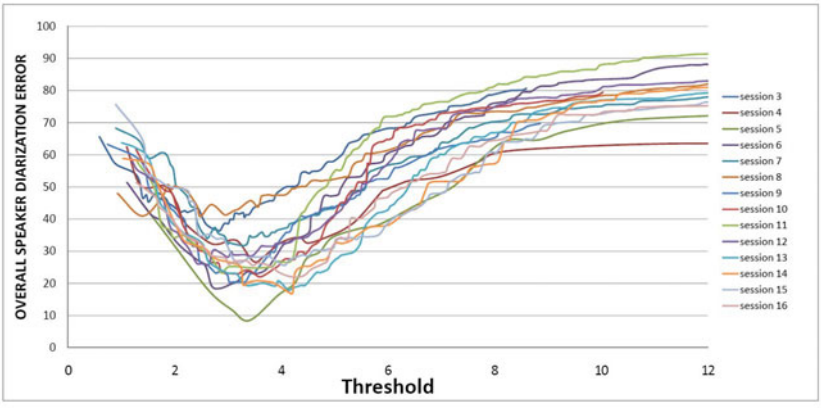


Fig. 1. Overall Speaker Diarization Error as a function of the similarity threshold applied as stopping criterion in the clustering algorithm, for sessions 3-16 of the development set using a 256-mixture GMM system

4 Results

Experiments carried out showed almost no difference in performance among the GMM systems using 256, 512 and 1024 mixtures. Therefore, the 256-mixture GMM system was selected for further analyses, due to the lower cost of sufficient statistics and similarity matrix computations. Table 2 shows the performance of the clustering algorithm described above on the evaluation set, using four different segmentations:

- Seg1: Reference Speaker Segmentation.
- Seg2: Reference Speaker Segmentation + GTTS Acoustic Change Detection.
- Seg3: Reference Speech/Non-Speech Segmentation + GTTS Acoustic Change Detection.
- Seg4: GTTS Speech/Non-Speech Detection + GTTS Acoustic Change Detection.

Table 2. Overall Speaker Diarization Error obtained by applying the clustering algorithm on four different segmentations of the evaluation set (see text for details)

DER %	**Seg1**	**Seg2**	**Seg3**	**Seg4**
256-m GMM	20.48	26.14	29.61	33.16

The Overall Speaker Diarization Error obtained with the Reference Speaker Segmentation (Seg1, 20.48%) would be the best performance that our clustering system could reach for the evaluation set. The difference between this result and the result obtained with the fully automated system (Seg4, 33.16%) may be explained as follows:

- Difference between Seg3 and Seg4: 3.55%. Seg3 starts from a perfect Speech/Non-Speech classification, whereas Seg4 applies the GTTS Speech/Non-Speech detection system. So, the difference can be explained by the Speech/Non-Speech classification error.
- Difference between Seg1 and Seg2: 5.66%. Since both systems take the reference speaker segmentation as a starting point, the difference in performance can only be due to over-segmentation introduced by the GTTS acoustic change detector. Applying the acoustic change detector on the optimal speaker segmentation does not remove speaker boundaries but produces many short segments whose statistics strongly depend on local variabilities. This explains why the performance of the clustering algorithm, which is based on those statistics, degrades for short segments.
- Difference between Seg2 and Seg3: 3.47%. Seg2 includes all the speaker boundaries (plus a number of acoustic changes inside speaker turns), whereas Seg3 may be missing some of them. This explains the difference.

4.1 Processing Time

Table 3 shows the CPU time (expressed as real-time factor, ×RT) employed in six separate operations: (1) feature extraction for segmentation; (2) speech/non-speech segmentation; (3) acoustic change detection; (4) feature extraction for clustering; (5) computation of sufficient statistics; and (6) hierarchical clustering of speech segments, for both the reference speaker segmentation and the automatic segmentation. Note that the CPU time employed in clustering is almost four times higher for the automatic segmentation than for the reference segmentation, because of the different number of speech segments: 7.24 and 3.62 segments/minute, respectively. The total CPU time of the speaker diarization system is 0.2932×RT.

Computations were made in two servers. The first one, devoted to speech/non-speech segmentation and acoustic change detection, was a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. The second one, devoted to clustering, was a Dell PowerEdge R610, equipped with 2 Xeon 5550 (each featuring 4 cores) at 2.66GHz and 32GB of RAM.

Table 3. CPU time (real-time factor, ×RT) employed by the different modules of the speaker diarization system

	Ref. segm.	**GTTS segm.**
Features (segmentation)	-	0.0033
Speech/non-speech segmentation	-	0.0375
Acoustic change detection	-	0.1058
Features (clustering)	0.0026	
Statistics	0.0050	
Clustering	0.038	0.139

5 Conclusions

In this paper a new speaker diarization approach, which applies agglomerative hierarchical clustering based on dot scoring, has been described.The system consists on a chain of four uncoupled modules: speech/non-speech segmentation, acoustic change detection, computation of sufficient statistics and hierarchical clustering of speech segments. Despite its simplicity, the proposed system attained competitive results in the Albayzin 2010 Speaker Diarization Evaluation.

Experiments carried out on different segmentations showed: (1) that the best performance that the clustering algorithm could attain for the evaluation set was around 20%; and (2) that over-segmentation introduced by the acoustic change detector was the main source of degradation, because of the lack of robustness in the estimation of statistics for short segments. Future work will involve trying to improve the robustness of the clustering algorithm to short segments, or alternatively, to avoid over-segmentation while keeping the detection rate of speaker boundaries. Besides, alternative feature parameterizations will be studied.

References

1. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing 14(5), 1979–1986 (2006)
2. Rodriguez-Fuentes, L.J., Penagarikano, M., Varona, A., Diez, M., Bordel, G.: GTTS Systems for the Albayzin 2010 Audio Segmentation Evaluation. In: FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop, Vigo, Spain (November 2010)
3. Rodríguez, L.J., Peñagarikano, M., Bordel, G.: A simple but effective approach to speaker tracking in broadcast news. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4478, pp. 48–55. Springer, Heidelberg (2007)
4. Penagarikano, M., Varona, A., Diez, M., Rodriguez-Fuentes, L.J., Bordel, G.: University of the Basque Country System for NIST 2010 Speaker Recognition Evaluation. In: Proceedings of the II Iberian SLTech Workshop, Vigo, Spain (2010)
5. Penagarikano, M., Bordel, G.: Sautrela: A Highly Modular Open Source Speech Recognition Framework. In: Proceedings of the ASRU Workshop, San Juan, Puerto Rico, pp. 386–391 (December 2005)
6. Strasheim, A., Brümmer, N.: SUNSDV system description: NIST SRE 2008. In: NIST Speaker Recognition Evaluation Workshop Booklet (2008)
7. Rodriguez-Fuentes, L.J., Penagarikano, M., Bordel, G., Varona, A., Diez, M.: KALAKA: A TV Broadcast Speech Database for the Evaluation of Language Recognition Systems. In: 7th International Conference on Language Resources and Evaluation, Valleta, Malta, May 17-23 (2010)
8. Zelenak, M., Schulz, H., Hernando, J.: Albayzin 2010 Evaluation Campaign: Speaker Diarization. In: FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop, Vigo, Spain (November 2010)
9. The 2009 NIST Rich Transcription Evaluation, `http://www.itl.nist.gov/iad/mig/tests/rt/`

A Bag-of-Paths Based Serialized Subgraph Matching for Symbol Spotting in Line Drawings

Anjan Dutta[1], Josep Lladós[1], and Umapada Pal[2]

[1] Computer Vision Centre, Edifici O, Campus UAB, 08193 Bellatera, Barcelona, Spain
[2] CVPR Unit, Indian Statistical Institute, 203, B.T.Road, Kolkata-700108, India
{adutta,josep}@cvc.uab.es,
umapada@isical.ac.in

Abstract. In this paper we propose an error tolerant subgraph matching algorithm based on bag-of-paths for solving the problem of symbol spotting in line drawings. Bag-of-paths is a factorized representation of graphs where the factorization is done by considering all the acyclic paths between each pair of connected nodes. Similar paths within the whole collection of documents are clustered and organized in a lookup table for efficient indexing. The lookup table contains the index key of each cluster and the corresponding list of locations as a single entry. The mean path of each of the clusters serves as the index key for each table entry. The spotting method is then formulated by a spatial voting scheme to the list of locations of the paths that are decided in terms of search of similar paths that compose the query symbol. Efficient indexing of common substructures helps to reduce the computational burden of usual graph based methods. The proposed method can also be seen as a way to serialize graphs which allows to reduce the complexity of the subgraph isomorphism. We have encoded the paths in terms of both attributed strings and turning functions, and presented a comparative results between them within the symbol spotting framework. Experimentations for matching different shape silhouettes are also reported and the method has been proved to work in noisy environment also.

Keywords: Symbol spotting, Serialization of graphs, Graph matching, Bag-of-paths, Attributed strings, Turning function, Graphical indexing, Mean paths.

1 Introduction

Information spotting is a major branch of indexing and retrieval methods. In document analysis, the research community is mainly focused in word spotting for textual documents and symbol spotting for graphical documents. Nowadays, symbol spotting has experienced a growing interest among the Graphics Recognition community, a subfield of Document Image Analysis and Recognition (DIAR). It can be defined as locating a given query symbol into a graphical document image, which is commonly referred as *focused retrieval*. The main

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 620–627, 2011.

application of symbol spotting is indexing and retrieval into large database of graphical documents, e.g. finding a mechanical part into a database of engineering drawings. The desired output for a particular query should be a ranked list of retrieved symbols in which the true positives should appear at the beginning.

Although symbol spotting is an emerging topic, several efforts have been made among the graphics recognition community for spotting symbols in graphical documents [10]. The algorithms proposed by Messmer [5], Müller and Rigoll [6] are among the first few approaches of symbol spotting. Graph based methods [4] are also popular, but they often suffer from computational complexity. Among the others, Rusiñol and Lladós have used a technique of splitting the symbols into several primitives and used graph matching [9] and off-the-shelf shape descriptors [8] to represent them. Recently Nayef and Breuel [7] proposed a branch and bound algorithm for spotting symbols in documents. The preprocessing is done by simple morphological operation and then thinning.

In this paper we propose a symbol spotting technique in line drawings based on attributed subgraph matching. Graphs are very suitable to represent graphical entities, in particular, line drawings. Also graph representations, when storing geometric information can efficiently handle various affine transformations viz. rotation, translation, scaling. Hence symbol spotting can be solved by inexact subgraph isomorphism techniques, which can take the advantage of the soundness of graph theory. Moreover, graphs are widely adapted by the research community as a robust tool since a long back, as a result lots of efficient methods and algorithms are available to handle the graph based methods. On the other hand, (sub)graph isomorphism is considered as a computationally hard problem, for that reason handling a large database of graphical documents using graphs is difficult as it increases the computational complexity. To avoid the computational burden, in this paper we propose a method based on factorization of graphs that represent the documents and find the common factorized substructures within the entire collection. The finding of the common substructures for the whole document collection and then performing the graph matching helps to reduce the computational complexity at the ultimate level. The proposed method can also be seen as a way to serialize graphs i.e. to represent them by one-dimensional structures. This further allows to reduce the complexity of subgraph isomorphism process to spot symbols on documents.

The rest of the paper is organized into three sections. In Section 2 we present the detailed methodology of our algorithm with a brief description of the spotting architecture. Section 3 presents the experimental results of the proposed methods. After that in Section 4, we conclude the paper and future research lines to extend the present work are defined.

2 Proposed Methodology

In this work we propose an efficient error tolerant subgraph matching algorithm based on the idea of bag-of-paths. Bag-of-paths for a particular graph is informally defined as the set of all acyclic paths between each pair of connected

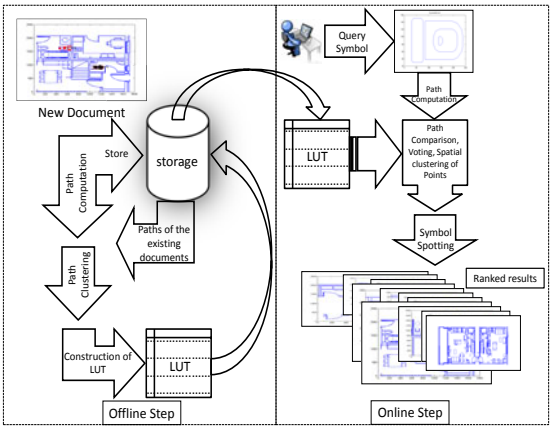


Fig. 1. Outline of the spotting architecture

nodes of that graph. The construction of bag-of-paths is based on the idea of graph factorization. For our case we factorize the graphs constructed from the documents within the database, as well as the query symbol. Similar paths of the whole document database are clustered into a lookup table so that the mean path of each of the clusters can serve as the key index of that cluster. Here the basic idea is to find the best matching cluster for each of the paths in the model bag-of-paths and apply a spatial voting scheme to the terminal points of the paths to detect the symbol in the whole database simultaneously.

Our entire framework can be divided into two different parts viz. offline and online (see Fig. 1). The offline part includes the computation of all the acyclic paths, clustering of those paths within the whole collection, construction of lookup table and computation of the mean path which acts as an index key for each of the table entries. Each time a new document is being included in the database the entire offline procedure is repeated to create the updated lookup table. For each of the documents in the database all the computed paths are stored to reduce the further path computation time. On the other hand, the online part includes the querying of the graphic symbol by an end user, computation of all the acyclic paths for that symbol, a voting scheme which is based on the similarity measure of the paths composed of the query symbol and a spatial clustering technique to detect the query symbol on the image database.

2.1 Bag-of-Paths

Our bag-of-paths approach can be motivated by an analogy to learning methods using the *bag-of-words* representation for text categorization. Bag-of-paths for a particular graph G can be defined as a set P of all acyclic paths between any two connected nodes of that graph. For instance in Fig 2(a) we have shown one sample symbol, in which all the acyclic paths from the point A to the point C are $A-B-C$, $A-D-C$ and $A-C$, which are shown in Fig 2(b) and all the acyclic

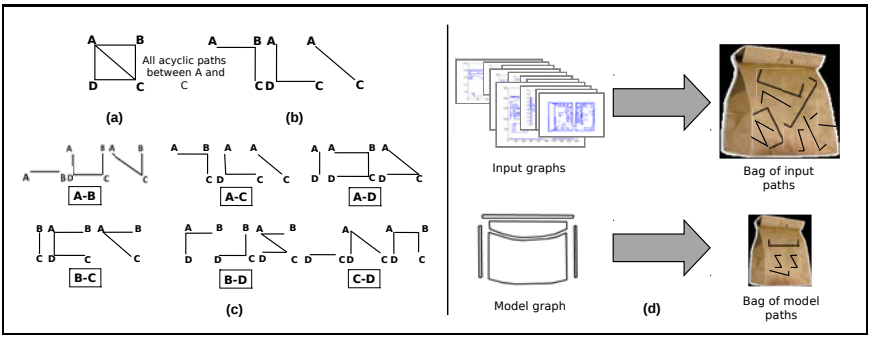


Fig. 2. Bag of paths representation (a) An example symbol. (b) All the acyclic paths between the points A and C of the symbol. (c) All the acyclic paths of the symbol. (d) Bag-of-paths representation for the document database and query symbol.

paths for that symbol are shown in Fig. 2(c). So for a graph corresponding to a vectorized document or model symbol, which can be thought of as an union of several symbols, we can find a set of paths which is referred as bag-of-paths.

In this work the paths are encoded in two different ways viz. (1) attributed string [11] and (2) turning function [1]. String edit distance [11] and L_p distance [1] based metric are respectively used to measure the similarity between different encoded paths. Finally, the performance of these two metrics are compared within a symbol spotting framework.

2.2 Construction of the LUT

The lookup table (LUT) is constructed by clustering the similar paths within the entire collection of documents. The clustering is intended to separate the structurally dissimilar paths into different clusters and accumulate the similar paths into same clusters. The LUT consists of two different items: a representative path of each cluster which acts as the indexing key and the list of locations where the paths belong in the document database. For our case the representative path is the mean path which is efficiently computed from the mean turning functions of the respective paths as detailed in [3]. The clustering of the paths is done in two steps: (1) all the paths in a single document are clustered, represented by the mean path and (2) all the mean paths in the document database are clustered and the final lookup table is constructed. In both the steps we have done the hierarchical clustering by computing the proximity matrix of all the candidate paths, where we have used string edit distance [3] as the distance measure.

2.3 Voting Scheme

A voting space is defined over the images of the database dividing them into grids of several sizes (10×10, 20×20 and 30×30). Multiresolution grids are used to detect the symbols accurately within the image and the sizes of them is experimentally determined. For a particular model path, we select the best

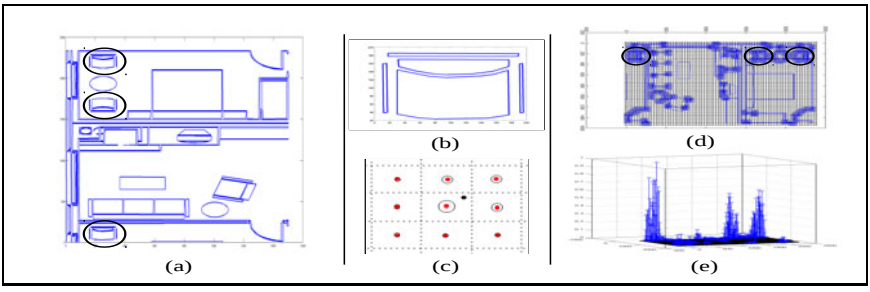


Fig. 3. Voting scheme for a given query and a particular document (a) The vectorized floorplan, here the circles denote the occurrences of the query symbol. (b) The given query symbol. (c) Nine neighboring grids for a particular terminal points. (d) Accumulated votes for a document, here the circles denote the higher frequencies. (e) Accumulated votes showing with a 3D plot.

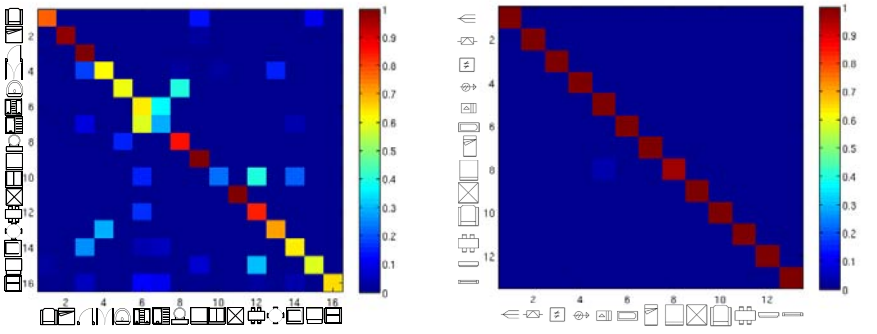
matching cluster or entry in the lookup table and accumulate the votes to the nine nearby grids (see Fig. 3(c)) of each of the two terminal vertices of each of the paths of the cluster. Vote to a particular grid is inversely proportional to the path distance metric and is weighted by the Euclidean distance to the centers of the respective grids from the terminal of the selected path. Fig. 3(e) shows the accumulation of votes by a 3D plot which clearly discriminates the occurrences of the query symbol on the document with the higher peaks. The grids constituting the higher peaks are filtered by the k-means algorithm applied in the voting space with k=2. Finally the occurrences of the query symbol on the documents are detected by another hierarchical clustering algorithm which clusters the spatial points contributed from all the grids considered.

3 Experimental Results

In order to evaluate the proposed spotting methodology, we present two different experiments. The first one only focuses on the bag-of-paths based shape matching algorithm as a distance measure between different shapes. The second experiment is designed to test the symbol spotting method in a document image database of real architectural drawings. This experiment also reports the comparative study between the two path comparison metrics viz. (i) attributed string and (ii) turning function.

3.1 Shape Matching Experiments

This experiment is done to test the efficiency of the bag-of-paths based shape matching algorithm as a shape descriptor. The algorithm is used to measure the distance between two shapes represented by bag-of-paths. It is expected that for a match the algorithm will give lower distance than a mismatch. We have used two different isolated symbol datasets for that purpose and they are (i) SESYD Queries (floorplans) [2] and (ii) GREC-POLY [8]. The results of the experiment



(a) SESYD Queries (floorplans) (b) GREC-POLY

Fig. 4. Confusion matrices shown for the two datasets

are represented in the confusion matrices (see Fig. 4). From the confusion matrices, we can conclude that the method has succeeded in most of the model classes, but has confused when the symbols contain significant structural similarity. This is due to the generation of similar factorized substructure (or paths).

3.2 Symbol Spotting Experiments

Finally, we have tested our method with a collection of ten floorplans and twelve different symbols as the queries. This dataset is a subset of FPLAN-POLY benchmark [8] which is available in the vectorized form and the vectorization is done by the Qgar software[1]. The floorplans in the database consists of approximately 90,000 paths and after lookup table construction these paths result in 7,135 entries. The amount of string comparison metric computation thus reduced by 12.6 times than the sequential access of the paths in the whole collection. The query symbols for the experiment are shown in Table 1.

Table 1. Query Symbols used for our experiments

Symbol-01	Symbol-02	Symbol-03	Symbol-04	Symbol-05	Symbol-06
					
Symbol-07	**Symbol-08**	**Symbol-09**	**Symbol-10**	**Symbol-11**	**Symbol-12**
					

The spotting experiments are performed by encoding the paths in terms of (i) attributed string and (ii) turning function. In Table 2 we present a detailed set of measures to evaluate the performance of the algorithm for the two metrics. We

[1] `http://www.qgar.org`

Table 2. Values of different measures of our spotting experiments

Symbols	Attributed string					Turning function				
	Precision	Recall	F-index	AveP	Time (secs./doc)	Precision	Recall	F-index	AveP	Time (secs./doc)
Symbol-01	57.14	100.00	72.72	70.95	26.20	50.00	100.00	66.67	43.33	2.99
Symbol-02	50.00	100.00	66.67	76.03	22.04	46.67	100.00	63.64	40.31	3.19
Symbol-03	57.14	100.00	72.72	47.62	31.21	66.67	100.00	80.00	77.08	4.45
Symbol-04	72.72	100.00	84.21	89.94	36.09	72.73	100.00	84.21	84.09	4.99
Symbol-05	57.14	100.00	72.72	74.70	15.36	10.81	100.00	19.51	45.97	2.24
Symbol-06	57.14	100.00	72.72	66.79	60.18	50.00	100.00	66.67	75.00	4.76
Symbol-07	80.00	100.00	88.89	95.00	107.93	100.00	100.00	100.00	100.00	7.08
Symbol-08	100.00	100.00	100.00	100.00	10.41	100.00	100.00	100.00	100.00	2.22
Symbol-09	23.53	100.00	38.10	66.30	2.97	30.77	100.00	47.06	60.19	1.17
Symbol-10	80.00	100.00	88.89	80.42	4.85	50.00	100.00	66.67	50.00	1.50
Symbol-11	100.00	100.00	100.00	100.00	2.46	100.00	100.00	100.00	100.00	0.90
Symbol-12	50.00	100.00	66.67	75.00	2.08	50.00	100.00	66.67	72.92	0.80
Mean	65.40	100.00	77.03	78.56	26.81	60.64	100.00	71.75	70.74	3.02

can see the recall values for all the symbols have reached 100% and this illustrates that the algorithm is able to retrieve all the occurrences of all the symbols in the whole database. However there is a good number of false positives appear and this affects the precision. But the important thing is that we obtained good average precision values for almost all the symbols and for both the metrics. This is crucial for any retrieval method because this ensures the occurrences of the true positives at the beginning of the ranked list. The performance of both the path encoding techniques are competitive to each other since attributed string achieves higher average precision than the turning function but it is less efficient than the other in terms of computation time.

4 Conclusions and Future Work

In this paper we have proposed an error tolerant subgraph matching algorithm based on the idea of bag-of-paths. Bag-of-paths for a particular collection of graphs is the set of all acyclic paths between each pair of connected nodes, which gives a factorized representation of graphs. Finding the common factorized substructures within the whole collection and then applying the serialized subgraph isomorphism reduces the computational complexity of the usual graph based methods.

Although the performance of our method is quite high both for symbol matching and symbol spotting, the method has an important limitation to deal with real-world database. This is due to the clustering technique we have used for clustering the paths. Clustering of structural information is a separate research issue and it needs further investigation. Our future research will also focus on other graph serialization methods which will reduce the computational complexity of usual graph based methods.

Acknowledgement

This work has been partially supported by the Spanish projects TIN2009-14633-C03-03, TIN2008-04998 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

1. Arkin, E.M., Paul Chew, L., Huttenlocher, D.P., Kedem, K., Mitchell, J.S.B.: An efficiently computable metric for comparing polygonal shapes. IEEE Trans. Pattern Anal. Mach. Intell. 13(3), 209–216 (1991)
2. Delalandre, M., Pridmore, T., Valveny, E., Locteau, H., Trupin, E.: Building synthetic graphical documents for performance evaluation, pp. 288–298. Springer, Heidelberg (2008)
3. Dutta, A.: Symbol spotting in graphical documents by serialized subgraph matching, Master's thesis, Computer Vision Centre, Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellatera, Barcelona, Spain (September 2010)
4. Lladós, J., Martí, E., Villanueva, J.J.: Symbol recognition by error-tolerant subgraph matching between region adjacency graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 1137–1143 (2001)
5. Messmer, B.T., Bunke, H.: A new algorithm for error-tolerant subgraph isomorphism detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 493–504 (1998)
6. Müller, S., Rigoll, G.: Engineering drawing database retrieval using statistical pattern spotting techniques. In: Chhabra, A.K., Dori, D. (eds.) GREC 1999. LNCS, vol. 1941, pp. 246–255. Springer, Heidelberg (2000)
7. Nayef, N., Breuel, T.M.: A branch and bound algorithm for graphical symbol recognition in document images. In: Proceedings of Ninth IAPR International Workshop on Document Analysis System (DAS 2010), pp. 543–546 (2010)
8. Rusiñol, M., Borràs, A., Lladós, J.: Relational indexing of vectorial primitives for symbol spotting in line-drawing images. Pattern Recognition Letters 31(3), 188–201 (2010)
9. Rusiñol, M., Lladós, J., Sánchez, G.: Symbol spotting in vectorized technical drawings through a lookup table of region strings. Pattern Analysis and Applications 13, 1–11 (2009)
10. Tombre, K., Lamiroy, B.: Pattern recognition methods for querying and browsing technical documentation. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 504–518. Springer, Heidelberg (2008)
11. Tsai, W.-H., Yu, S.-S.: Attributed string matching with merging for shape recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 7, 453–462 (1985)

Handwritten Word Spotting in Old Manuscript Images Using a Pseudo-structural Descriptor Organized in a Hash Structure

David Fernández, Josep Lladós, and Alicia Fornés

Computer Vision Center – Dept. Ciències de la Computació
Universitat Autònoma de Barcelona
{david.fernandez,josep.llados,alicia.fornes}@cvc.uab.es
http://www.cvc.uab.es/

Abstract. There are lots of historical handwritten documents with information that can be used for several studies and projects. The Document Image Analysis and Recognition community is interested in preserving these documents and extracting all the valuable information from them. Handwritten word-spotting is the pattern classification task which consists in detecting handwriting word images. In this work, we have used a query-by-example formalism: we have matched an input image with one or multiple images from handwritten documents to determine the distance that might indicate a correspondence. We have developed an approach based in characteristic Loci Features stored in a hash structure. Document images of the marriage licences of the Cathedral of Barcelona are used as the benchmarking database.

Keywords: Word spotting, Characteristic Loci, Structural descriptors, Handwritten document analysis.

1 Introduction

There is an increasing interest to digitally preserve and provide access to historical document collections in libraries, museums and archives. The conversion of historical document collections to digital archives is of prime importance to society both in terms of information accessibility, and long-term preservation. Historical archives usually contain handwritten documents. Examples are unique manuscripts written by well known scientists, artists or writers; letters, trade forms or administrative documents kept by parish or municipalities that help to reconstruct historical sequences in a given place or time, etc. While machine printed documents, under a minimum of conditions, are easy to be read by OCR systems, the recognition of handwriting is still a scientific challenge. The state of the art achieves only good performance in constrained domains or with small vocabularies (e.g. bank-checks, postal automation).

Handwriting recognition has a number of difficulties. First, the physical degradation of documents due to lifetime use and careless storage produces holes,

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 628–635, 2011.

stains, winkles, ink bleed. Second, the scanning process is difficult when digitizing old documents, books, parchments, etc. Several degradations can appear: non stationary noise due to illumination changes, show through effect, low contrast, warping effect, etc. Finally, in multi-writer problems, there are variations in word shapes due to the different writing styles.

Handwritten document image retrieval has attracted a lot of interest from the field of document analysis and digital libraries. The problem can be addressed following two main directions. First, full image-to-text transcription, followed by an ASCII string search. Second, when images present noise and distortion and the full transcription is not easy, word spotting [7] is a useful strategy. Handwritten word spotting is the pattern classification task which consists in detecting keywords in handwritten document images. The words are represented by shape features, so the problem is formulated in terms of shape-based comparison.

In the literature we can find two word-spotting approaches, depending on how the input is specified: query-by-string and query-by-example. In query-by-string, the input is a text string. Character models are learned off-line and at runtime the character models are combined to form words and the probability of each word is evaluated. In query-by-example the input is an image of the word to search, and the output is a set of the most representative (sub)images in the database containing a similar word shape.

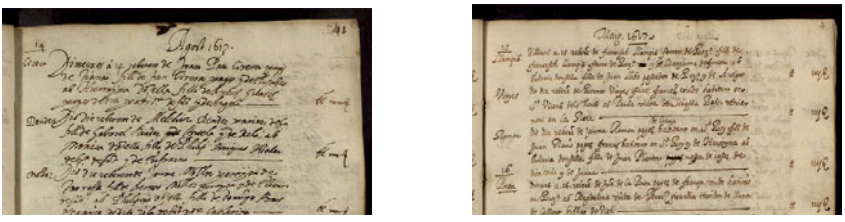
This work addresses the problem of handwritten word spotting in historical manuscripts following a query-by-example strategy. Classical approaches are based on contextual methods like Hidden Markov Model (HMM) [3] or Dynamic Time Warping (DTW) [6], using the sequential information of graphemes in a word. We propose a holistic approach using shape matching techniques. Our approach is inspired in Loci characteristic and allows to aggregate pseudo-structural information in the descriptor.

We have applied our work in an demography application. In particular, word spotting is applied to the manuscripts called *Llibre d'Esposalles*, a set of books written between 1451 and 1905. This corpus record marriage and the corresponding fees paid according to the social status of the families. It is conserved at the Archives of the Barcelona Cathedral and comprises 244 books with information on approximately 550.000 marriages celebrated in over 250 parishes. Each book contains the marriages of two years, and was written by a different writer. We show two examples of the documents in figure 1. Information extraction from these manuscripts is of key relevance for scholars in social sciences to study the demographical changes over five centuries.

The rest of the paper is structured as follows. In section 2 the proposed method of this work is described. Section 3 deals with the experimental results. Finally, the last section concludes the paper.

2 Word Spotting Approach

The objective of this work is word spotting for indexation and retrieval purposes. Thus given a query word image, we intend to locate instances of the same word



(a) (b)

Fig. 1. *Llibre d'esposalles* (Archive of Barcelona Cathedral, ACB)

class into the documents to be indexed. In this work inspired in some literature approaches, words are considered as shapes, and spotting is achieved through shape dissimilarity functions.

A spotting strategy has two requirements that can be separated in two major modules: the learning and the retrieval stage (Fig. 2). First, word images have to be mapped to a feature space considering shape features. Second, it is necessary to design an indexation structure allowing to formulate queries of word images and retrieve similar instances from the database in terms of shape similarity. Hence, features describing words have to be chosen so that they allow to cluster the space in classes in an unsupervised way, and also the features have to satisfy properties as coping with distortions and obtaining compact representations.

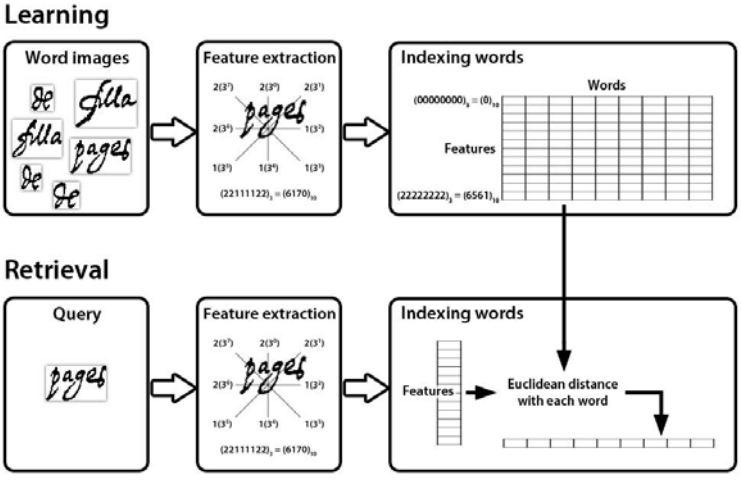


Fig. 2. Outline of the approach

Our approach is inspired in characteristic Loci feature [1,2]. We propose a descriptor based on pseudo-structural features. Given a word image, a feature vector based on Loci characteristics is computed at some characteristic points. Loci characteristics encode the frequency of intersection counts for a given

key-point in different direction paths starting from this point. As key-points it can be used contours, foreground pixels, background pixels or skeletons, depending on the application.

Once word images are encoded using a Loci-based descriptor, the indexation structure is organized as a hashing-like way where features are encoded as index keys and words are stored in a hashing structure. Afterwards, the word spotting is performed by a voting process after Loci vectors from the query word are indexed in the hashing table. Let us further describe the different steps in the following subsections.

2.1 Pre-processing Step

The quality of old documents can be affected by degradations. We perform a pre-processing step in order to improve the image quality for the subsequent processing and to segment the relevant parts. We first binarize the document. Then, we remove margins of the document that are likely to interfere with subsequent operations. The page is then segmented into lines using projection analysis techniques [5]. Once the lines are segmented, word segmentation is done using a similar technique. The projection function is smoothed with an Anisotropic Gaussian Filter [4].

In our approach, for each considered word, we extract the bounding box and do a fast rejection with the words that are very big or very small with regard to the mean of all the words of the document. As well, the bounding box that has few pixels of information is ruled out. This allows to drastically reduce the search space. Then, a noise removal operation is performed.

2.2 Feature Extraction

The characteristic Loci features were devised by Glucksman and applied to the classification of mixed-font alphabetic, as described in [2]. A characteristic Loci feature is composed by the number of the intersections in the four directions (up, down, right and left) (Fig. 3). For each background pixel in a binary image, and each direction, we count the number of intersections (an intersection means a black/white transition between two consecutive pixels). Hence, each key-point generates a codeword (called *Locu number*) of length 4.

This work presents a new feature descriptor based in the characteristic Loci features. We have introduced three variations of the basic descriptor:

- We have added the two diagonal directions, as we can see in figure 3. This gives more information to the descriptor and more sturdiness to the method.
- The number of the intersections is quantized. We have bounded the number of intersections in intervals. Each direction has a different quantization. This bounding generates a more robust feature.
- Two modes are implemented to compute the feature vector, namely background and foreground pixels are taken as key-points.

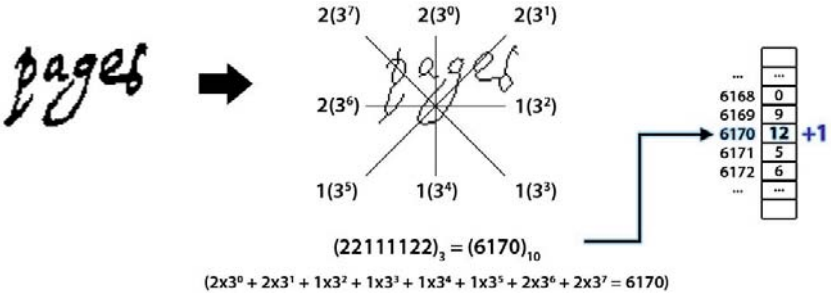


Fig. 3. Characteristic Loci feature of a single point of the word page

The skeleton of the image is computed by an iterative thinning until reaching lines of 1 pixel width.

The feature vector is computed by assigning a label to each background (or foreground) pixel as show in Fig. 3. The features are computed according to the number of intersections with the background pixels of the image in right, upward, left and downward directions. In previous works, the characteristic Loci method has been applied for digits and isolated letter recognition. In this work, to reduce the dimension of the feature space, the maximum number of intersections has been limited to 3 values (0, 1 and 2). By delimiting the number of possible values we reduce the number of combinations. The length of the feature vector is proportional to the number of possible values. For example, with 3 possible values and 8 directions, we obtain 3^8 (6.561) combinations; with 4 possible values we have 3^4 (65.536). It increases in exponential way and the computational cost (and time) increases in the same way.

Characteristic Loci feature was designed for digit and isolated letter recognition, and the number of intersections was bounded. The original approach uses the same interval in all directions. In this work we have also bounded the number of intersections. For each direction we have defined a different interval for each value. The horizontal direction has a larger interval than the vertical direction. In the original approach the digits or characters have a similar height and width, but in our approach the width of the words is usually bigger than the height. According to the dimensions of the words the range of the intervals are directly proportional. Diagonal directions are a combination of the two other directions. Table 1 shows the intervals for each direction.

Table 1. Intervals for each direction in characteristic Loci feature

	Values		
direction	**0**	**1**	**2**
Vertical	{0}	[1, 2]	[3, +∞]
Horizontal	{0}	[1, 4]	[5, +∞]
Diagonal	{0}	[1, 3]	[4, +∞]

According to the above encoding, for each background pixel, an eight-digit number in base 3 is obtained. For instance, the *Locu number* of point P in Fig. 3 is $(22111122)_3 = (6170)_{10}$. The *Locu numbers* are between 0 and 6.561 ($= 3^8$). This is done for all background pixels. In this case, the dimension of the feature space becomes 6.561. Each element of this vector is a *Locu number*, and represents the total number of background pixels with this *Locu number*.

2.3 Matching Words

The retrieval process of this approach consists in organizing the feature codewords in a look up table M (Fig. 2). Columns of M represent the words (w) of the database. Rows correspond to all the possible combinations that can appear using characteristic Loci features (f) . *M(f, w)* means that the feature f is presented in word w. For this work, we have 8 directions and each one has three different values. So, we have $3^8 (= 6.561)$ possible combinations. The feature vector is like a histogram of *Locu numbers*.

Classification process consists in searching the best matching of the query with all the words of M (Fig. 2). The chosen query is used to extract the vector of features. This vector is used to match the query with all the words of the database. In the retrieval step we have applied two distance formulations, namely Euclidean and Cosine, to match similar words.

3 Experimental Results

The experiments have been performed using 30 documents of the volume 69 of the Cathedral of Barcelona archives, using 10 key-words as queries.

We have implemented the well know approach developed by Rath and Manmatha [6] to compare it with our proposed methodology. They present an algorithm for matching handwritten words in noisy historical documents using Dynamic Time Warping (DTW). To evaluate the retrieval performance of the system with a query word against a handwritten document image, we use the standard precision and recall measures. We have done the following experiments: in the first one we have evaluated the performance using different characteristic pixels, background and foreground pixels of the word image, using different mask sizes, size of regions of interest to compute the number of intersections for each key-point; in the second experiment we compare different distance measures with different key-points.

In table 2, we show the performance of the word retrieval system using different characteristic pixels as reference. The precision and recall table is computed according to different mask size. We observe that when we increase the mask size, the results are better. Using 80 or more pixels we obtain similar results. Comparing the performance using background and foreground pixels, we observe that using background pixels we obtain better results. Using background pixels as reference, the number of pixels that gives information to the feature vector is higher than using foreground pixels.

Table 2. Accuracy using different characteristic pixels and mask sizes

	Background pixels		Foreground pixels	
size mask	Precision	recall	precision	recall
15	68, 3%	66, 6%	39, 1%	77, 9%
20	76, 3%	68, 6%	43, 1%	79, 9%
40	78, 6%	64, 5%	44, 9%	79, 6%
80	82, 5%	67, 0%	46, 5%	79, 1%
100	82, 5%	67, 0%	46, 5%	79, 1%

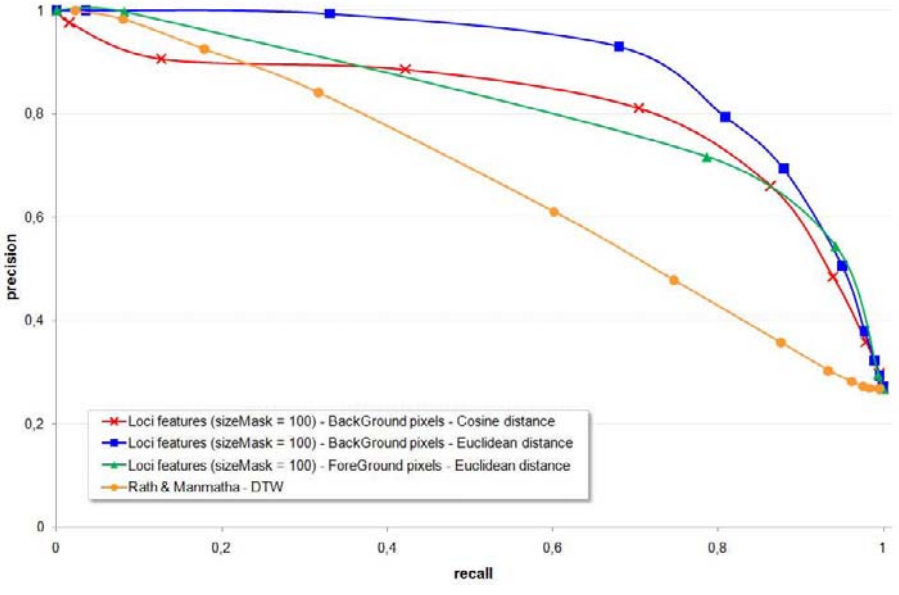


Fig. 4. Experimental results. Precision-Recall curve.

Figure 4 shows the results of our experiments compared with the algorithm of Rath and Manmatha [6] employing the same database in all of them. In our experiments we have used different distance measures to compute the distance between words descriptors. All the experiments are done using a mask size of 100 pixels. First, we have compared the performance using background and foreground pixels using Euclidean distance as measure. We can observe that we obtain better results using background pixels. We have used these measures because Euclidean distance gives us the magnitude of difference between two word images, while cosine distance is a normalized measure which gives us a measure of how similar two word images are.

The next experiment consists in evaluating the performance using different distance measures (euclidean and cosine distance). The results are better using Euclidean distance than using cosine distance. Finally, we evaluate the performance in comparison to the algorithm of Rath and Manmatha. In their work they use Dynamic Time Warping to compute the distance between words. We

can observe that our approach obtains better results. The better performance of our approach is due to the nature of the descriptor. While Rath and Manmatha use a pixel-based column descriptor, our descriptor captures more global information, including the structure of the word strokes.

4 Conclusions

In this paper we have presented a handwritten word spotting approach based on a pseudo-structural descriptor (inspired in characteristic Loci [1,2]) organized in a hash structure.The queried word image is matched with segmented words of handwritten documents to determine the distance that might indicate a correspondence. We have tested our method with a database of old handwritten documents from the Cathedral of Barcelona. Our approach outperforms other state-of-art methods, in particular the well known approach of Rath and Manmatha [6]. The results of our experiments show that using background pixels as reference pixels we obtain better results than using foreground pixels. Moreover using Euclidean distance we obtain better results than using cosine distance.

Acknowledgments. The authors thank to the *CED-UAB*, for providing the images. This work has been partially supported by the Spanish projects TIN2008-04998, TIN2009-14633-C03-03 and CSD2007-00018, by the EU project ERC-2010-AdG-20100407-269796 and by a research grant of the UAB (471-01-8/09).

References

1. Ebrahimi, A., Kabir, E.: A pictorial dictionary for printed Farsi subwords. Pattern Recognition Letters 29, 656–663 (2008)
2. Glucksman, H.A.: Classification of mixed-font alphabets by characteristic loci. In: Proc. IEEE Comput. Conf., pp. 138–141 (1967)
3. Kessentini, Y., Paquet, T., Hamadou, A.B.: Off-line handwritten word recognition using multi-stream hidden Markov models. Pattern Recognition Letters 31(1), 60–70 (2010)
4. Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1212–1225 (2005)
5. Papavassiliou, V., Stafylakis, T., Katsouros, V., Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition 43, 369–377 (2010)
6. Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), vol. 2, pp. 521–527 (2003)
7. Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal of Document Analysis and Recognition (IJDAR) 9, 139–152 (2006)

Identification of Erythrocyte Types in Greyscale MGG Images for Computer-Assisted Diagnosis

Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin,
Faculty of Computer Science and Information Technology,
Zolnierska 49, 71-210, Szczecin, Poland
dfrejlichowski@wi.zut.edu.pl

Abstract. In the paper an algorithm for the recognition of erythrocytes is presented and experimentally evaluated. The objects of interest are localised and extracted from digital microscopic images, stained by means of the MGG (May-Grunwald-Giemsa) method in greyscale. The area covering a single red blood cell (RBC) is transformed from Cartesian to polar co-ordinates. Later, the two-dimensional Fourier transform is applied to the resultant image. Finally, the subpart of the spectrum is selected in order to represent an object. This description (Polar-Fourier Greyscale Descriptor) is matched with the templates represented in the same way. The smallest dissimilarity measure indicates the recognised erythrocyte type. When using this approach every RBC is investigated, and basing on the whole knowledge about the number of particular types of erythrocytes present in an image a diagnosis can be made.

Keywords: Computer-Assisted Diagnosis, Erythrocyte Recognition, MGG images.

1 Introduction

The abnormalities in blood particles may cause some serious diseases. Amongst them the deformations of erythrocytes can lead to various kinds of anaemia or malaria. It comes from the fact that deformed red blood cells (RBCs) cannot deliver oxygen properly, which results in that the blood circulation is non-regulated. For this reason the computer assisted automatic diagnosis of selected diseases can be based on erythrocyte shapes. In that case, the computer-based analysis can use the digital microscopic images stained by means of the MGG (May-Grunwald-Giemsa) method. Few examples of this kind of input images (in greyscale) are provided in Fig. 1.

Obviously, the problem of automatic diagnosis based on microscopic images is not new. However, in most cases the analysis of various particles is investigated, e.g. in [1] 12 categories of particles in human urine were classified. It is desirable, yet more difficult, and therefore less efficient than the limitation to one type of particles. Moreover, even when this limitation is assumed, the leukocytes are usually analysed ([2,3]). Nevertheless, lately the scientific interest in automatic

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 636–643, 2011.

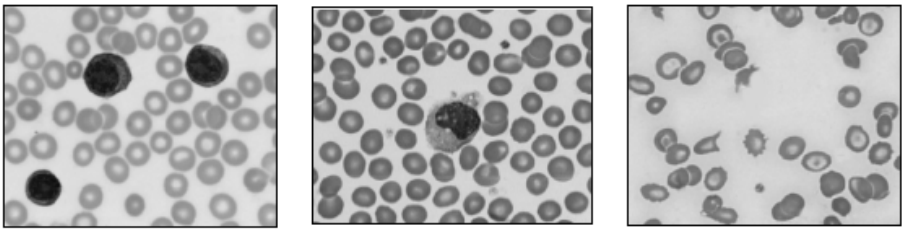


Fig. 1. Examples of digital MGG images (in greyscale) of human blood as an input data for the approach presented in the paper

recognition of red blood cells is arising. For example, in [4] some histogram features were applied for this purpose. Another two approaches developed so far were based on deformable templates ([5]) and morphological operators ([6]).

Apart from the several examples mentioned above two algorithms are especially worthy of attention. They are the most similar to the approach provided in this paper. The first one is based on mathematical morphology and polar-logarithmic transform ([7]). The authors were concentrated on the analysis of circular properties of the erythrocyte shape. Therefore, they used only five kinds of this blood cell and gave them very general names ('Normal', 'Mushroom', 'Spicule', 'Echinocyte', and 'Bitten', see Fig. 2), which is clearly not adequate for the automatic diagnosis.

In [8] the template matching for recognition of extracted erythrocyte shapes was used. For this purpose three shape description techniques were applied. In comparison with the previous paper, it is important to note that twelve distinct types of erythrocytes were used (schistocyte, dacrocyte, acantocyte, echinocyte, ovalocyte, normocyte, stomatocyte, Mexican hat cell, spherocyte, leptocyte, annular erythrocyte, drepanocyte, see Fig. 2), which is closer to the medical approach for the diagnosis. The same set of classes was applied in the work presented in this paper.

The remaining part of the paper is organised as follows. The second section describes the pre-processing technique applied in order to localise and extract

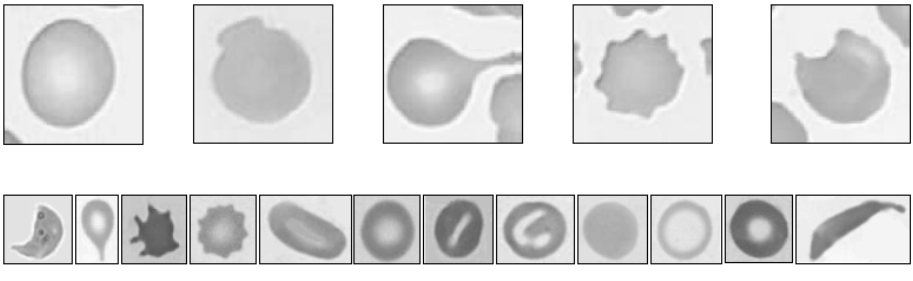


Fig. 2. Two various sets of erythrocytes used in the previous works, covering five classes (top row, [7]), and twelve (bottom row, [8])

particular erythrocytes. The third section provides in details the proposed algorithm for red blood cells representation and recognition. The fourth section presents the experimental results, and finally, the last section concludes the paper and discusses some suggestions for further work.

2 Description of the Used Pre-processing Method

Although the identification of an erythrocyte's type constitutes the main topic of the paper, in order to preserve the completeness of the description, in this section the algorithm used for the localisation of each particular cell is briefly provided. It is based on the approach presented in [8] and starts with the conversion of the image into greyscale, if it is necessary. Then, the modified histogram thresholding is performed. It starts with the derivation of the histogram $h(l_k)$ ([8]):

$$h(l_k) = \sum_{k=1}^{m} b(k, l_k), \tag{1}$$

where:

$$b(k, l_k) = \begin{cases} 1, & \text{if } k = l_k \\ 0, & \text{if } k \neq l_k \end{cases}. \tag{2}$$

Later, the achieved histogram is smoothed through bins averaging, with the length of the window established experimentally as ([8]):

$$c(j) = \frac{\sum_{i=j-m}^{j+m} h(i)}{2m+1}, \tag{3}$$

where:
j — number of a bin,
$c(j)$ — averaged histogram value for bin,
$h(i)$ — histogram value before averaging,
m — number of bins taken in left and right neighborhood of j-th bin.

The threshold value depends on the number of bins with values other than zero in a histogram (distinct grey-levels found for an image). If it is higher than 150, the threshold t is derived as the number of a bin with the minimal histogram value amongst the bin numbers belonging to the interval ([8]):

$$t \in (c_{\max} - \left\lfloor \frac{v}{4} \right\rfloor, c_{\max}), \tag{4}$$

where:
c_{max} — number of the highest bin,
v — number of gray-levels found in particular image (number of non-zero bins in histogram).

If it is smaller, the threshold t is derived as the number of a bin with the minimal histogram value amongst the bin numbers belonging to the interval ([8]):

$$t \in (c_{\max} - \left\lfloor \frac{v}{4} \right\rfloor, c_{\max} - 20). \tag{5}$$

After the thresholding particular cells were localised. For this purpose regions of each separate objects were traced. Obviously, only objects entirely placed within an image were analysed. In order to reject thrombocytes and leukocytes, and work only with erythrocytes the area of extracted regions was analysed. For each case, if it was significantly bigger (leukocytes) or smaller (thrombocytes) then it was rejected. Additionally, this process rejects some occluded shapes, difficult to recognise. However, it is possible that some undesirable particular remain, if they have the area similar to erythrocytes. In order to avoid this problem, the analysis of their histogram is applied, since it is different for particular particles. It comes from the fact that thrombocytes and leukocytes have almost black parts inside. Moreover, the histogram equalisation of the image before the binarisation is also performed, what reduces the number of occluded shapes.

As a result of the process described above, by using the established co-ordinates of a cell, the rectangular subpart of the greyscale image with it could be extracted and used in the next step.

3 Representation and Recognition of a Cell

The most important element of the approach is the use of two transforms for a cell image. Firstly, the transformation from Cartesian to polar co-ordinates is used, and later the 2D Fourier transform of the achieved polar image. Finally, the subspectrum is extracted in order to represent an object. However, few additional steps have to be performed in order to avoid some problems which may occur.

Sometimes, the extracted subimage has a low quality. This may happen for example when the cells within an image are small, if noise is present, or if lossy compression of an image was used. In order to reduce the influence of the above problem, firstly the median filtering (with square mask 3×3 pixels) and smoothing using low-pass filtering (again mask with 3×3 size, containing ones, norm factor equal to 9) were applied.

The second problem to consider was the varying size of the subimage, especially the irregular size of an erythrocyte itself. In order to achieve the best polar representation of a subimage, it was expanded according to the maximal distances from the centre of an erythrocyte, calculated by means of the moment theory. The new areas of the subimage which appeared were filled in with the colour of the background. The prepared subimage was later transformed into polar co-ordinates. That gave in result a greyscale image again. In order to make the representation invariant to scaling, the achieved image is resized to the constant size. Here, the 128×128 size was assumed. However, other values can be used as well. The resultant image was subjected to the two-dimensional Fourier transform. Finally, the square subpart from the absolute spectrum with the size of 10×10 was extracted, concatenated, and stored as a vector containing 100 elements.

The above general discussion leads to the following formulation of the algorithm for representation of erythrocytes extracted from digital greyscale images — ***Polar-Fourier Greyscale Descriptor***:

Step 1. Perform the 3px median filter on the input subimage I.
Step 2. Perform low-pass filter on the subimage I, using convolution with $norm = 9$ and mask 3×3 size, containing ones.
Step 3. Calculate the centroid O basing on the simple moment theory:
Step 3a. Calculate m_{00}, m_{10}, m_{01} using the general formula ([9]):

$$m_{pq} = \sum_x \sum_y x^p y^q I(x, y). \tag{6}$$

Step 3b. Calculate the centroid $O(x_c, y_c)$ using the formulas ([9]):

$$x_c = \frac{m_{10}}{m_{00}}, \qquad y_c = \frac{m_{01}}{m_{00}}. \tag{7}$$

Step 4. Find the maximal distances d_{maxX} , d_{maxY} from the centroid to the boundaries of the subimage I:
Step 4a. Calculate the distances from O to particular boundaries of the subimage I:

$$d_1 = x_c, \qquad d_2 = M - x_c, \qquad d_3 = y_c, \qquad d_4 = N - y_c, \tag{8}$$

where: (M, N) — size of the subimage I.
Step 4b. Select the highest values of (d_1, d_2, d_3, d_4) for particular axes:

$$d_{maxX} = max(d_1, d_2), \qquad d_{maxY} = max(d_3, d_4). \tag{9}$$

Step 5. Expand I into X direction by $d_{maxX} - x_c$ pixels, and into Y direction by $d_{maxY} - y_c$ pixels. Fill in the new areas in I with the colour of the background.
Step 6. Transform I into polar co-ordinates (into new image P), using the formulas:

$$\rho_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}, \qquad \theta_i = atan\left(\frac{y_i - y_c}{x_i - x_c}\right). \tag{10}$$

Step 7. Resize the P image into the constant rectangular size, e.g. 128×128.
Step 8. Calculate the absolute spectrum of the 2D Fourier transform ([10]):

$$C(k, l) = \frac{1}{HW} \left| \sum_{h=1}^{H} \sum_{w=1}^{W} P(h, w) \cdot e^{(-i\frac{2\pi}{H}(k-1)(h-1))} \cdot e^{(-i\frac{2\pi}{W}(l-1)(w-1))} \right|, \tag{11}$$

where:
H, W — height and width of the polar image P,
k — sampling rate in vertical direction ($k \geq 1$ and $k \leq H$),

l — sampling rate in horizontal direction ($l \geq 1$ and $l \leq W$),
$C(k, l)$ — value of the coefficient of discrete Fourier transform in the coefficient matrix in k row and l column,
$P(h, w)$ — value in the image plane with coordinates h, w.
Step 9. Select the square subpart of the achieved absolute spectrum with the indices $1, \ldots, 10$ for both axes and after concatenation put it into vector V.

The achieved vector V represents the object. Fig. 3 presents some examples of erythrocytes and their representations achieved by means of the proposed algorithm.

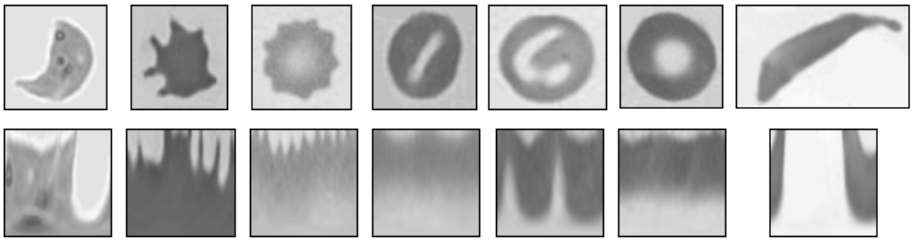


Fig. 3. Examples of tested erythrocytes (top) and their representations achieved by means of the Polar-Fourier Greyscale Descriptor (bottom)

According to the template matching approach the description of an object is matched with base templates, described using the same algorithm. For this purpose the Euclidean distance as the dissimilarity measure is used. The smallest value simply indicates the recognised class; in our case — type of a red blood cell.

4 Conditions and Results of the Experiment

The algorithm, described in the previous section, was tested using 55 MGG images, converted into greyscale. Using the pre-processing method described in Section 2, every cell was localised and extracted separately. Then, it was represented by means of the proposed algorithm and matched using the dissimilarity measure with the templates. Obviously, they were represented in the same way. In order to increase the efficiency of the algorithm five various objects were used to represent a class. That gave sixty template objects, five for each of the twelve classes.

The number of cells within a single MGG image varies significantly, from several dozens to several hundreds or even more than a thousand. Hence, the total number of extracted and analysed cells during the experiments was close to ten thousand. The achieved recognition rates for particular erythrocyte types are provided in Table 1. As we can see they are different for various classes. The best results, close to 100 percent of recognition, were achieved for acanthocytes,

normocytes, and leptocytes. For spherocytes the result was ideal. On the other hand, in four cases (i.e. schistocyte, ovalocyte, Mexican hat cell, and annular erythrocyte) the recognition rate was merely close to 70%. Nevertheless, the total average recognition result was close to 86%. On first sight it seems to be far from an ideal. However, two items have to be taken into account. First of all, the poor quality of microscopic images may cause difficulties in the proper recognition of an object. An ideal identification of particular cells is not possible. Secondly, for the automatic (or semi-automatic) diagnosis the results do not have to be ideal. In most cases even the presence of one abnormal red blood cell is sufficient to perform a diagnosis or indicate a possibility of a disease. Considering the large number of particles in an image it is easy to achieve, even with the performance of the algorithm far from one hundred per cent. As it was discussed in [8] the average recognition result which is higher than 80% is enough to perform a correct diagnosis, what was confronted there with the diagnoses made by humans.

Table 1. Average recognition rates (RR) achieved for particular classes of erythrocytes

Class	schistocyte	dacrocyte	acantocyte	echinocyte	ovalocyte
RR	69%	83%	98%	89%	74%
Class	normocyte	stomatocyte	Mexican hat cell	spherocyte	leptocyte
RR	97%	90%	72%	100%	97%
Class	annular erythrocyte	drepanocyte			**TOTAL**
RR	76%	81%			**86%**

5 Conclusions and Future Plans

In the paper the problem of automatic diagnosis based on digital microscopic images of human blood was touched on. The automatic recognition of erythrocytes' types was suggested. The idea is based on the fact that there is a group of diseases caused by abnormal red blood cells, which can be easily diagnosed by means of their shapes.

The proposed approach for the diagnosis is based on the calculation of the particular types of red blood cells present within an image. For this purpose twelve types of erythrocytes were used. Their identification was performed using the method based in general on two transforms. For the localised and extracted subpart of the image containing a cell firstly the transformation from Cartesian to polar co-ordinates is used, and later the two-dimensional Fourier transform of the achieved polar image. Finally, the subspectrum is taken as a description of the object. This description is matched with sixty templates (five for each type of erythrocyte). The smallest value of the dissimilarity measure indicates the recognised class.

The presented algorithm was tested by means of cells localised and extracted from 55 greyscale MGG images. The average recognition rate for all classes was close to 86%, varying from 69% for schistocyte to 100% for spherocyte. This

result can be considered as sufficient enough because of the large number of objects to identify within an image. Thanks to this the overall diagnosis is less dependent on single incorrect identifications. Moreover, the ideal result is almost impossible due to the character of real microscopic blood images.

The future plans are above all concentrated on performing an automatic diagnosis based on the approach presented in this paper on a much bigger collection of images. Obviously, the precise formulation of all decision rules for the diagnosis also has to be stated.

References

1. Ranzato, M., Taylor, P.E., House, J.M., Flagan, R.C., LeCun, Y., Perona, P.: Automatic recognition of biological particles in microscopic images. Pattern Recognition Letters 28(1), 31–39 (2007)
2. Song, X.B., Abu-Mostafa, Y., Sill, J., Kasdan, H., Pavel, M.: Robust image recognition by fusion of contextual information. Information Fusion 3(4), 277–287 (2002)
3. Sabino, D.M.U., da Fontoura Costa, L., Rizzatti, E.G., Zago, A.M.: A texture approach to leukocyte recognition. Real-Time Imaging 10(4), 205–216 (2004)
4. Díaz, G., González, F.A., Romero, E.: A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic imageses. Journal of Biomedical Informatics 42(2), 296–307 (2009)
5. Bronkorsta, P.J.H., Reinders, M.J.T., Hendriks, E.A., Grimbergen, J., Heethaar, R.M., Brakenhoff, G.J.: On-line detection of red blood cell shape using deformable templates. Pattern Recognition Letters 21(5), 413–424 (2000)
6. Di Ruberto, C., Dempster, A., Khan, S., Jarra, B.: Analysis of infected blood cell images using morphological operators. Image and Vision Computing 20(2), 133–146 (2002)
7. Luengo-Oroz, M.A., Angulo, J., Flandrin, G., Klossa, J.: Mathematical Morphology in Polar-Logarithmic Coordinates. Application to Erythrocyte Shape Analysis. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3523, pp. 199–206. Springer, Heidelberg (2005)
8. Frejlichowski, D.: Pre-processing, Extraction and Recognition of Binary Erythrocyte Shapes for Computer-Assisted Diagnosis Based on MGG Images. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) ICCVG 2010. LNCS, vol. 6374, pp. 368–375. Springer, Heidelberg (2010)
9. Hupkens, T.M., de Clippeleir, J.: Noise and intensity invariant moments. Pattern Recognition Letters 16(4), 371–376 (1995)
10. Kukharev, G.: Digital Image Processing and Analysis. SUT Press (1998) (in Polish)

Classification of High Dimensional and Imbalanced Hyperspectral Imagery Data*

Vicente García, J. Salvador Sánchez, and Ramón A. Mollineda

Institute of New Imaging Technologies
Department of Computer Languages and Systems, Universitat Jaume I
Av. Sos Baynat s/n, 12071
Castelló de la Plana, Spain
{jimenezv,sanchez,mollined}@lsi.uji.es

Abstract. The present paper addresses the problem of the classification of hyperspectral images with multiple imbalanced classes and very high dimensionality. Class imbalance is handled by resampling the data set, whereas PCA is applied to reduce the number of spectral bands. This is a preliminary study that pursues to investigate the benefits of using together these two techniques, and also to evaluate the application order that leads to the best classification performance. Experimental results demonstrate the significance of combining these preprocessing tools to improve the performance of hyperspectral imagery classification. Although it seems that the most effective order of application corresponds to first a resampling algorithm and then PCA, this is a question that still needs a much more thorough investigation.

1 Introduction

Hyperspectral sensors are characterized by a very high spectral resolution that usually results in hundreds of observation channels [22]. Although this allows to address many applications requiring very high discrimination capabilities in the spectral domain [3], the huge amount of data available makes complex the classification of hyperspectral images. In this classification context, another important drawback is that the hyperspectral information is commonly represented by a very large number of features (spectral bands), which are usually highly correlated [22, 23].

A complex situation frequently ignored in hyperspectral imaging refers to the presence of severely skewed class priors. This situation is generally known as the class imbalance problem [12]. A data set is said to be imbalanced when one of the classes (the minority one) is heavily under-represented in comparison to the other (the majority) class. Because of samples of the minority and majority classes usually represent the presence and absence of rare cases respectively, they are also known as positive and negative examples. It has been observed that class imbalance often leads to poor classification performance in many real-world applications, especially for the minority classes.

* Partially supported by the Spanish Ministry of Education and Science under grants CSD2007–00018, AYA2008–05965–0596–C04–04/ESP and TIN2009–14205–C04–04, and by Fundacio Caixa Castello–Bancaixa under grant P1–1B2009–04.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 644–651, 2011.

Most of the approaches to tackle the imbalance problem have been proposed both at the data and algorithmic levels. Data-driven methods consist of balancing the original data set, either by over-sampling the minority class [4, 11] and/or by under-sampling [9, 17] the majority class until the classes are approximately equally represented. Within this group, we can also find several algorithms for feature selection [1, 15, 18, 25, 26]. At the algorithmic level, solutions include internally biasing the discrimination-based process [7, 8] and assigning distinct costs to the classification errors [19, 20, 30].

Although class imbalance has been extensively studied for binary classification problems, very few approaches deal with multi-class imbalanced data sets, as is the case of remote sensing applications. In the particular context of hyperspectral imagery, some proposals are adjustments of conventional learning algorithms [2, 16, 28], whereas others use classifier ensembles [24, 27] or feature selection techniques [5].

In this paper, some well-known strategies to cope with class imbalance are investigated for the classification of hyperspectral imagery acquired by the Airborne Visible/Infrared Imaging Spectrometrer (AVIRIS[1]). The problem is of great relevance since these image data present both very high dimensionality and multiple imbalanced classes, what certainly provides additional challenges in the framework of remote sensing classification. In order to face such a problem, this work focuses on the joint use of feature extraction and resampling techniques, and explores the order in which they should be applied to achieve the best classification results.

The rest of the paper is organized as follows. Section 2 describes the methodology proposed to handle class imbalance and high dimensionality, and also briefly reviews the classifiers used in this work. Next, Sect. 3 contains the experiments on a real hyperspectral image database and discusses the most important findings. Finally, Sect. 4 concludes the present study and outlines possible directions for future research.

2 Methodology

This section provides an overview of the method here proposed to handle and classify remote sensing data according to the two issues of interest previously pointed out. In a first stage, the hyperspectral image data set will be preprocessed with the double aim of balancing the skewed classes and reducing the number of features/bands, albeit not necessarily in this order. The second stage will consist of classifying the resulting set after overcoming those two problems. Note that only those algorithms that will be further used in the experiments are described in the present section.

2.1 Preprocessing

Taking the particular characteristics of hyperspectral data sets into account, most imaging tasks could usually benefit from the application of some preprocessing techniques. Here we concentrate on a common situtation in which the data set consists of multiple imbalanced classes in a high dimensional representation space.

[1] http://aviris.jpl.nasa.gov/

Balancing the Classes. Data level methods for balancing the classes consists of resampling the original data set, either by over-sampling the minority class or by under-sampling the majority class, until the classes are approximately equally represented. Both strategies can be applied in any learning system since they act as a preprocessing phase, thus allowing the system to receive the training instances as if they belonged to a well-balanced data set. By using this strategy, any bias of the learning system towards the majority class due to the skewed class priors will hopefully be eliminated.

The simplest method to increase the size of the minority class corresponds to random over-sampling, that is, a non-heuristic method that balances the class distribution through the random replication of positive examples. Nevertheless, since this method replicates existing examples in the minority class, overfitting is more likely to occur. Chawla et al. [4] proposed an over-sampling technique that generates new synthetic minority samples by interpolating between several preexisting positive examples that lie close together. This method, called SMOTE (Synthetic Minority Over-sampling TEchnique), allows the classifier to build larger decision regions that contain nearby samples from the minority class.

On the other hand, random under-sampling [14, 29] aims at balancing the data set through the random removal of negative examples. Despite its simplicity, it has empirically been shown to be one of the most effective resampling methods. Unlike the random approach, many other proposals are based on a more intelligent selection of the negative examples to be eliminated. For instance, the one-sided selection technique [17] selectively removes only those negative samples that either are redundant or that border the minority class examples (assuming that these bordering cases are noise).

Dimensionality Reduction. The reduction in the hyperspectral representation space can be carried out by means of feature selection or extraction techniques. In both approaches, the aim is to reduce the number of bands, without much loss of information. The process of feature selection is to choose a representative subset of features from the original data by assessing its discrimination capabilities according to statistical distance measures among classes (e.g., Bhattacharyya distance, Jeffries-Matusita distance, and the transformed divergence measure). The feature extraction approach addresses the problem of dimensionality reduction by projecting the data from the original feature space onto a low-dimensional subspace, which contains most of the original information [13].

Probably the most widely-known feature extraction method corresponds to Principal Component Analysis (PCA), which seeks to reduce the dimension of the data by finding a few orthogonal linear combinations of the original variables with the largest variance. It involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components.

2.2 Classification

We assume that there exists a set of n previously labeled examples (training set, TS), say $X = \{(x_1, \omega_1), (x_2, \omega_2), \ldots, (x_n, \omega_n)\}$, where each element has an attribute vector x_i and a class label ω_i. Two traditional classification techniques will be used in the experimental study: the nearest neighbor rule and a decision tree.

Nearest Neighbor Rule. One of the most popular non-parametric classification approaches corresponds to the k nearest neighbor (kNN) decision rule [6]. In brief, this classifier consists of assigning a new input sample $\mathbf{x}$ to the class most frequently represented among the k closest examples in the TS, according to a certain dissimilarity measure (e.g., the Euclidean distance). A particular case is when $k = 1$, in which an input sample is decided to belong to the class indicated by its closest neighbor.

The characteristics of the kNN classifier need the entire TS stored in computer memory, what causes large time and memory requirements. On the other hand, the kNN rule is extremely sensitive to the presence of noisy, atypical and/or erroneously labeled examples in the TS.

Decision Tree. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. The nodes of the tree evaluate the existence or significance of individual features. Following a path from the root to the leaves of the tree, a sequence of such tests is performed resulting in a decision about the appropriate class of new objects.

The decision trees are constructed in a top-down fashion by choosing the most appropriate attribute each time. An information-theoretic measure is used to evaluate features, which provides an indication of the "classification power" of each feature. Once a feature is chosen, the training data are divided into subsets, corresponding to different values of the selected feature, and the process is repeated for each subset until a large proportion of the instances in each subset belongs to a single class.

3 Experiments and Results

The experiments were carried out on the 92AV3C data set[2], which corresponds to a hyperspectral image (145×145 pixels) taken over Northwestern Indiana's Indian Pines by the AVIRIS sensor in June 1992 and employed to recognize different land-cover classes. Although the AVIRIS sensor collects 224 spectral bands, four of these contain only zero values and so they can be removed, leaving a total of 220 non-zero bands. The ground truth data show that the image has 17 classes, although only 16 classes belonging to different crop types, vegetation, man-made structures or other kinds of land were used (see Table 1). The omitted class contains unlabeled pixels, which presumably correspond to uninteresting regions or were too difficult to label.

In order to increase the statistical significance of the experimental results, classification accuracies were averaged over 30 different random partitions (2/3 of pixels for training and the rest for testing) of the original data set, preserving the prior class probabilities of each and the statistical independence between the training and test sets of every partition. The training sets were preprocessed by two different resampling algorithms, SMOTE and random under-sampling (RUS), to handle the class imbalance, and also by PCA for dimensionality reduction by retaining those principal components with a variance of 0.95. Because of difficulty to determine which classes to resample, we divided the biggest class (Soybeans–min) into four blocks (each one with 25% of

[2] `https://engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html`

Table 1. Number of training and test pixels per class, along with the relative percentage of samples that belong to each class

Class	Training	Test	%
1. Stone–steel towers	63	32	0.92
2. Hay–windrowed	326	163	4.72
3. Corn–min	556	278	8.05
4. Soybeans–notill	645	323	9.34
5. Alfalfa	36	18	0.52
6. Soybeans–clean	409	205	5.92
7. Grass/Pasture	331	166	4.79
8. Woods	863	431	12.48
9. Bldg–Grass–Trees—Drives	253	127	3.67
10. Grass/pasture–mowed	17	9	0.25
11. Corn	156	78	2.26
12. Oats	13	7	0.19
13. Corn–notill	956	478	13.83
14. Soybeans–min	1645	823	23.81
15. Grass/Trees	498	249	7.21
16. Wheat	141	71	2.05

samples). Based on this, the remaining classes were over-sampled to reach 25%, 50% and 75% the size of the majority class. Similarly, the under-sampling was applied removing 25%, 50% and 75% of samples according to the size of the biggest class.

The J48 decision tree (an open source Java implementation of the very popular C4.5 algorithm) and the 1NN classifier were applied to sets that were preprocessed and also to each original training set (without any preprocessing). All hyper-parameters of the classifiers were set to the default values suggested in the WEKA toolkit [10]. Apart from calculating the average accuracy of each individual class to evaluate the effect of the preprocessing techniques on the majority and minority classes separately, the mean of these individual accuracies was also computed in order to have an overall estimate of the performance. For the sake of clarity, we averaged the three percentages of resampling (25%, 50% and 75%) in one single result.

3.1 Analysis of Results

Table 2 reports the mean of the accuracies measured separately on each class when using J48 and 1NN to classify the test samples. As can be seen, the use of PCA (individually or jointly with some resampling algorithm) produces an important decrease in 1NN performance, whereas both resampling techniques outperform the accuracies achieved on the original set. These results are much less significant with the decision tree. In the case of the 1NN classifier, PCA probably fails because the database here used includes noisy bands due to the effect of atmospheric absorption [22, 21]. It is known that the kNN classifiers are very sensitive to noise in the training set and thus, the 1NN classifier seems to require a previous step consisting of the removal of those noisy bands or the application of some editing/filtering algorithm.

Table 2. Mean of accuracies of the 16 classes

	J48	1NN
Original	0.697	0.622
RUS	0.691	0.631
SMOTE	0.707	0.719
PCA	0.620	0.393
PCA+RUS	0.619	0.397
PCA+SMOTE	0.631	0.440
RUS+PCA	0.644	0.387
SMOTE+PCA	0.710	0.456

Table 3. Average classification accuracy on each class

	J48															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Original	0.867	0.953	0.598	0.663	0.637	0.516	0.833	0.916	0.521	0.599	0.467	0.432	0.632	0.736	0.884	0.897
RUS	0.888	0.952	0.616	0.677	0.596	0.576	0.843	0.880	0.556	0.597	0.500	0.412	0.594	0.593	0.868	0.904
SMOTE	0.898	0.925	0.602	0.675	0.599	0.561	0.827	0.898	0.580	0.724	0.522	0.395	0.615	0.717	0.866	0.904
PCA	0.964	0.918	0.517	0.627	0.380	0.398	0.642	0.878	0.360	0.612	0.416	0.329	0.515	0.670	0.840	0.891
PCA+RUS	0.923	0.929	0.538	0.647	0.386	0.469	0.685	0.824	0.396	0.640	0.427	0.310	0.487	0.521	0.818	0.896
PCA+SMOTE	0.945	0.863	0.532	0.631	0.507	0.440	0.681	0.806	0.457	0.634	0.521	0.304	0.481	0.642	0.760	0.895
RUS+PCA	0.905	0.930	0.580	0.672	0.419	0.490	0.703	0.836	0.402	0.653	0.463	0.464	0.529	0.534	0.822	0.902
SMOTE+PCA	0.927	0.903	0.622	0.685	0.664	0.483	0.754	0.834	0.513	0.782	0.611	0.652	0.569	0.568	0.794	0.904
	1NN															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Original	0.918	0.943	0.505	0.582	0.313	0.457	0.743	0.887	0.362	0.462	0.413	0.325	0.528	0.687	0.912	0.923
RUS	0.929	0.945	0.548	0.647	0.326	0.541	0.788	0.827	0.404	0.440	0.469	0.366	0.514	0.539	0.892	0.927
SMOTE	0.957	0.823	0.554	0.636	0.722	0.551	0.832	0.747	0.524	0.800	0.682	0.839	0.470	0.597	0.830	0.946
PCA	0.854	0.814	0.238	0.275	0.111	0.199	0.367	0.773	0.236	0.075	0.127	0.100	0.365	0.514	0.677	0.566
PCA+RUS	0.860	0.818	0.291	0.334	0.108	0.275	0.440	0.672	0.271	0.073	0.170	0.103	0.348	0.334	0.652	0.599
PCA+SMOTE	0.892	0.572	0.267	0.294	0.328	0.247	0.433	0.652	0.319	0.384	0.283	0.358	0.325	0.444	0.569	0.679
RUS+PCA	0.859	0.826	0.277	0.325	0.116	0.263	0.413	0.661	0.253	0.066	0.164	0.094	0.331	0.331	0.643	0.569
SMOTE+PCA	0.892	0.584	0.314	0.291	0.331	0.255	0.514	0.667	0.319	0.353	0.312	0.385	0.329	0.433	0.597	0.716

It is also remarkable that SMOTE excels all the other approaches, irrespective of the classifier used. On the other hand, when comparing the different combinations of resampling and PCA, one can observe that the application of SMOTE and PCA in this order leads to the highest performance in terms of mean of the accuracy of each class. Note that the average number of bands given by PCA is 13 in all cases, that is, it obtains a very high dimensionality reduction.

In order to assess the effect of the preprocessing approaches on each class separately, Table 3 shows the average classification accuracy achieved for each individual class. Both resampling techniques consistently improve the accuracy of the classes with less than 1% of samples (1, 5, 10, and 12), but entail a slight reduction on the performance of the most represented classes (8, 13, and 14). It is worth noting that this degradation on the majority classes appears to be less significant when using SMOTE, what is in keeping with the mean of accuracies reported in Table 2.

If we focus on the results of PCA, it is interesting to note that this algorithm leads to a decrease in the performance of most classes, especially when used with the 1NN classifier. Surprisingly, the application of SMOTE before using PCA mitigates this effect, suggesting that it is important to balance the classes before reducing the dimensionality of hyperspectral data.

4 Conclusions and Further Extensions

The present paper has focused on classification of hyperspectral imagery with two complex characteristics: high dimensionality and severe skewed class distributions. The experimental study has allowed to draw some preliminary conclusions: (i) It results more important to balance the classes rather than reduce the dimensionality, at least in terms of accuracy; (ii) The best choice seems to be the application of SMOTE followed by PCA; and (iii) The J48 decision tree appears to be a more robust classifier than the 1NN for this particular hyperspectral database.

In hyperspectral imaging, selection is generally preferable to feature extraction because of two main reasons. On the one hand, feature extraction would need the whole (or most) of the original data representation to extract the new features, forcing to always deal with the whole initial representation of the data. Besides, since the data are transformed, some crucial and critical information might be compromised and distorted. Thus future research will be addressed to use some feature selection algorithm instead of PCA. Another direction for future studies would be incorporating an editing/filtering phase to remove possible noisy data before any other process.

References

1. Blagus, R., Lusa, L.: Class prediction for high-dimensional class-imbalanced data. Bioinformatics 11(1), 523–540 (2010)
2. Bruzzone, L., Serpico, S.B.: Classification of imbalanced remote-sensing data by neural networks. Pattern Recogn. Lett. 18(11-13), 1323–1328 (1997)
3. Camps-Valls, G.: Machine learning in remote sensing data processing. In: Proc. IEEE Int'l. Workshop Machine Learning for Signal Processing, Grenoble, France, pp. 1–6 (2009)
4. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic minority oversampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
5. Chen, X., Fang, T., Huo, H., Li, D.: Semisupervised feature selection for unbalanced sample sets of VHR images. IEEE Geosci. Remote Sens. Lett. 7(4), 781–785 (2010)
6. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13(1), 21–27 (1967)
7. Ezawa, K., Singh, M., Norton, S.: Learning goal oriented bayesian networks for telecommunications risk management. In: Proc. 13th Int'. Conf. Machine Learning, pp. 139–147 (1996)
8. Fawcett, T., Provost, F.: Adaptive fraud detection. Data Min. Knowl. Disc. 1(3), 291–316 (1997)
9. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evol. Comput. 17(3), 275–306 (2009)
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: an update. SIGKDD Explor. Newslett. 11, 10–18 (2009)

11. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In: Proc. Int'l. Conf. Intelligent Computing, Hefei, China, pp. 878–887 (2005)
12. He, H., Garcia, E.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21(9), 1263–1284 (2009)
13. Hsu, P.H., Tseng, Y.H., Gong, P.: Dimension reduction of hyperspectral images for classification applications. Geogr. Inf. Sci. 8(1), 1–8 (2002)
14. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intell. Data Anal. 6(5), 429–449 (2002)
15. Kamal, A., Zhu, X., Narayanan, R.: Gene selection for microarray expression data with imbalanced sample distributions. In: Proc. Int'l. Joint Conf. Bioinformatics, Systems Biology and Intelligent Computing, Shanghai, China, pp. 3–9 (2009)
16. Kubat, M., Holte, R., Matwin, S.: Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. 30(2-3), 195–215 (1998)
17. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proc. 14th Int'l. Conf. Machine Learning, Nashville, USA, pp. 179–186 (1997)
18. Lin, L., Ravitz, G., Shyu, M.L., Chen, S.C.: Effective feature space reduction with imbalanced data for semantic concept detection. In: Proc. Int'l. Conf. Sensor Networks, Ubiquitous, and Trustworthy Computing, Taichung, Taiwan, pp. 262–269 (2008)
19. Liu, X.Y., Zhou, Z.H.: The influence of class imbalance on cost-sensitive learning: An empirical study. In: Proc. 6th Int'l. Conf. Data Mining, Hong Kong, pp. 970–974 (2006)
20. Maloof, M.: Learning when data sets are imbalanced and when costs are unequal and unknown. In: Workshop Learning from Imbalanced Data Sets II, Washington, DC (2003)
21. Martínez-Usó, A., Pla, F., Sotoca, J.M., García-Sevilla, P.: Clustering-based hyperspectral band selection using information measures. IEEE Trans. Geosci. Remote Sens. 45(12), 4158–4171 (2007)
22. Melgani, F., Bruzzone, L.: Classification of hyperspectral remote sensing images with support vector machines. IEEE Trans. Geosci. Remote Sens. 42(8), 1778–1790 (2004)
23. Richards, J., Jia, X.: Using suitable neighbors to augment the training set in hyperspectral maximum likelihood classification. IEEE Geosci. Remote Sens. Lett. 5(4), 774–777 (2008)
24. Trebar, M., Steele, N.: Application of distributed SVM architectures in classifying forest data cover types. Comput. Electron. Agr. 63(2), 119–130 (2008)
25. Van Hulse, J., Khoshgoftaar, T., Napolitano, A., Wald, R.: Feature selection with high-dimensional imbalanced data. In: IEEE Int'l. Conf. Data Mining Workshops, Miami, USA, pp. 507–514 (2009)
26. Wasikowski, M., Chen, X.W.: Combating the small sample class imbalance problem using feature selection. IEEE Trans. Knowl. Data Eng. 22(10), 1388–1400 (2010)
27. Waske, B., Benediktsson, J.A., Sveinsson, J.R.: Classifying remote sensing data with support vector machines and imbalanced training data. In: Benediktsson, J.A., Kittler, J., Roli, F. (eds.) MCS 2009. LNCS, vol. 5519, pp. 375–384. Springer, Heidelberg (2009)
28. Williams, D., Myers, V., Silvious, M.: Mine classification with imbalanced data. IEEE Geosci. Remote Sens. Lett. 6(3), 528–532 (2009)
29. Zhang, J., Mani, I.: kNN approach to unbalanced data distributions: a case study involving information extraction. In: Proc. Workshop Learning from Imbalanced Datasets, Washington DC (2003)
30. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18(1), 63–77 (2006)

Active Learning for Dialogue Act Labelling

Fabrizio Ghigi[1], Vicent Tamarit[2], Carlos-D. Martínez-Hinarejos[2], and José-Miguel Benedí[2]

[1] Dpto Electricidad y Electrónica, Universidad de Ciencia y tecnologia, Universidad del País Vasco, Sarriena s/n, 48940, Leioa, Spain
fabrizio.ghigi@gmail.com
[2] Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022, Valencia, Spain
{vtamarit,cmartine,jbenedi}@iti.upv.es

Abstract. Active learning is a useful technique that allows for a considerably reduction of the amount of data we need to manually label in order to reach a good performance of a statistical model. In order to apply active learning to a particular task we need to previously define an effective selection criteria, that picks out the most informative samples at each iteration of active learning process. This is still an open problem that we are going to face in this work, in the task of dialogue annotation at dialogue act level. We present two different criteria, weighted number of hypothesis and entropy, that we have applied to the Sample Selection Algorithm for the task of dialogue act labelling, that retrieved appreciably improvements in our experimental approach.

1 Introduction

Dialogue systems are an important application in the field of Natural Language Processing. A dialogue system is usually defined as a computer system that interacts with a human by using dialogue to achieve a defined objective [Dybkjær and Minker, 2008]. The computer system interprets the user input in the form of dialogue meaningful units, which are usually known as Dialogue Acts (DA) [Bunt, 1994], and that are used by the system to determine its reaction to user input (this reaction can be coded in DA labels as well). The reaction of the system is defined by the dialogue strategy, which indicates what actions the system must perform, including the response generation to the user. These strategies can be rule-based strategies [Gorin et al., 1997] (based on a set of predefined rules) or data-based strategies [Young, 2000] (based on statistical models). In any case, these strategies are based on the study of dialogues of the task to be fulfiled, and in their annotation in terms of DA.

The goal of this work is to explore various sample selection criteria and employ them in an active learning strategy framework for dialogue annotation. The results prove that we can achieve a good annotation model performance by using only a subset of the initial set of samples (the most effective data samples), reducing the effort needed to label the dialogues that will be used to train the final

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 652–659, 2011.

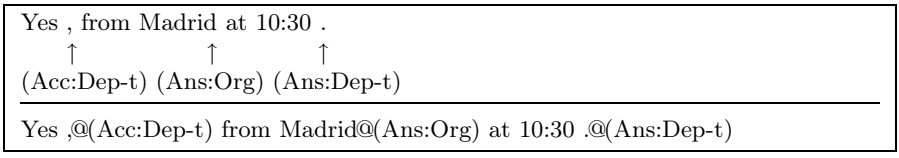


Fig. 1. An alignment between a dialogue turn and its corresponding DA labels (from the DIHANA task) and the result of the re-labelling process, where @ is the attaching metasymbol.

dialogue model. The automatic annotation method used in this work is the N-Gram Transducer (NGT) annotation model, described in [Tamarit et al., 2009]. We report experiments to find a good selection criterion for the Active Learning Algorithm [Hwa, 2000] for the task of automated DA labelling of the DIHANA corpus [Benedí et al., 2006].

This document is organised as follows: In Section 2, the statistical model for labelling the unsegmented dialogue turns is presented. In Section 3, active learning strategy is introduced. In Section 4, the selection criteria choosen are presented. In Section 5, the experimental setting used to test the learning criteria and the obtained results are detailed. In Section 6, final conclusions and future work are presented.

2 The NGT Annotation Model

The dialogue annotation problem can be presented as, given a word sequence $\mathcal{W}$ that represents a dialogue, obtain the sequence of DA $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. This probability can be modelled by a Hidden Markov Model approach using the Bayes rule [Stolcke et al., 2000] or directly modeling the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$.

The NGT model directly estimates the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$ by means of an n-gram model which acts as a transducer. The definition of this model is based on a Stochastic Finite-State Transducer (SFST) inference technique known as GIATI[1] [Casacuberta et al., 2005]. GIATI starts from a corpus of aligned pairs of input-output sequences. These alignments are used in a re-labelling process that produces a corpus of extended words as a result of a combination of the words of the input and output sentences. This corpus is used to infer a grammatical model (usually a smoothed n-gram).

In the case of dialogues, the input language is the sequence of words of the dialogue, the output language is the sequence of DA of the dialogue, and the alignment is between the last word of the segment and the corresponding DA. Thus, for each turn $w_1 w_2 \dots w_l$ and its associated DA sequence $u_1 u_2 \dots u_r$, the re-labelling step attaches the DA label to the last word of the segment using

[1] GIATI is the acronym for Grammatical Inference and Alignments for Transducer Inference.

a metasymbol (@), providing the extended word sequence $e_1 e_2 \dots e_l$, where: $e_i = w_i$ when w_i is not aligned to any DA, $e_i = w_i @ u_k$ when w_i is aligned to the DA u_k. Figure 1 presents an example of alignment for a dialogue turn and the corresponding extended word sequence. After the re-labelling process, a grammatical model is inferred. The usual option is a smoothed n-gram.

In the case of dialogues, the alignments between the words in the turn and the corresponding DA labels are monotonic (no cross-inverted alignments are possible). Consequently, no conversion to SFST is necessary to efficiently apply a search algorithm on the n-gram, since for each input word we can decide whether to emit or not a DA label without referring to posterior words. Therefore, this n-gram acts as a transducer and gives the name to the technique (NGT: N-Gram Transducers) [Martínez-Hinarejos et al., 2009].

The decoding in the NGT model is a Viterbi search which forms a search tree. The i-th level of the tree corresponds to the i-th input word in the sequence. Each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. The probability of each branch is updated according to the corresponding parent node, the n-gram probability of the corresponding extended word sequence and the n-gram probability of the corresponding DA sequence (in case a new DA is produced).

In the final step, the search on the NGT model produces a search tree where each leaf node represents a possible solution (an annotation hypothesis) to the annotation problem for the input word sequence (a dialogue). Each leaf node has associated a probability calculated by the method described above, and the leaf node with highest probability is taken as the optimal solution for the annotation problem. The solution is obtained by going up from the leaf node till the root node of the constructed tree, giving an annotation and a segmentation on the dialogue.

3 Active Learning

Active learning selects more data at each iteration of the learning process from the unlabeled set by asking someone to manually label that data. The algorithm stops when no more data or no more human resources are available, or a sufficient performance is reached.

In order to apply the active learning algorithm, a criterion that allows our system to assign a “priority" to each sample in the unlabeled set data is needed; then we can use the given scores to sort the set of unlabeled data, and choose a subset with higher priority (according to the selected criterion). The selected samples are manually labelled and they are used to reestimate the model parameters. If the accuracy goal is not overtaken, the reestimated model is used in the next step of sample selection. Otherwise, the process is finished.

In our implementation of Active Learning Algorithm [Hwa, 2000], U is a set of unlabeled candidates; L is a small set of labeled training samples; M is the current model.

Initialize
  $M \leftarrow Train(L)$
Repeat
  $N \leftarrow Select(n,\ U,\ M,\ f)$
  $U \leftarrow U\text{-}N$
  $L = L \cup Label(N)$
  $M{=}Train(L)$
Until $(M{=}M_{true})$ or $(U{=}\varnothing)$ or (Human Stops)

4 Sample Selection Criteria

In our case, the training process of the model is the usual training for the NGT model, and the labelling process in the human annotation of the dialogues. Consequently the key point of the Active Learning Algorithm presented in Section 3 is the sample selection criterion. Depending on the task, various criteria could be used. In this work we tested the algorithm with two different criteria: Weighted Number of Hypothesis and entropy. Both criteria are based on the idea that the more significant samples that we can add to the training set are those samples that are more difficult to assign a correct label. These "difficult" samples can be measured by the "uncertainty" in finding a correct label for the sample.

4.1 Weighted Number of Hypothesis

The first criterion is the number of hypothesis retrieved by the NGT decoding. Each hypothesis gets a weight, that depends on the feasibility of the hypothesis: the most probable hypothesis have more weight on the final decision, while the less probable hypothesis not strongly affect our uncertainty. We use for each sample the following equation:

$$\sum_i \frac{\Pr_i(x)}{\Pr_{max}(x)} \tag{1}$$

where $\Pr_i$ represents probability of i-th hypothesis obtained by the decoding of sample x (in our case a possible decodification of the current dialogue in DA) with the current model, and $\Pr_{max}(x)$ is the maximum probability among all hypothesis of the current sample. When this value is computed for each unlabeled sample, we select the dialogues with highest scores of "uncertainty". We decide to assign this value to each hypothesis because not every hypothesis retrieved by the model adds the same uncertainty: hypothesis with higher probability get a weight close to 1, while hypothesis less probable get less weight.

4.2 Entropy

The second criterion used is that of *Entropy*. The *Entropy* is a common way in language processing of evaluating language models. It measures how difficult is for the model to recognize a specific sample: the smaller the entropy, the easier

for the model to decode correctly the sample. The entropy for a dialogue is computed according to the following expression [Robinson, 2008]:

$$H_m(t) = -\frac{1}{\Pr_m(s)} \left(\sum_{t \in T} \Pr_m(t) \log \Pr_m(t) \right) + \log \Pr_m(s) \tag{2}$$

where $\Pr_m(s)$ is the word sequence probability by the model M(in our case, given by a n-gram of words), $\Pr_m(t)$ is the probability of the decodification retrieved by the model M(i.e., the probability given by the NGT model), and T is the dialogue set.

To have homogenous values, the computed value of *Entropy* is normalized by the lenght (number of words) of the current sample, because the entropy value is influenced by the length of the sample. Like previous criterion, *Entropy* gives us an indication of how much we know about the current sample, i.e., an uncertainty level.

5 Experiments

Experiments are developed for the dialogue act annotation task.The automatic annotation method used in this work is the NGT model.The learning criteria described in Section 4 are tested on the Dihana corpus that will be described in Section 5.1. In order to evaluate results we use DAER and SegDAER metrics. DAER is the average edit distance between the reference DA sequences of the turns and the DA sequences assigned by the labelling model. SegDAER is an average edit distance between sequences derived from the reference and the annotation result; in this case, sequences are a combination of the DA label and its position (segmentation). Incremental selection of training samples is lead by the Active Learning Algorithm described in Section 3.

5.1 Dihana Corpus

The Dihana corpus [Benedí et al., 2006] is a set of spoken dialogues in Spanish language, between a human and a simulated machine, acquired with the Wizard of Oz (WoZ) technique. It is restricted at the semantic level (dialogues are related to the task of obtaining information about train tickets), but natural language is allowed (there are no lexical or syntactical restrictions). The Dihana corpus is composed of 900 dialogues about a telephone train information system. It was acquired from 225 different speakers (153 male and 72 females), with small dialectal variants. There are 6,280 user turns and 9,133 system turns. The vocabulary size is 823 words. The total amount of speech signal is about five and a half hours. The annotation scheme used in the corpus is based on the Interchange Format (IF) defined in the C-STAR project [Lavie et al., 1997], which was adapted to dialogue annotation. Details on the annotation process are available in [Alcácer et al., 2005].

5.2 Experiment Strategy

We have used the same partition (720 dialogue to pick up for training, 180 dialogues for test) of the Dihana corpus for every experiment developed, maintaining the following strategy:

1. Start the experiment with a small training set, picked out by a general criteria (in fact we picked out the two largest dialogues).
2. Train a model with this small training set, and verify the accuracy of the system, calculating DAER and SegDAER for the NGT model predictions.
3. With current model compute score for each remaining dialogue in unlabeled set, using criteria *Weighted Number of Hypothesis*, equation (1), or *Entropy*, equation (2).
4. Select a subset of the remaining dialogues with higher scores.
5. Include the selected dialogues in the training samples.
6. Return to step 2.

The Sample Selection Algorithm described in Section 3 is used to manage incremental selection of training samples.

5.3 Results

In Figure 2 we can see DAER and SegDAER trends for Random Baseline, and for the two selection criteria, *Weighted Number of Hypothesis*, equation (1), and *Entropy*, equation (2).

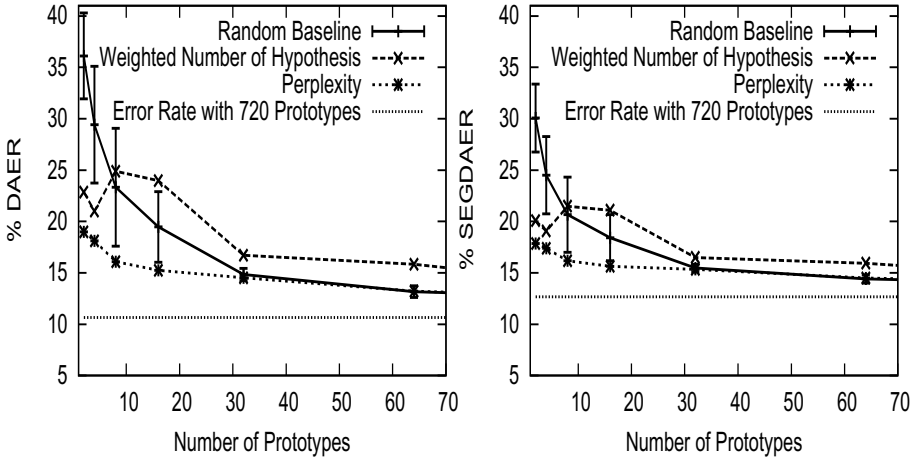


Fig. 2. Comparison among Random Baseline, *Weighted Number of Hypothesis*, equation (1), and *Entropy*, equation (2), error rate (DAER and SegDAER) behaviour while incrementing the training set size. The lowest line represent the error rate using the entire set of availables dialogues (in our case 720).

As we can clearly see from Figure 2, *Entropy* criterion is a very effective selection criterion for this task, it has better performance than Random Baseline until the asymptote is reached. Moreover, performance is close to that obtained with the whole training set. We tried more experiments incrementing training set size by one dialogue at each iteration, but this does not change significantly the error rate trend.

6 Conclusions and Future Work

In this document we have shown results of applying Active Learning to a dialogue act labelling task. We have seen that choosing a well founded criterion (*Entropy*) to implement Active Learning Algorithm, significant performance boost can be achieved. In the experiments developed *Entropy* criterion obtained really good results, while *Weighted Number of Hypothesis* criterion had a variable behaviour, although more experiments should be perform in the future to confirm its properties.

Future work contemplates the application of presented criteria against other corpora (such as SwitchBoard) to confirm goodness of criteria, the parallelization of the Active Learning Algorithm to speed up the selection process, the exploration of other selection criteria, the application of this work in an interactive framework and the analysis of the error rate for each single dialogue act label taking into account its frecuency in the corpus.

Acknowledgments. Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), MITTRAL (TIN2009-14633-C03-01) projects and the FPI scholarship (BES-2009-028965). Also supported by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067.

References

Alcácer et al., 2005. Alcácer, N., Benedí, J.M., Blat, F., Granell, R., Martínez, C.D., Torres, F.: Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In: SPECOM, Greece, pp. 583–586 (2005)

Benedí et al., 2006. Benedí, J.M., Lleida, E., Varona, A., Castro, M.J., Galiano, I., Justo, R., López, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: DIHANA. In: Fifth LREC, Genova, Italy, pp. 1636–1639 (2006)

Bunt, 1994. Bunt, H.: Context and dialogue control. THINK Quarterly 3 (1994)

Casacuberta et al., 2005. Casacuberta, F., Vidal, E., Picó, D.: Inference of finite-state transducers from regular languages. Pat. Recognition 38(9), 1431–1443 (2005)

Dybkjær and Minker, 2008. Dybkjær, L., Minker, W. (eds.): Recent Trends in Discourse and Dialogue. Text, Speech and Language Technology, vol. 39. Springer, Dordrecht (2008)

Gorin et al., 1997. Gorin, A., Riccardi, G., Wright, J.: How may I help you? Speech Comm. 23, 113–127 (1997)

Hwa, 2000. Hwa, R.: Sample selection for statistical grammar induction. In: Proceedings of the 2000 Joint SIGDAT, pp. 45–52. Association for Computational Linguistics, Morristown (2000)

Lavie et al., 1997. Lavie, A., Levin, L., Zhan, P., Taboada, M., Gates, D., Lapata, M.M., Clark, C., Broadhead, M., Waibel, A.: Expanding the domain of a multi-lingual speech-to-speech translation system. In: Proceedings of the Workshop on Spoken Language Translation, ACL/EACL 1997 (1997)

Martínez-Hinarejos et al., 2009. Martínez-Hinarejos, C.D., Tamarit, V., Benedí, J.M.: Improving unsegmented dialogue turns annotation with N-gram transducers. In: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23), vol. 1, pp. 345–354 (2009)

Robinson, 2008. Robinson, D.W.: Entropy and uncertainty, vol. 10, pp. 493–506 (2008)

Stolcke et al., 2000. Stolcke, A., Coccaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M.: Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational Linguistics 26(3), 1–34 (2000)

Tamarit et al., 2009. Tamarit, V., Benedí, J., Martínez-Hinarejos, C.: Estimating the number of segments for improving dialogue act labelling. In: Proceedings of the First International Workshop of Spoken Dialog Systems Technology (2009)

Young, 2000. Young, S.: Probabilistic methods in spoken dialogue systems. Philosophical Trans. Royal Society (Series A) 358(1769), 1389–1402 (2000)

Multi-class Probabilistic Atlas-Based Segmentation Method in Breast MRI

Albert Gubern-Mérida[1], Michiel Kallenberg[2], Robert Martí[1], and Nico Karssemeijer[2]

[1] University of Girona, Spain
{agubern,marly}@eia.udg.edu
[2] Radboud University Nijmegen Medical Centre, The Netherlands
{n.karssemeijer,m.kallenberg}@rad.umcn.nl

Abstract. Organ localization is an important topic in medical imaging in aid of cancer treatment and diagnosis. An example are the pharmacokinetic model calibration methods based on a reference tissue, where a pectoral muscle delineation in breast MRI is needed to detect malignancy signs. Atlas-based segmentation has been proven to be powerful in brain MRI. This is the first attempt to apply an atlas-based approach to segment breast in T1 weighted MR images. The atlas consists of 5 structures (fatty and dense tissues, heart, lungs and pectoral muscle). It has been used in a Bayesian segmentation framework to delineate the mentioned structures. Global and local registration have been compared, where global registration showed the best results in terms of accuracy and speed. Overall, a Dice Similarity Coefficient value of 0.8 has been obtained which shows the validity of our approach to Breast MRI segmentation.

Keywords: breast MRI segmentation, probabilistic atlas, Bayesian framework, Markov Random Field.

1 Introduction

Atlas-based segmentation is a powerful generic technique for automatic delineation of objects in volumetric images that can take into account neighbourhood relationships between several different structures. Many atlas-based segmentation methods have been proposed in the literature for 3D medical imaging applications especially applied to segment MR brain images ([1,14] and references therein). However, a variety of these algorithms have also been used to different 3D image modalities and different body structures such as prostate MRI [5,8], abdominal CT [9], chest CT [10] and head and neck CT [3].

Neither previous work in atlas-based segmentation for the delineation of breast structures nor public breast MRI probabilistic atlases are present in the literature. In this work we have developed a multi-class probabilistic atlas-based segmentation method for breast MRI. The method segments the pectoral muscle, fatty and dense breast tissues, the heart and lungs. Segmentation of these structures is useful for cancer diagnosis or treatment applications where organ localization is needed. More specifically, our proposal aims to segment the pectoral

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 660–667, 2011.

muscle to use it as a reference tissue in pharmacokinetic model calibration [7,15]. Note that in this work a "probabilistic atlas" refers to the pair of an anatomical image and a probability tissue distribution volume. The former defines the atlas reference space, while the latter provides a volume with the complete spatial distribution of probabilities that a voxel belongs to one or more organs. Next sections describe the material employed (Sec. 2), the corrections applied to the volumes (Sec. 3.1), how the probabilistic atlas was built (Sec. 3.2) and how it is incorporated in a Bayesian voxel classification framework providing very rich information to find the organ location (Sec. 3.3). Qualitative and quantitative results are shown in Sec. 4 and finally conclusions are given discussed in Sec. 5.

2 Material

The data set used to construct the atlas and evaluate the segmentation results consists of 9 breast T1 weighted MR scans obtained from clinical data. Breast MRI examinations were performed on a 1.5 T system (Siemens 1.5T, Magnetom Vision), with a dedicated breast coil (CP Breast Array, Siemens, Erlangen). A dynamic contrast enhanced T1-weighted Flash-3D sequence was used, with repetition time of 8.1 ms, an echo time of 4 ms, and a flip angle of 20 degrees. The pixel spacing was 1.25 mm x 1.25 mm, and the slice thickness 1.5 mm. Per series, 108 slices were acquired, without interslice gap. Patients were scanned in prone position.

The pre-contrast series were used for the segmentation and each MR volume was manually segmented by an expert into 7 classes: background, fatty tissue, glandular tissue, pectoral muscle, lung area and the heart. The seventh class is the "non-of-above" class. Fig. 1 shows an example of a MRI slice on an axial view and the manual delineation of the mentioned classes.

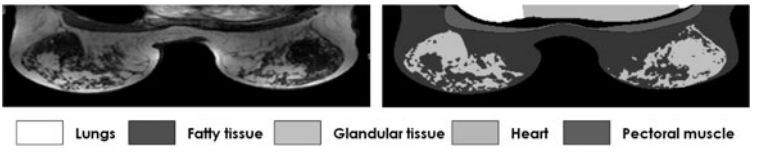


Fig. 1. MR scan on an axial slice of a clinical breast MR T1 weighted volume with the manual annotation of the different structures

3 Methodology

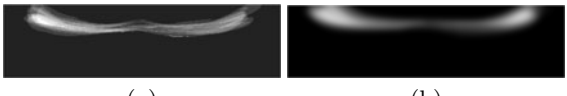
This section describes the parts that compose the methodology. Firstly, a brief description of the preprocessing algorithms applied to the data set is given in Sec. 3.1. Secondly, Sec. 3.2 describes the method developed for building a breast probabilistic atlas. Then, adopting the classification made by Rohlfing et al. [11], Sec. 3.3 shows the Average Shape Atlas segmentation approach implemented in this work, which uses atlas information on a Bayesian framework.

3.1 Data Preprocessing

Because of the inhomogeneity of the breast coil sensitivity, intensity values are corrupted. Signal intensity homogeneity is required because image artifacts can considerably affect registration and segmentation results. For this reason the first step of the methodology consisted in correcting variability between images and inhomogeneities. Hence, image normalization algorithm was developed and applied to each scan in order to compensate inter-patient signal intensity variability. In addition, Non-parametric intensity Non-uniformity Normalization (N3) [13] bias field correction method was also employed to each scan.

3.2 Construction of the Atlas

In this work, for each patient segmentation, a full probabilistic atlas was built with the 8 remaining patients following a leave-one-out evaluation strategy. Firstly, the 8 patients and their segmentations were mapped onto the same reference space and the probabilistic atlas was created computing the frequency with which each location was labelled as a specific organ. A common reference space was used for all tests by selecting an extra patient that was not included in the evaluation data set. This extra image became the anatomical image of the atlas. Secondly, the final smooth probabilistic atlas was obtained using a 3D Gaussian convolution, basically to compensate for the small number of cases and the local registration errors. Variance σ^2 of 50 and 20 mm were used for global and local registration respectively. Fig. 2 shows an example of a pectoral probabilistic atlas before (a) and after (b) smoothing.



(a) (b)

Fig. 2. Pectoral probabilistic atlas example before (a) and after (b) smoothing

For the first registration stage we employed two warping transforms to compensate for inter-patient differences. The first transform we evaluated was affine registration focused on a Volume of Interest (VOI). Although we were aware that such transform does not offer enough degrees of freedom (DOF) to compensate for the large differences that are present in this type of images (see Fig. 3), we observed that we could globally align pectoral muscles (which was the main goal of the segmentation) by aligning thoracic areas. For this reason, a VOI in each volume was defined by manually selecting the thoracic area with a single annotation point. Such approach was selected having in mind to obtain a good execution time to make the method suitable for clinical use.

The second evaluated transform was a non-rigid registration based on B-Splines proposed by Rueckert et al. [12]. This algorithm has been vastly applied by atlas-based segmentation methods [1,5,10,11] to minimize inter-individual

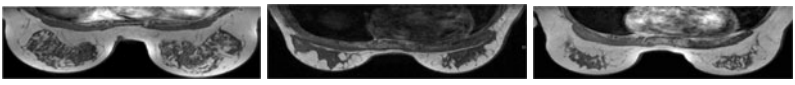


Fig. 3. Three breast MRI axial slices from three different subjects: variation between structures is easily observed in these three examples

variability in the shapes of anatomical structures and it is usually initialized by a global registration method like the affine registration described earlier.

Both approaches maximize the similarity measure of Mutual Information (MI) in an multi-resolution scheme using a stochastic gradient descent optimizer. For non-rigid registration, B-spline grid spacing of 32, 16, 8, 4 and 2 mm was used for each of the 5 resolutions respectively. Elastix [6] was used for the implementation.

3.3 Segmentation

As we mentioned previously, the atlas is used in a Bayesian framework in order to segment the different structures of the breast. The approach is based on the work of Park et al. [9] and Karssemeijer [4], who segmented abdominal structures in computed tomography (CT) and three-dimensional X-ray images respectively. Figure 4 shows a schema of the Bayesian voxel classification algorithm incorporating the use of the probabilistic atlas.

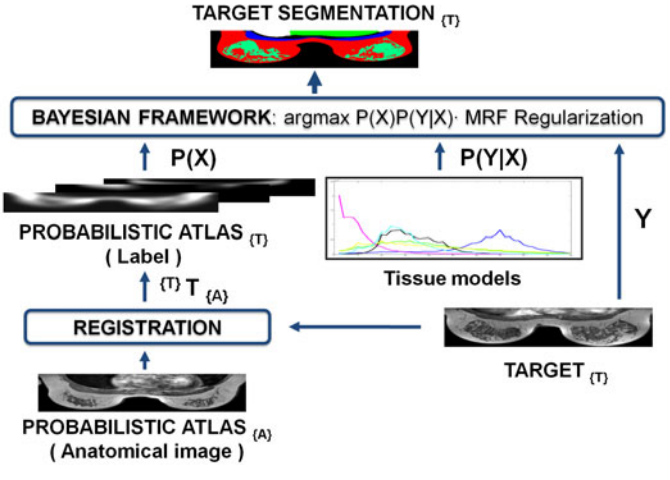


Fig. 4. Voxel classification algorithm overview: from bottom to top, the labels of the probabilistic atlas are mapped onto target image space {T} using the anatomical image of the atlas. The probabilistic atlas, the tissue models and the target are provided to the Bayesian framework as a prior probability $P(X)$, conditional probability $P(Y|X)$ and set of intensity values Y, respectively. The Bayesian framework estimates the segmentation X that maximizes $P(X)P(Y|X)$.

In this paper the true label (the segmentation or set of labels) is denoted by X and the target image (data set of intensity values) is denoted by Y. Elements of X and Y are arranged by a spatial position denoted by $i \in I$, where I is the simple index (x, y, z) in a 3D rectangular grid. Sample realizations of X and Y are represented throughout this work as $\mathrm{x} = (x_1, x_2, ..., x_N)$ and $\mathrm{y} = (y_1, y_2, ..., y_N)$, respectively, where N is the total number of voxels. Sample space of X is denoted Ω_x where $\Omega_x = \{\mathrm{x} : x_i \in \{1, 2, 3, 4, 5, 6, 7\}, \forall i \in I\}$. Labels 1, 2, 3, 4, 5, 6 and 7 are background, fatty tissue, glandular or dense tissue, heart, lungs, pectoral muscle and "None of the above" label respectively.

The problem consists in estimating the label X that best explains the given observation Y according to some cost function. As a decision rule, MAP (maximum a posteriori) was chosen: segmentation X was estimated by maximizing the global a posteriori probability $P(X|Y)$ by searching the most probable labeling given the image Y and some prior model. Using Bayes theorem, the posterior probability to be maximized can be written as $P(Y|X)P(X)$. The probability distribution $P(Y|X)$ of the image Y, given a particular segmentation X, can be estimated from training data or from the image at hand. For example, Park et al. modelled probabilities of y_i's as conditional Gaussians given mean and variance of the true label X. In this work $P(Y|X)$ was specified by signal intensity tissue models directly built from the scans and manual segmentations of the data set. For each structure, a histogram of intensity values was built considering the voxels of the MRI volumes which belong to it using the manual segmentations.

On the other hand, the probability distribution $P(X)$ is given by the probabilistic atlas once it has been mapped onto the target space using the same registration procedure used in its construction. A Markov Random Field (MRF) regularization is included to smooth the segmentation taking into account neighbourhood information. It introduces the probability of finding a particular label at i that depends only on the labels of voxels close to i (26 nearest neighbours in this work). Considering the addition of the MRF regularization, the posteriori probability, optimized by Iterated Conditional Mode (ICM), is defined as follows

$$\arg\max P(y_i|x_i = k)\, P(x_i = k) \exp\left(\sum_{n=1}^{K} B(k, n) g_i(n)\right), \tag{1}$$

where $g_i(n)$ denotes the number of neighbours labelled as n, K is the number of classes and the interaction parameters $B(k, n)$ determines if regions labelled k and n are likely to neighbour each other.

4 Results

In a leave-one-out experiment we evaluated both registration methods by comparing the segmentation results. The quality of the segmentation was measured by determining the similarity with the ground truth. As a performance measure the Dice Similarity Coefficient (DSC) was calculated. Figure 5 shows a box plot

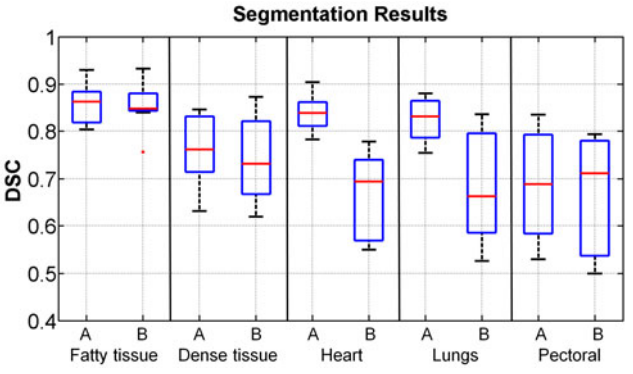


Fig. 5. Box plot with segmentation DSC values for each organ using VOI affine (A) and B-splines (B) registrations

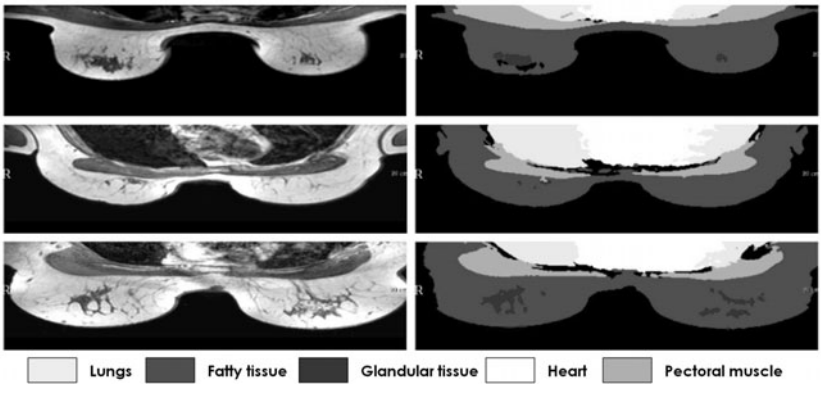


Fig. 6. Intermediate slices from 3 different patients and their segmentation using affine registration with VOI

with DSC values for the patient segmentations of each organ using affine registration (A) or B-Splines (B). For lungs and heart, VOI affine registration clearly outperforms B-Splines because non-rigid transform seems to introduce errors (p-value < 0.05, two-sided paired t-test). Segmentations of fatty and dense tissues and pectoral muscle provides similar values with no significant differences for both registration methods (p-values of 0.66, 0.43 and 0.15 respectively for a two-sided paired t-test). To sum up, mean DSC values were 0.8 for affine registration and 0.72 for B-Splines. The execution times of affine and B-Splines, measured on Intel(R) Core(TM)2 Quad CPU Q9550 2.83GHz were 15 min. to 1 hour aprox. respectively.

Overall, segmentation results can be considered satisfactory in all cases (normally a DSC > 0.7 is considered good segmentation [2]), which is illustrated by Fig. 6, where intermediate slices and their segmentations from 3 different patients are shown.

5 Discussion and Conclusions

In this work we have presented a framework for the segmentation of breast structures based on atlas. To the best of our knowledge this is the first proposal using atlas-based methodology for breast segmentation. Firstly we have constructed a probabilistic atlas by registering 8 patient data sets onto a single patient. Thereupon, we have integrated it into a Bayesian framework with MRF regularization for segmentation of breast MRI structures. Affine registration focused on VOI and B-Splines registration algorithms have been evaluated. The former has presented satisfactory results (general DSC average of 0.8) and acceptable execution time to be suitable for routine clinical use in the future.

For the pectoral muscle segmentation, which is the organ of main interest in this paper, the mean of DSC values is approximately 0.7 and good delineations have been obtained in intermediate slices. Even though the DSC value is not as good as for the other structures, this result is encouraging, especially considering that the pectoral muscle is a small structure compared with the others segmented and presents high inter-patient variability. We believe that the obtained segmentations are suitable as a reference tissue in pharmacokinetic model calibration, where a specific percentage of voxel well labelled is needed.

Further research will be focused on solving the problems we found that affect the segmentation results. Firstly, we were aware of the low number of cases for the evaluation, strictly related with the difficulty of acquiring annotations for the 7 different classes in large volumes (108 slices per volume). However, evaluation will be extended to more cases by increasing the data set. Secondly, although bias field on images was corrected for, they still presented inhomogeneities. Thus, other correction methods will be studied. In addition, B-Splines registration has not provided good deformation alignment and has introduced additional registration errors. This could be explained by the fact that inter-patient variability is larger than the one which could be minimized by the implemented local registration. Other fast non-rigid registration algorithms, which incorporate user intervention such as control points or adding local deformation constraints, will be studied. Finally, signal intensity has not enough discriminative power to separate the pectoral muscle and glandular tissue classes. The use of other features, such as the ones obtained from Dynamic Contrast Enhanced breast MRI, and other voxel classification methods that use atlas information will also be considered.

References

1. Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., Rueckert, D.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46(3), 726–738 (2009)
2. Bartko, J.: Measurement and reliability: statistical thinking considerations. Schizophr. Bull. 17(3), 483–489 (1991)

3. Han, X., Hoogeman, M.S., Levendag, P.C., Hibbard, L.S., Teguh, D.N., Voet, P., Cowen, A.C., Wolf, T.K.: Atlas-based auto-segmentation of head and neck ct images. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5241, pp. 434–441. Springer, Heidelberg (2008)
4. Karssemeijer, N.: A statistical method for automatic labeling of tissues in medical images. Mach. Vision Appl. 3(2), 75–86 (1990)
5. Klein, S., van der Heide, U., Lips, I., van Vulpen, M., Staring, M., Pluim, J.P.W.: Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. Medical Physics 35(4), 1407–1417 (2008)
6. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: Elastix: a toolbox for intensity based medical image registration. IEEE Transactions on Medical Imaging 29(1), 196–205 (2010)
7. Kovar, D.A., Lewis, M., Karczmar, G.S.: A new method for imaging perfusion and contrast extraction fraction: input functions derived from reference tissues. J. Magn. Reson. Imaging 8(5), 1126–1134 (1998)
8. Martin, S., Troccaz, J., Daanenc, V.: Automated segmentation of the prostate in 3d mr images using a probabilistic atlas and a spatially constrained deformable model. Med. Phys. 37(4), 1579–1590 (2010)
9. Park, H., Bland, P.H., Meyer, C.R.: Construction of an abdominal probabilistic atlas and its application in segmentation. IEEE Trans. Med. Imaging 22(4), 483–492 (2003)
10. van Rikxoort, E., Isgum, I., Arzhaeva, Y., Staring, M., Klein, S., Viergever, M.A., Pluim, J.P.W., van Ginneken, B.: Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. Medical Image Analysis 14, 39–49 (2010)
11. Rohlfing, T., Brandt, R., Menzel, R., Russakoff, D.B., Maurer Jr., C.R.: Quo vadis, atlas-based segmentation? In: Suri, J., Wilson, D.L., Laxminarayan, S. (eds.) The Handbook of Medical Image Analysis. Registration Models, vol. III, ch. 11, pp. 435–486. Kluwer Academic / Plenum Publishers, New York (2005)
12. Rueckert, D., Hayes, C., Studholme, C., Summers, P., Leach, M., Hawkes, D.J.: Non-rigid registration of breast mr images using mutual information. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 1144–1152. Springer, Heidelberg (1998)
13. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
14. Wu, M., Rosano, C., Lopez-Garcia, P., Carter, C.S., Aizenstein, H.J.: Optimum template selection for atlas-based segmentation. NeuroImage 34(4), 1612–1618 (2007)
15. Yankeelov, T.E., Luci, J.J., Lepage, M., Li, R., Debusk, L., Lin, P.C., Price, R.R., Gore, J.C.: Quantitative pharmacokinetic analysis of dce-mri data without an arterial input function: a reference region model. Magn. Reson. Imaging 23(4), 519–529 (2005)

Impact of the Approaches Involved on Word-Graph Derivation from the ASR System*

Raquel Justo, Alicia Pérez, and M. Inés Torres

University of the Basque Country,
Sarriena s/n. 48940. Leioa. Spain
{raquel.justo,alicia.perez,manes.torres}@ehu.es

Abstract. Finding the most likely sequence of symbols given a sequence of observations is a classical pattern recognition problem. This problem is frequently approached by means of the Viterbi algorithm, which aims at finding the most likely sequence of states within a trellis given a sequence of observations. Viterbi algorithm is widely used within the automatic speech recognition (ASR) framework to find the expected sequence of words given the acoustic utterance in spite of providing a suboptimal result. Word-graphs (WGs) are also frequently provided as the ASR output as a means of obtaining alternative hypotheses, hopefully more accurate than the one provided by the Viterbi algorithm. The trouble is that WGs can grow up in a very computationally inefficient manner. The aim of this work is to fully describe a specific method, computationally affordable, for getting a WG given the input utterance. The paper focuses specifically on the underlying approaches and their influence on both the spatial cost and the performance.

Keywords: Lattice, word-graphs, automatic speech recognition.

1 Introduction

Statistical decision theory is applied in a wide variety of problems within pattern recognition framework that aim at minimising the probability of erroneous classifications. The maximisation of the posterior probability $P(\bar{w}|\bar{x})$ allows to get the most likely sequence of symbols $\bar{w}$, that matches a given sequence of input observations, $\bar{x}$, as shown in eq. (1).

$$\hat{\bar{w}} = \arg\max_{\bar{w}} [P(\bar{w}|\bar{x})] \tag{1}$$

In many pattern recognition tasks, such as computer vision or automatic speech recognition (ASR), the knowledge source involved in the optimisation problem can be represented by stochastic finite-state models. Thus, the search problem is

* This work has been partially funded by the Spanish Ministry of Science and Innovation under the Consolider Ingenio 2010 programme (MIPRCV CSD2007-00018) and SD-TEAM project (TIN2008-06856-C05-01); and by the Basque Government (under grant GIC10/158 IT375-10).

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 668–675, 2011.

formulated as a decoding problem trough a finite-state network. In this context, the sequence of symbols that most probably causes the sequence of observations should be decoded considering all possible paths in the network that match the given input (denoted by $\mathcal{D}_{\bar{x}}$).

$$\hat{\bar{w}} = \arg\max_{\bar{w}} \left[\sum_{\bar{d} \in \mathcal{D}_{\bar{x}}} P(\bar{w}, \bar{d}|\bar{x}) \right] \approx \arg\max_{\bar{w}} \left[\max_{\bar{d} \in \mathcal{D}_{\bar{x}}} P(\bar{w}, \bar{d}|\bar{x}) \right] \tag{2}$$

Nevertheless, a typical approximation is to consider the sequence of symbols associated to the most probable path as in eq. (2). The Viterbi algorithm [1] is widely used to implement the maximum approximation and find the most probable path within the finite-state search space. This approach allows efficient implementation by means of dynamic programming. Nevertheless, for problems in which the sequence of symbols happen to be compatible with multiple sequences of states, this approach turns to be suboptimal. This is the case of ASR where given the acoustic representation of input signal ($\bar{x}$) the same sequence of words ($\bar{w}$) can be explained by a number of different ways within throughout search network.

All together, Viterbi-based decoding algorithm does not provide the most likely sequence of words, as it is the goal, but the word sequence associated to the most likely path in the search network. Moreover, additional assumptions made at the implementation stage also have important impact in the final sequence of words.

In order to get a better ASR performance it seems of interest not to restrict ourselves only to the hypothesis provided by the Viterbi-based implementation and take more hypotheses into account. Alternative hypotheses can be provided by means of a word-graph (WG), which can be seen as a fuzzy transcription of the acoustic utterance. WGs are an efficient means of representing a vast number of hypotheses. Numerous applications can benefit from WGs, by means of confidence measures [2,3], or extracting the list of n-best hypotheses that could be re-ordered with a different knowledge source [4].

However, the WGs grow quickly with the length of the input utterance, thus, the computational cost associated to its generation could result unfeasible. To face this, some approximations, closely related with a loose of accuracy, must be made. Since these approximations are poorly described in the literature, a standard method to compare the results obtained with WGs can be hardly adopted.

The aim of this work is to provide a full description of a method to derive a WG from the ASR system and above all, the approximations carried out to make this process computationally affordable. It is hardly ever referred in the literature the practical approaches adopted to carry out this sort of strategies, and this is precisely our aim. The paper focuses specifically on the influence of these approaches on both the spatial cost and the performance of the system.

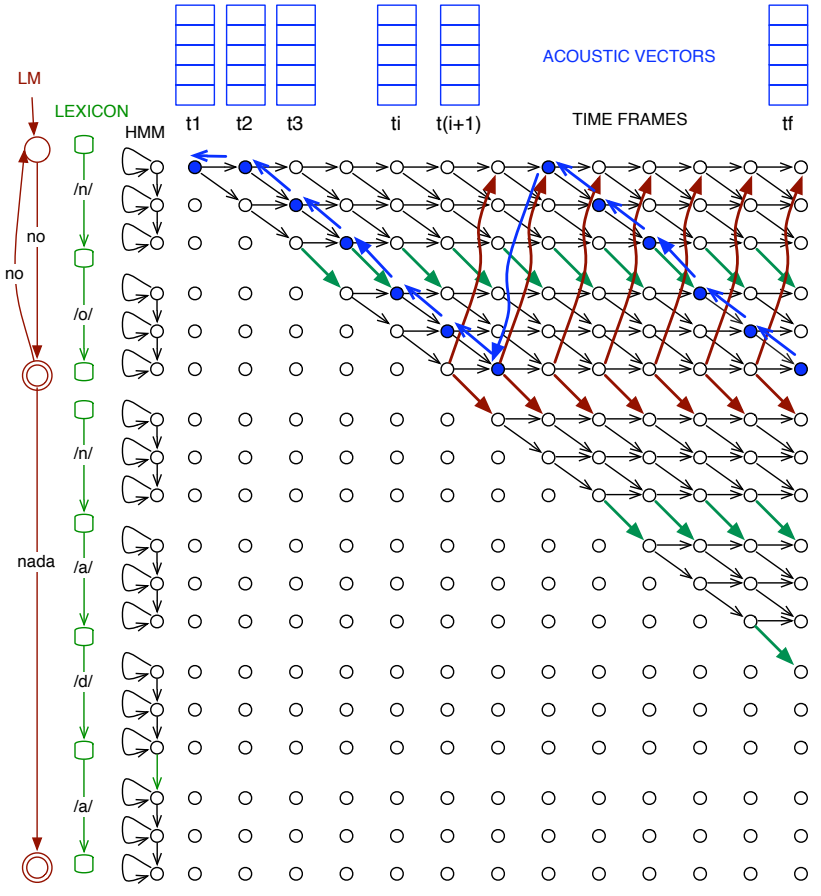


Fig. 1. Search space: cartesian product of acoustic observations and models' states

2 ASR Decoding: Viterbi Algorithm

The goal of an ASR system is to obtain the most likely word sequence given the acoustic signal uttered by the speaker. In order to do this the Bayes' decision rule is applied in eq.(1) giving as a result eq. (3).

$$\hat{\bar{w}} = \arg\max_{\bar{w}} P(\bar{w}|\bar{x}) = \arg\max_{\bar{w}} P(\bar{w})P(\bar{x}|\bar{w}) \tag{3}$$

where $P(\bar{w})$ is the probability associated by the language model (LM) to the string $\bar{w}$ and $P(\bar{x}|\bar{w})$ is the probability associated by the lexical-acoustic model. The decoding process in ASR can be represented by means of a search problem [5,6] where the search space consists of a finite-state network [7] comprising the aforementioned models (see Fig. 1).

To implement the search problem represented by eq. (2), each node in the search network is associated to the most likely path that reaches the node in a specific time-frame t. Thus, each node is linked with a predecessor node reached at time $t-1$. If we turn to Fig. 1, the blue node reached at time t_3 has two possible predecessors, but only the information related to the one associated with the most likely path will be kept, that is blue node reached at time t_2. After the forward decoding phase the final node storing the highest accumulated probability can be traced backwards predecessor by predecessor giving as a result the sequence of words associated to the most-likely path.

There are a few approximations well worthy to bear in mind when it comes to making this implementation efficient. This algorithm is typically speed up by means of a beam search. That is, at each time frame not all the nodes remain alive for subsequent search, instead, the nodes where the accumulated probability does not exceed a percentage of the higher accumulated probability are not explored any longer.

3 Approaches Involved Getting the Word-Graph

3.1 Forward Decoding

The key issue to derive a word-graph from the trellis (see Fig. 1) is to allow each node storing n predecessors instead of a single one. Note that in the implementation of Viterbi algorithm each node only would store a single predecessor (n=1), and as a result a single sequence of words is obtained.

Intuitive though this procedure might seem, there are implementation nuances that are worthy of further consideration. For instance, being strict, by allowing to all the nodes of the search space comprising the involved models, a graph of phone like units would be derived. Alternative pronunciations and time segmentations of the same word would be contained in such a graph. Note that the alternative ways in which a word could be uttered has an impact on the lexical-acoustic models but it has no relevance at word level. As a result, in order to derive a word-graph computational cost can be significantly reduced by allowing to store n predecessors only to the nodes that match with the states of the LM.

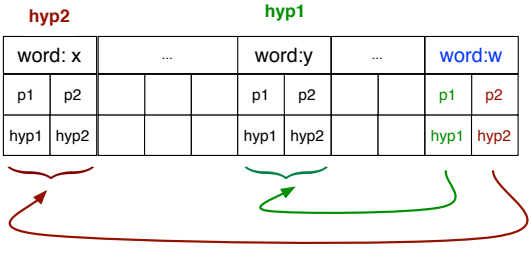


Fig. 2. Structure generated in the forward decoding phase for n=2

The alternative word-level hypotheses are stored in a structure that encompasses the list of the n-best predecessors and the probabilities associated to the corresponding paths. As an example, Fig. 2 shows the structure created in the forward decoding phase assuming n=2. Thus, when the word w is reached the information related to the best 2 paths (hyp1 and hyp2) is stored.

In an attempt to cut down the search-time the beam-search strategy is applied. At each time-frame the nodes storing paths with low probability are pruned. Note the assumption intrinsic to this node-pruning strategy. While a node is allowed to store up to n predecessors with their respective accumulated probabilities, it is the the probability of the most probable amongst the n predecessors that is associated to the node as far as pruning is regarded. As a result it might happen that a node storing a predecessor with a very high probability and the remaining predecessors with very low probabilities would remain alive, in contrast to other nodes in which the probabilities of all the predecessors are high, while not high enough compared to the most likely one in a given time-frame. On this account, beam search strategy implies an additional approach that did not occur over the standard decoding when the allowed number of predecessor stored was only one.

3.2 Backward Decoding: Getting the Word-Graph from the Lattice

Once the entire observation has been analyzed in the forward decoding phase, the n best final states are considered and traced backwards until the initial state in the lattice is reached. As a result a WG is generated conveying the following information: word label, probability of the edge, time-frame segmentation (see as an example Fig. 3).

Although the WG is an efficient means of storing a vast number of hypotheses, it grows quickly with the length of the input utterance. In order to avoid this inefficient behavior a new approach has been adopted. The idea is based on merging nodes of the WG associated to the same state in the LM, but not all of them, only those within a neighborhood of m time-frames. Being the duration of a time-frame denoted as Δt_w:

$$t_n \in t_s \pm \frac{1}{2} m \Delta t_w \Rightarrow \text{merge nodes n and s} \tag{4}$$

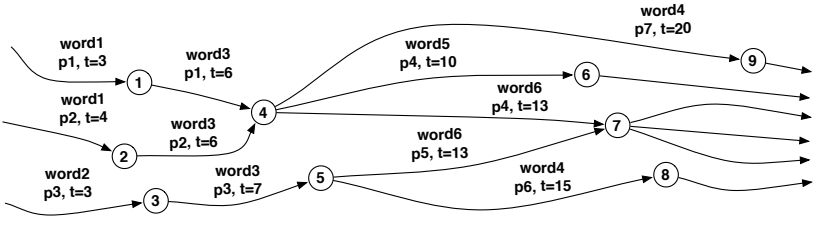


Fig. 3. Example of a part of a WG showing the word log-probability and time-frame

For example in Fig. 3, given a time frame constant $\Delta t_w = 1$, if nodes 4 and 5 are associated with the same state of the LM, they would be merged for values of $m \geq 2$. The result of merging nodes has the advantage of providing a WG of smaller size. In general, this constraint prevents us from generating fake hypotheses, but this is not ensured for particular states of the LM such as unigrams. The drawback of merging nodes of the WG is that hypotheses with fake probabilities can be generated.

On the other hand, as mentioned above, the WG is an efficient means of storing a vast number of hypotheses, while some of them might result to be redundant. Thus, in this work we have only explored a number of hypotheses below a threshold. This additional approach is also necessary to become this process computationally efficient. Specifically, a depth search has been carried out over the WG as an attempt to extract the most likely hypothesis. Nevertheless, this method leads to a number of hypotheses that share the same initial words, while different hypothesis with low probabilities in the beginning might be left aside.

4 Experimental Results

The experimental results were carried out on two different corpora:

Dihana corpus consists of 900 human-machine dialogues in Spanish regarding information about train timetables, fares, destinations and services [8]. This task has intrinsically a high level of difficulty due to the spontaneity of the speech. It comprises 5,590 different sentences to train the LM with a vocabulary of 865 words. The test set includes 1,348 spoken utterances.

MetEus corpus is a weather forecast corpus in Spanish, Basque and English [9]. In this work we focus on the Basque ASR task. The training corpus consists of 7,523 different sentences with a vocabulary of 1,135 words, and the test includes 1,798 spoken utterances.

In these tasks for each input utterance in the test a series of WGs were generated for different values of n (being "n" the number of allowed predecesors) and m (being "m" the number of time-frames merged). In the following section the influence of both parameters (n and m) are evaluated in terms of spatial cost and performance.

The influence of merging the time frames in the number of edges of the WG for Dihana (MetEus task follows the same trend) is represented in Fig. 4. The ordinate shows the number of word-graphs with the edge-range represented in the abscissa. The figure includes three series of histograms associated to a time-merging of width ranging from m=0 to m=3. It is shown that as the number of merged frames increases, the number of small WGs gets higher and the number of big WGs decreases. From these results it comes out that merging procedure allows to reduce the size of the WG.

The performance of the system, in terms of word error-rate (WER), is given in Table 1. The WGs were evaluated in a twofold manner: on the one hand, the most probable hypothesis within the WGs (denoted by "p-best"), and on

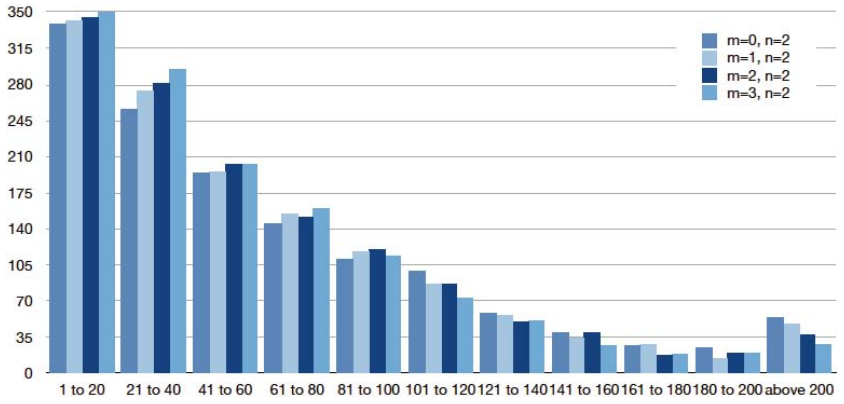


Fig. 4. Number of edges of the word-graphs obtained with n=2 and varying $0 \leq m \leq 3$

the other hand the oracle-best (denoted by "oracle"). The oracle score provides the best performance achievable considering all the hypotheses involved in the WGs. Comparing "oracle" with "p-best" results presented in Table 1, it comes out that the most probable hypothesis is not necessarily the most accurate one. Thus, the aim of using alternative hypotheses rather than the most-likely one becomes apparent.

Comparing the performance for different time-merging choices it turns out that the increment of "m" hardly degrades the performance, with the additional advantage of dealing with smaller graphs. Similarly, comparing the differences for the most likely hypothesis, it turns out that the more states we merge the more degraded the first hypothesis is. In conclusion, while the probabilities are blurred in the merging process, the most accurate hypothesis remains amongst the hypotheses provided by the graph.

If we compare the results obtained through the Viterbi-like implementation (m=0, n=1) with the "p-best" obtained with m=0 and n=2, it seems as though the WG would not help. However, this decrement in performance is due to the different approaches (described in the previous sections) devoted to make the computation affordable and not by the scope of WG itself. Moreover, the oracle results reveal that the WGs offer significantly more accurate hypotheses than the Viterbi-like approach (a reduction of %28 and %33 in WER for Dihana and MetEus respectively). As a result, resorting to re-scoring techniques would allow to extract hypotheses with better system performance.

Table 1. WER for several values of the stored predecessors (n) and number of frames merged (m) in both Dihana and Meteus corpora

	Dihana					MetEus				
(n,m)	(1,0)	(2,0)	(2,1)	(2,2)	(2,3)	(1,0)	(2,0)	(2,1)	(2,2)	(2,3)
p-best	**18.37**	20.01	21.15	22.01	22.68	**12.78**	12.90	13.47	14.38	15.05
oracle	18.37	**13.09**	13.26	13.30	13.42	12.78	8.45	8.45	**8.44**	8.45

The difference in performance between MetEus and Dihana tasks has to do with the difficulty of the task itself. While MetEus is read speech, Dihana is spontaneous speech with very long sentences at times.

5 Concluding Remarks and Future Work

We have presented a way of getting a WG from the lattice explored in the ASR process by storing a number (n) of predecessors at each node. Some approaches have been carried out (merging states, obtaining the n-best list of hypotheses from the WG, etc.) in order to make the computation affordable.

It has turned out that merging the states allows to alleviate the computational cost within the WG at the expense of generating hypotheses with possibly fake probabilities. Nevertheless, the explored WGs still contain more accurate hypotheses than the hypothesis obtained without considering a WG. Other minimisation strategies that have proven successful in other fields of PR (such as graph cut theory in computer vision [10]) could be explored.

Currently we are focusing on both re-scoring methods taking advantage of both time-segmentation and probabilities derived from the WG and also the integration of this sort of WG with other kind of WG aiming at speech translation or understanding systems.

References

1. Forney Jr., G.D.: The Viterbi Algorithm. Proc. of the IEEE 61, 268–278 (1973)
2. Hazen, T.J., Seneff, S., Polifroni, J.: Recognition confidence scoring and its use in speech understanding systems. Computer Speech & Language 16, 49–67 (2002)
3. Ferreiros, J., Segundo, R.S., Fernández, F., D'Haro, L., Sama, V., Barra, R., Mellén, P.: New word-level and sentence-level confidence scoring using graph theory calculus and its evaluation on speech understanding. In: Proc. Interspeech, pp. 3377–3380 (2005)
4. Blackwood, G.: Lattice Rescoring Methods for Statistical Machine Translation. PhD thesis, University of Cambridge (2010)
5. Jelinek, F.: Statistical Methods for Speech Recognition, 2nd edn. Language, Speech and Communication series. The MIT Press, Cambridge (1999)
6. Huang, X., Acero, A., Hon, H.: Spoken Language Processing: A guide to Theory, Algorithm, and System Development. Prentice Hall, Englewood Cliffs (2001)
7. Caseiro, D., Trancoso, I.: A specialized on-the-fly algorithm for lexicon and language model composition. IEEE TASLP 14, 1281–1291 (2006)
8. Benedí, J., Lleida, E., Varona, A., Castro, M., Galiano, I., Justo, R., López, I., Miguel, A.: Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA. In: Proc. of LREC 2006, Genoa, Italy (2006)
9. Pérez, A., Torres, M.I., Casacuberta, F., Guijarrubia, V.: A Spanish-Basque weather forecast corpus for probabilistic speech translation. In: Proc. of the 5t SALTMIL, Genoa, Italy (2006)
10. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 1222–1239 (2001)

Visual Word Aggregation

R.J. López-Sastre, J. Renes-Olalla, P. Gil-Jiménez, and S. Maldonado-Bascón

GRAM, Department of Signal Theory and Communications, University of Alcalá
robertoj.lopez@uah.es

Abstract. Most recent category-level object recognition systems work with visual words, *i.e.* vector quantized local descriptors. These visual vocabularies are usually constructed by using a single method such as K-means for clustering the descriptor vectors of patches sampled either densely or sparsely from a set of training images. Instead, in this paper we propose a novel methodology for building efficient codebooks for visual recognition using clustering aggregation techniques: the Visual Word Aggregation (VWA). Our aim is threefold: to increase the stability of the visual vocabulary construction process; to increase the image classification rate; and also to automatically determine the size of the visual codebook. Results on image classification are presented on the testbed PASCAL VOC Challenge 2007.

Keywords: clustering aggregation; visual words; object recognition.

1 Introduction

A popular strategy for representing images within the context of category-level object recognition is the *Bag-of-Words* (BoW) approach [3]. The basic idea behind this type of representation is to characterize an image by the histogram of its visual words, *i.e.* vector quantized local features (see Figure 1). Popular candidates for these local features are local descriptors [12] that can be extracted at specific interest points [3], densely sampled over the image [7], or via a hybrid scheme called *dense interest points* [18]. There are various clustering methods for creating these visual words. K-means or variants thereof, such as approximate K-means [15] or vocabulary trees [14], are currently the most common.

Subsequently, each local feature in an image is mapped to a cluster so as to represent any image as a histogram over the clusters. The BoW representation has been shown to characterize the images and objects within them in a robust yet descriptive manner, in spite of the fact that it ignores the spatial configuration between visual words. Moreover, variations on these BoW models have shown impressive results lately [17], wining the PASCAL Visual Object Classes Challenge on object classification.

Although such ideas appear to be quite exciting, there are 2 main challenges that need to be overcome. Since the clustering into visual words is unsupervised, this representation does not group semantically meaningful object parts (*e.g.* wheels or eyes). In practice, if the dataset is sufficiently coherent (*e.g.* images

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 676–683, 2011.

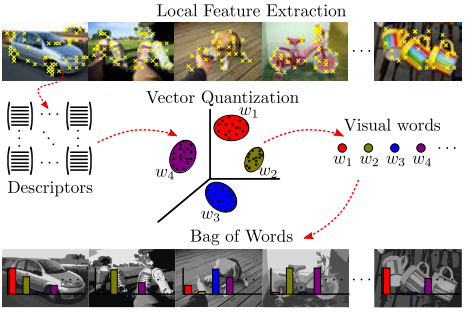


Fig. 1. BoW approach overview. It starts with the extraction of local features followed by robust description of the features, e.g. using SIFT [11]. The following step consists in vector quantizing the high dimensional space of local image descriptors to obtain a visual vocabulary. A BoW is then built as a histogram of visual word occurrences.

of only one particular class), only a reduced number of visual words represent semantic object parts. Moreover, when an unsupervised quantization is applied to a more diverse dataset, synonyms and polysemies are the norm rather than the exception [16].

On the other hand, there are the limitations of the clustering algorithms themselves. In general, data clustering usually has associated the stability problem: it is not possible to use cross validation for tuning the clustering parameters because of the absence of ground truth; the dependence on the initialization is a common problem for most of the iterative methods; the objectives pursued by each clustering approach are different and different structures in data may be discovered.

Specifically, K-means clustering output depends on the initialization as the procedure only undertakes the search for a local optimum and it requires the user to specify the number of clusters. Furthermore, it is computationally expensive for big values of K. Other approaches use efficient hierarchical clustering schemes (*e.g.* [9]) where one fixes a cut-off threshold on the cluster compactness. It may happen that some *real* clusters are split in several clusters, so that the visual words are not representative of all features. Furthermore, run-time and memory requirements are often significantly higher for these hierarchical methods.

Several attempts have been made to create efficient codebooks for visual recognition. There are some unsupervised approaches based on frequent itemset mining (*e.g.* [20]). Typically, finding representative visual words boils down to finding frequent co-occurring groups of descriptors in a transaction database obtained from the training images. Some supervised approaches use image annotation and class labels to guide the semantic visual vocabulary construction (*e.g.* [13,10]).

In this paper, we introduce a new methodology to obtain efficient visual words with a threefold objective: to overcome the problem of clustering stability; to increase the image classification rate; and also to automatically determine the size of the visual codebook. We propose to adapt the clustering aggregation techniques described in [6] to the visual vocabulary construction process. To the

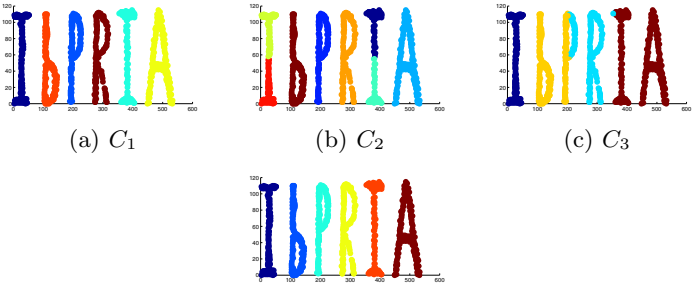
best of our knowledge, this is the first paper to describe such a clustering aggregation based approach within this context. We analyze how these techniques perform in discovering visual words using different combinations of quantization algorithms.

The rest of this paper is organized as follows. In Section 2 we introduce the clustering aggregation theory. Section 3 gives a detailed description of the novel approach we propose to adapt the clustering aggregation techniques to the visual vocabulary construction process. Experiments in image categorization are described in Section 4 and Section 5 concludes the paper.

2 Clustering Aggregation

The problem of clustering aggregation has been considered under a variety of names: consensus clustering, clustering combination and cluster ensembles. Many approaches have been proposed (*e.g.* the graph cut method [5] and the Bayesian method [19]).

In [6], clustering aggregation is defined as an optimization problem where, given a set of m clusterings, the objective is to find the clustering that minimizes the total number of disagreements with the m clusterings. Clustering aggregation can be considered as a metaclustering method to improve stability and robustness of clustering by combining the results of many clusterings. Moreover, it can determine the appropriate number of clusters while detecting outliers. A toy example to illustrate how the clustering aggregation works is depicted in Figure 2.



(a) C_1 (b) C_2 (c) C_3

(d) C

Fig. 2. Toy example. (a)-(c) are 3 different clusterings $\{C_1, C_2, C_3\}$ over the IbPRIA dataset of 2D points. (d) depicts the result of the clustering aggregation algorithm, the clustering C. Note that the solution C improves the clustering robustness and finds the 6 clusters in the IbPRIA dataset. We have used different colors to denote different clusters.

Gionis *et al.* [6] propose an approach to this problem based on correlation clustering techniques [1]. We are given a set of m clusterings $\{C_1, C_2, \ldots, C_m\}$. Our objective is to obtain a single clustering C that agrees as much as possible

with the m input clusterings. It is possible to define a distance $d(u, v)$ between two vectors u and v as the fraction of the m clusterings that place u and v in different clusters. Our objective is to find a clustering C that minimizes the function $d(C) = \sum_{C(u)=C(v)} d(u, v) + \sum_{C(u)\neq C(v)}(1 - d(u, v))$, where $C(v)$ denotes the label of the cluster to which v belongs to. In the experiments we have used the *Balls* and the *Agglomerative* (*Agg*) algorithms described in [6]. Both algorithms take as input a complete graph with all the distances between vectors. The *Balls* algorithm tries to find groups of nodes that are within a ball of fixed radius and far from other nodes. Once such a set is found, the algorithm considers it a new cluster and proceeds with the rest. The *Agg* is a bottom-up algorithm which starts with every node in a cluster. It merges two vertices if the average distance between them is less than a fixed value.

3 Visual Word Aggregation

In this work we propose to analyze how clustering aggregation algorithms work for building efficient visual vocabularies. We propose a novel BoW approach via Visual Word Aggregation (VWA). Our aim with this approach is threefold: to increase the stability of the codebook construction process, to automatically determine the size of the vocabulary, and to obtain better results in categorization. Figure 3 depicts the major steps of our proposal. In the first step, images are represented using local features (*e.g.* SIFT [11]). Then, the vector quantization processes start. We define m as the number of clustering algorithms that are executed, *i.e.* m is the number of codebooks. Different quantization algorithms and/or several executions of the same algorithm can be used. The VWA uses these m initial codebooks to build the vocabulary to be used in the BoW approach.

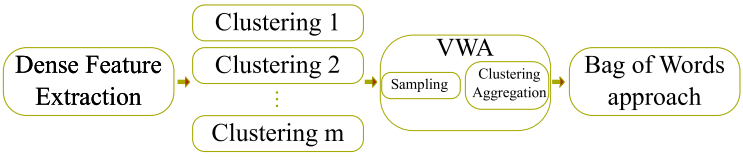


Fig. 3. Flowchart of our novel approach for image classification via VWA

However, a direct application of the clustering aggregation algorithms in [6] to the m codebooks is not feasible. Every clustering defines a vocabulary that organizes the local descriptors in a high dimensional space (*e.g.* 128 dimensions for SIFT descriptors). Furthermore, thousands of descriptors are extracted from each image, so we have to deal with large datasets of vectors, where the number of clusters is high too. The algorithms described in [6] take the distance matrix as input so their complexity is quadratic in the number of data objects in the dataset, which makes them inapplicable to large datasets. Gionis *et al.* [6] presented a sampling algorithm to overcome this problem. In a preprocessing step,

their algorithm samples a set of nodes S uniformly at random from the dataset. The set S is the input for the clustering aggregation algorithm. In the postprocessing step, the algorithm goes through the nodes not in S and decides whether to place it on one of the existing clusters or to create a singleton. Nonetheless, we observed experimentally that the time complexity of their approach is high within our context, *i.e.* when the number of clusters and the dimensionality of vectors are high.

In order to reduce the run-time of the visual vocabulary construction, we define a new sampling strategy. Let O be the dataset of local descriptors of size N, $O = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$. We start with a uniform and random sample $R \subset O$ of size $M = \beta N$, where $\beta \in [0, 1]$ is the sampling factor. As in [6], the set R is sampled to obtain the subset $S \subset R$. The set S is given as input to the clustering aggregation algorithm which builds a clustering $C = \{c_1, c_2, \ldots, c_K\}$. Note that with our sampling scheme, the postprocessing step only needs to evaluate the elements in R and not in S, which significantly reduces the run-time of the original approach. Finally, we inspect the vectors in O and not in R and assign them to the nearest cluster. Using this double sampling strategy we can handle large datasets letting VWA converge into a final codebook.

4 Results

Experimental Setup. Our aim is to evaluate, within the context of image classification, the performance of the VWA approach. So as to obtain reliable results, we use the PASCAL VOC Challenge 2007 database [4]. This challenge is widely acknowledged as a difficult testbed for both object detection and image categorization. We select the *trainval* and *test* set for training and testing the classifier respectively. See [4] for further details.

For image representation, we use SIFT [11] descriptors of 16×16 pixel patches computed over a grid with spacing of 8 pixels. With these descriptors we perform the vocabulary construction via VWA. Specifically, we use our own implementations of the K-means and the Jurie and Triggs (J&T) [7] clustering algorithms. In the clustering aggregation step we integrate our novel sampling methodology with the *Balls* and the *Agg* algorithms [6].

Support Vector Machines (SVMs) are used for classification. We experiment with the Histogram Intersection Kernel (HIK) which has shown good results in object recognition [8]. The HIK applied to two feature vectors $\mathbf{x}$ and $\mathbf{x}'$ of dimension d is defined as $k(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \min(\mathbf{x}(i), \mathbf{x}'(i))$. Specifically, we use libSVM [2]. A 10-fold cross-validation on the *trainval* set to tune SVM parameters is conducted to train each classifier. We follow the image classification evaluation procedure proposed by the PASCAL VOC Challenge [4] using the Mean Average Precision (MAP), which is computed by taking the mean of the average precisions for the 20 classes for each method.

Codebooks performance in image classification. We evaluate the MAP in image categorization for the codebooks described in Table 1. Note that codebooks C1 and C4 have been obtained without using the VWA approach,

Table 1. Codebooks obtained for the experiments in image classification

	Codebook description
C1	K-means ($K = 200$)
C2	3 K-means ($K = 200$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)
C3	3 K-means ($K = 200$) and *Agg* + Sampling ($\beta = 0.33$)
C4	J&T ($r = 0.83, N = 3000$)
C5	3 J&T ($r = 0.8, N = 3000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.25$)
C6	2 K-means ($K = 200$) + J&T ($r = 0.92, N = 3000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)
C7	J&T ($r = 0.8, N = 3000$) + K-means ($K = 2000$) and *Balls* ($\alpha = 0.25$) + Sampling ($\beta = 0.5$)

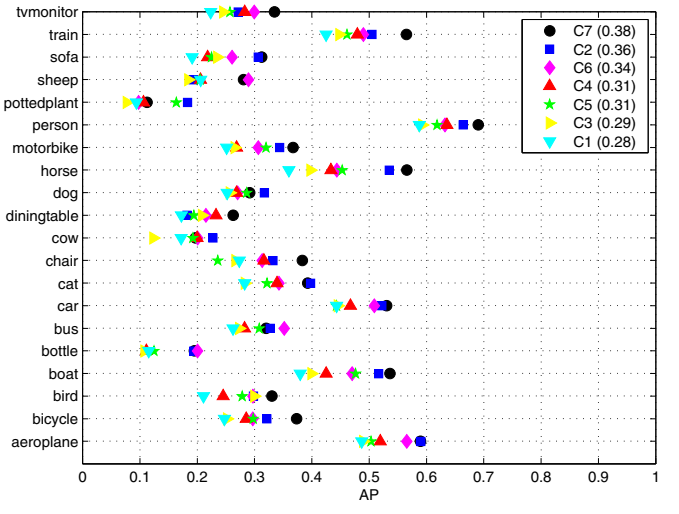


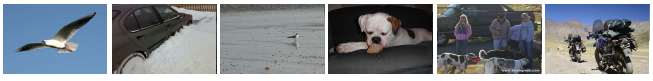
Fig. 4. Evaluation of codebooks on image categorization over the PASCAL VOC 2007 Challenge. Average precision per class for each method is shown.

i.e. following a traditional BoW approach. Results per object category are shown in Figure 4. The aggregation of 1 K-means and 1 J&T, *i.e.* codebook C7, obtains the best MAP (0.38). Furthermore, all the codebooks generated via VWA using the *Balls* algorithm and our sampling approach (vocabularies C2, C5, C6 and C7), obtain better results than when a traditional BoW is used (C1 and C4). Comparing C2 and C3 we also have observed that the *Balls* algorithm performs better than the *Agg*. Moreover, for the *Balls* algorithm, we have found that $\alpha \leq 0.25$ leads to better results in image categorization. We observed experimentally that the sampling factor β directly affects to the classification performance: the best results are obtained for $\beta \geq 0.5$. Finally, Figure 5 shows ranked images for 4 different classes.

Discussion. Results confirm that the VWA technique can be used to obtain better vocabularies. It is also useful for large sets of vectors in high-dimensional spaces. Such spaces are sparse with the data points far away from each other.

(a) highest ranked positive images

(b) lowest ranked positive images



(c) highest ranked negative images

Fig. 5. Ranked images for the classes aeroplane, bicycle, boat, cat, horse and person. (a) positive images assigned the highest rank. (b) positive images assigned the lowest rank. (c) negative images assigned the highest rank, *i.e.* images which confuse the classifiers.

Furthermore, all pairwise distances in a high-dimensional data set seem to be very similar. The phenomenon is known in the statistical literature as the *curse of dimensionality*. This may lead to problems when searching for clusters. K-means is a popular algorithm for its simplicity. Unfortunately, centers tend to be tightly clustered near dense regions and sparsely spread in sparse ones. The J&T [7] is a mean-shift based approach that can be used to overcome some of the limitations of K-means. The VWA technique can be used to combine the properties of K-means and J&T clustering algorithms to obtain better visual vocabularies.

5 Conclusion

We have introduced the VWA methodology which incorporates the clustering aggregation techniques to the visual codebook construction process. To the best of our knowledge, this is the first paper to describe such a clustering aggregation based methodology within this context. Also, a novel sampling strategy has been designed in order to use the VWA approach with large sets of vectors in high dimensional spaces. Results show that the MAP increases when the vocabularies are obtained via VWA. Exploring other clusterings as well as other datasets is one interesting avenue of future research.

Acknowledgements. This work was partially supported by projects TIN2010-20845-C03-03 and CCG10-UAH/TIC-5965.

References

1. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Machine Learning 56, 89–113 (2004)
2. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)

3. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV (2004)
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC 2007) Results (2007), http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html
5. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: ICML (2004)
6. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data 1(1), 4 (2007)
7. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: CVPR (2005)
8. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
9. Leibe, B., Mikolajczyk, K., Schiele, B.: Efficient clustering and matching for object class recognition. In: BMVC (2006)
10. López-Sastre, R.J., Tuytelaars, T., Acevedo-Rodríguez, J., Maldonado-Bascón, S.: Towards a more discriminative and semantic visual vocabulary. Computer Vision and Image Understanding 115(3), 415–425 (2011)
11. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. PAMI 27(10), 1615–1630 (2005)
13. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS (2006)
14. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR, pp. 2161–2168 (2006)
15. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
16. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Van Gool, L.: Modeling scenes with local descriptors and latent aspects. In: ICCV (2005)
17. van de Sande, K., Gevers, T., Snoek, C.: Evaluation of color descriptors for object and scene recognition. In: CVPR (2008)
18. Tuytelaars, T.: Dense interest points. In: CVPR (2010)
19. Wang, H., Shan, H., Banerjee, A.: Bayesian cluster ensembles. In: SDM (2009)
20. Yuan, J., Wu, Y.: Context-aware clustering. In: CVPR (2008)

Character-Level Interaction in Multimodal Computer-Assisted Transcription of Text Images

Daniel Martín-Albo, Verónica Romero, Alejandro H. Toselli, and Enrique Vidal*

Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Spain
{dmartinalbo,vromero,ahector,evidal}@iti.upv.es

Abstract. To date, automatic handwriting text recognition systems are far from being perfect and heavy human intervention is often required to check and correct the results of such systems. As an alternative, an interactive framework that integrates the human knowledge into the transcription process has been presented in previous works. In this work, multimodal interaction at character-level is studied. Until now, multimodal interaction had been studied only at whole-word level. However, character-level pen-stroke interactions may lead to more ergonomic and friendly interfaces. Empirical tests show that this approach can save significant amounts of user effort with respect to both fully manual transcription and non-interactive post-editing correction.

1 Introduction

At present time, the use of automatic handwritten text recognition systems (HTR) for the transcription of manuscript document images is far from being useful, mainly because of the unrestricted vocabulary and/or handwriting styles involved in such documents. Typically, the automatic transcriptions obtained by these HTR systems need a heavy human *post–editing* process in order to obtain transcriptions of standard quality. In practice, such a *post–editing* solution becomes rather inefficient, expensive and hardly acceptable by professional transcribers.

In previous works [7,5], a more effective, *interactive* on-line approach was presented. This approach, called "Computer Assisted Transcription of Handwritten Text Images" (CATTI), combines the accuracy ensured by the human transcriber with the efficiency of the HTR systems to obtain final perfect transcriptions. Empirical results show that the use of CATTI systems can save a substantial quantity of human effort with respect to both pure manual transcriptions and post-editing.

So far, human corrective feedback for CATTI has been studied at two different levels: a) whole–word interactions (both typed and handwritten using an e–pen interface [7]) and b) (typed) character–level corrections [5]. According to the results of these works, keystroke corrections can save a significant quantity of human effort with respect to whole-word corrections, while multimodal, e-pen interaction seems more ergonomic for human transcribers, which is a key point in the design of friendly and usable user interfaces.

* Work supported by the Spanish Government (MICINN and "Plan E") under the MITTRAL (TIN2009-14633-C03-01) research project and under the research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and by the Generalitat Valenciana under grant Prometeo/2009/014.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 684–691, 2011.

In this work, we focus on character level interaction using the more ergonomic *e–pen handwriting* modality, which is perhaps the most natural way to provide the required feedback in CATTI systems. It is important to note, however, that the use of this kind of non–deterministic feedback typically increases the overall interaction cost in order to correct the possible feedback decoding errors. Nevertheless, by using informations derived from the interaction process, we will show how the decoding accuracy can be significantly improved over using a plain e–pen handwriting recognizer which can not take advantage of the interaction context.

2 CATTI Overview

In the original CATTI framework, the human transcriber (named *user* from now on) is directly involved in the transcription process since he is responsible of validating and/or correcting the HTR outputs. The process starts when the HTR system proposes a full transcription of a feature vector sequence x, extracted from a handwritten text line image. The user validates an initial part of this transcription, p', which is error-free and introduces a correct word, v, thereby producing correct transcription *prefix*, $p = p'v$. Then, the HTR system takes into account the available information to suggest a new suitable continuation *suffix*, s. This process is repeated until a full correct transcription of x is accepted by the user [7].

At each step of this process, both the image representation, x, and a correct transcription prefix p are available and the HTR system should try to complete this prefix by searching for the most likely suffix $\hat{s}$ as:

$$\hat{s} = \arg\max_{s} P(s \mid x, p) = \arg\max_{s} P(x \mid p, s) \cdot P(s \mid p) \tag{1}$$

Since the concatenation of p and s constitutes a full transcription hypothesis, $P(x \mid p, s)$ can be approximated by concatenated character Hidden Markov Models (HMMs) [2,4] as in conventional HTR. On the other hand, $P(s \mid p)$ is usually approximated by dynamically modifying a n-gram in order to cope with the increasingly consolidated prefixes [7]. Let $p = p_1^k$ be a consolidated prefix and $s = s_1^l$ a possible suffix:

$$P(s \mid p) \simeq \prod_{j=1}^{n-1} P(s_j \mid p_{k-n+1+j}^{k}, s_1^{j-1}) \cdot \prod_{j=n}^{l} P(s_j \mid s_{j-n+1}^{j-1}) \tag{2}$$

In order to make the system more ergonomic and friendly to the user, interaction based on characters (rather than full words) has been studied in [5] with encouraging results. Now, as soon as the user types a new keystroke (character), the system proposes a suitable continuation following the same process described above. As the user operates now at the character level, the last word of the prefix may be incomplete. In order to *autocomplete* this last word, it is assumed that the prefix p is divided into two parts: the fragment of the prefix formed by complete words (p'') and the last incomplete word of the prefix (v_p). In this case the HTR decoder has to take into account x, p'' and v_p, in order to search for a transcription suffix $\hat{s}$, whose first part is the continuation of v_p:

$$\hat{s} = \arg\max_{s} P(s \mid x, p'', v_p) = \arg\max_{s} P(x \mid p'', v_p, s) \cdot P(s \mid p'', v_p) \tag{3}$$

Again, the concatenation of p'', v_p and s constitutes a full transcription hypothesis and $P(x\,|\,p'', v_p, s)$, can be modelled with HMMs. On the other hand, to model $P(s \mid p'', v_p)$ we assume that the suffix s is divided into two fragments: v_s and s''. v_s is the first part of the suffix that corresponds with the final part of the incomplete word of the prefix, i.e, $v_p v_s = v$ where v is an existing word in the task dictionary (Σ), and s'' is the rest of the suffix. So, the search must be performed over all possible suffixes s of p, and the language model probability $P(v_s, s'' \mid p'', v_p)$ must ensure that the concatenation of the last part of the prefix v_p, and the first part of the suffix, v_s, form an existing word (v) in the task dictionary. This probability can be decomposed into two terms:

$$P(v_s, s'' \mid p'', v_p) = P(s'' \mid p'', v_p, v_s) \cdot P(v_s \mid p'', v_p) \tag{4}$$

The first term accounts for the probability of all the whole-words in the suffix, and can be modelled directly by (2). The second term should ensure that the first part of the suffix (usually a word-ending-part) v_s, will be a possible suffix of the incomplete word v_p, and can be stated as:

$$P(v_s \mid p'', v_p) = \begin{cases} \frac{P(v_p, v_s|p'')}{\sum_{v'_s} P(v_p, v'_s|p'')} & \text{if } v_p v_s \in \Sigma \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

3 Multimodal CATTI (MM-CATTI) at the Character Level

One way to increase the ergonomy and the usability in CATTI is to allow the user to provide his or her validating and/or corrective feedback by means of more comfortable peripheral devices, such as e–pen or touchscreen.

Clearly, decoding this kind of non-deterministic feedback consists in *on-line* HTR. As previously mentioned, the information available in the interaction process, allows us to boost the accuracy of this on-line HTR subsystem with respect to a conventional on-line HTR decoder (which do not make use of the interaction-derived information).

Let x be the representation of the input image and p' a user-validated prefix of the transcription. Let t be the on-line touchscreen pen strokes provided by the user. These data are related to the suffix suggested by the system in the previous interaction step, s', and are typically aimed at accepting or correcting parts of this suffix. Using this information, the system has to find a new suffix, $\hat{s}$, as a continuation of the previous prefix p', considering all possible decodings, d, of the on-line data t and some information from the previous suffix s'. That is:

$$\begin{aligned} \hat{s} &= \arg\max_s P(s \mid x, s', p', t) = \arg\max_s \sum_d P(s, d \mid x, p', s', t) \\ &\approx \arg\max_s \max_d P(t \mid d) \cdot P(d \mid p', s') \cdot P(x \mid s, p', d) \cdot P(s \mid p', d) \end{aligned} \tag{6}$$

An approximate two-step solution to this difficult optimization problem is followed (see Figure 1). In the first step, an *"optimal"* decoding, $\hat{d}$, of the on-line pen-strokes t is computed using only the first two terms of equation (6). After observing this decoding, $\hat{d}$, the user may type additional keystrokes, κ, to correct possible errors in $\hat{d}$. In the

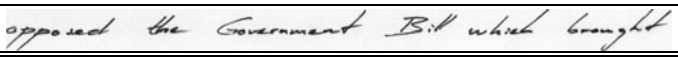
		x	
INTER-0		**p** $\hat{\mathbf{s}} \equiv \hat{\mathbf{w}}$	**opposite this Comment Bill in that thought**
INTER-1	Step-1	**p', t** $\hat{d}$ κ	**oppos**e l e
	Step-2	**p** $\hat{\mathbf{s}} \equiv$ **s'**	**oppose** **d the Government Bill in that thought**
INTER-2	Step-1	**p', t** $\hat{d}$ κ	**opposed the Government Bill** ω w
	Step-2	**p** $\hat{\mathbf{s}} \equiv$ **s'**	**opposed the Government Bill w** **hich brought**
FINAL		κ $\mathbf{p} \equiv \mathbf{T}$	# **opposed the Government Bill which brought**

Fig. 1. Example of multimodal CATTI at character level interaction. The process starts when the HTR system proposes a full transcription of the handwritten text image x. Then, each interaction consists in two steps. In the first step the user handwrites some touchscreen to amend the suffix proposed by the system in the previous step. This defines a correct prefix **p'**, which can be used by the on-line HTR subsystem to obtain a decoding of **t**. After observing this decoding, $\hat{d}$, the user may type additional keystrokes, κ, to correct possible errors in $\hat{d}$. On the second step, a new prefix is built from the previous correct prefix **p'**, the decoded on-line handwritten text, $\hat{d}$, and the typed text κ. Using this information, the system proposes a new potential suffix. The process ends when the user enters the special character "**#**". System suggestions are printed in boldface and typed text in typewriter font. In the final transcription, **T**, underlined italic characters are those which were typed by the user.

second step, the first two terms of (6) are ignored and d is replaced with $\hat{d}$ in the last two terms. This way, a new consolidated prefix $p = p'\hat{d}$ is obtained, which leads to a formulation identical to (1). These two steps are repeated until p is accepted by the user as a full correct transcription of x.

Assuming whole-word e–pen feedback, this approach was studied and tested in [7], with good results. Here we consider single-character e–pen strokes, which we think may lead to improved productive and usability. Therefore, we henceforth assume that $\hat{d}$ consists of a single character. As in section 2, the prefix p' is divided into two parts: p'' (fragment of p' formed by complete words) and v'_p (the last incomplete word of p'). Therefore the first step of the optimization (6) can be written as:

$$\hat{d} = \arg\max_{d} P(t \mid d) \cdot P(d \mid p'', v'_p, s') \tag{7}$$

where, $P(t \mid d)$ is provided by a morphological (HMM) model of the character d and $P(d \mid p'', v'_p, s')$ can be approached by a language model dynamically constrained by information derived from the interaction process. Equation (7) may lead to several scenarios depending on the assumptions and constraints adopted for $P(d \mid p'', v'_p, s')$. We examine some of them bellow.

The first and simplest scenario corresponds to a naive approach where any kind of interaction-derived information is considered; that is, $P(d \mid p'', v'_p, s') \equiv P(d)$.

In a slighty more restricted scenario, we take into account just the information from the previous off-line HTR prediction s'. The user interacts providing t in order to correct the wrong character of s', e, that follows the validated prefix p'. Clearly, the erroneous character e should be prevented to be a decoding on-line HTR result. This *error-conditioned model* can be written as $P(d \mid p'', v'_p, s') \equiv P(d \mid e)$.

Another, more restrictive scenario, using the information derived from the validated prefix p', arises when we regard the portion of word already validated (v'_p), i.e. $P(d \mid p'', v'_p, s') \equiv P(d \mid v'_p, e)$. In this case the decoding should be *easier* as we know beforehand what should be a suitable continuation of the part of word accepted so far.

Finally, the most restrictive scenario corresponding to the additional consideration of the information provided by p'', is left for future studies.

3.1 Dynamic Language Modelling for Character-Level MM-CATTI

Language model restrictions are implemented on the base of n-grams, depending on each multimodal scenario considered. As mentioned above, the simplest scenario is that which does not take into account any information derived from the interaction. In this case, $P(d)$ can be modelled directly using uni-grams. This is the *baseline* case.

The second case, $P(d \mid e)$, only considers the first wrong character. The language model probability is given by

$$P(d \mid e) = \begin{cases} 0 & \text{if } d = e \\ \frac{P(d)}{1-P(e)} & \text{if } d \neq e \end{cases} \tag{8}$$

The next scenario, given by $P(d \mid v'_p, e)$, the on-line HTR subsystem counts not only on the first wrong character but also on the last incomplete word of the validated prefix v'_p. This scenario can be approached in two different ways: using a character language model or a word language model. In the first one, the on-line HTR subsystem uses a modified character n-gram model:

$$P(d \mid v'_p, e) = \begin{cases} 0 & \text{if } d = e \\ \frac{P(d|v'^{\,k}_{p\,k-n+2})}{1-P(e|v'^{\,k}_{p\,k-n+2})} & \text{if } d \neq e \end{cases} \tag{9}$$

In the second approach (10), we use a word language model to generate a more refined character language model. This can be written as:

$$P(d \mid v'_p, e) = \begin{cases} 0 & \text{if } d = e \\ \frac{P(d|v'_p)}{1-P(e|v'_p)} & \text{if } d \neq e \end{cases}$$

where:

$$P(d \mid v'_p) = \frac{P(v'_p, d)}{\sum\limits_{d'} P(d', v'_p)} = \frac{\sum\limits_{v_s} P(v'_p, d, v_s)}{\sum\limits_{v_s}\sum\limits_{d'} P(v'_p, d', v_s)} \tag{10}$$

being $v'_p d v_s$ an existing word of Σ.

4 Off- and On-Line HTR System Overview

Both the off-line and on-line HTR systems employ a similar conceptual architecture composed of three modules: *preprocessing*, *feature extraction* and *recognition*. The first two entail different well-known standard techniques depending on the data type, but the last one is identical for both systems. The Off-line HTR preprocessing involves skew and slant corrections and size normalization operation [8]. On the other hand, on-line handwriting preprocessing encompasses repeated points elimination and noise reduction. Regarding feature extraction, the off-line case converts the preprocessed text into a sequence of 60-dimensional feature vectors, whereas the on-line preprocessed coordinates are transformed into a sequence of 7-dimensional feature vectors [6].

As explained above, the recognition process is similar in both cases. Characters are modelled by continuous density left-to-right HMMs with a Gaussian mixture per state. Each lexical word is modelled by a Stochastic Finite-State automaton, and text sentences are modelled using bi-grams with Kneser-Ney back-off smoothing. All these finite-state models can be easily integrated into a single global model in which decoding process is efficiently performed by the Viterbi algorithm.

5 Experimental Framework

For test the effectiveness of MM-CATTI at character level different experiments were carried out. The corpora and the performance measures used are explained below.

5.1 Assessment Measures

Some types of measures have been adopted to assess the performance of character-level transcription. On the one hand, to make the post-editing process more accurately comparable to CATTI at character level, we introduce a *post-editing autocompleting* approach. Here, when the user enters a character to correct some incorrect word, the system automatically completes the word with the most probable word on the task vocabulary. Hence we define the *Post-editing Key Stroke Ratio* (PKSR), as the number of keystrokes that the user must enter to achieve the reference transcription, divided by the total number of reference characters. On the other hand, the effort needed by a human transcriber to produce correct transcriptions using CATTI at character level is estimated by the *Key Stroke Ratio* (KSR), which can be defined as the number of (character level) user interactions that are necessary to achieve the reference transcription of the text image considered, divided by the total number of reference characters. These definitions make PKSR and KSR comparable and the relative difference between them gives us a good estimate of the reduction in human effort that can be achieved by using CATTI at character level with respect to using a conventional HTR system followed by human autocompleting postediting. This *estimated effort reduction* will be denoted as "EFR".

Finally, since only single-character corrections are considered, the conventional classification error rate (ER) will be used to assess the accuracy of the on-line HTR feedback subsystem under the different constraints entailed by the MM-CATTI at character level interaction process.

5.2 Corpus

The character level CATTI was evaluated on the IAMDB corpus. For the MM-CATTI, the on-line UNIPEN corpus was employed to simulate the user touchscreen interactions.

The IAMDB [3] is a publicly accessible corpus composed of 1,539 scanned text pages, handwritten by 657 different writers. No restriction was imposed related to the writing style or with respect to the pen used. The database is provided at different segmentation levels: characters, words, lines, sentences and page images. Here we use sentence-segmented images. Each sentence is accompanied by its ground through transcription as the corresponding sequence of words. To better focus on the essential issues of the considered problems, no punctuation marks, diacritics, or different word capitalizations are included in the transcriptions. From 2,324 sentences that forms the corpus, 200 were used as test, leaving the rest as training partition.

The UNIPEN corpus [1] comes organized in several categories: lower and upper-case letters, digits, symbols, isolated words and full sentences. For our experiment, three UNIPEN categories were used: *1a* (digits), *1c* (lowercase letters) and *1d* (symbols). Three arbritary writers were chosen as test partition and 17 as training data [7].

5.3 Results

Different experiments have been carried out to asses the feasibility and potential of CATTI at character level. Two types of results are reported for CATTI at character level: the PKSR (first column of table 2) and the KSR (second column of table 2). The 12.5% of KSR corresponds to a total of 1,627 characters that the user has to correct. In the MM-CATTI at character level these characters would have to be handwritten by the user on the touchscreen. It is simulated here using character samples belonging to a same writer from the UNIPEN corpus.

As we mentioned earlier, the introduction of multimodal interactivity leads, on the one hand, to an ergonomic and easier way of working, but on the other hand, to a situation where the system has to deal with non-deterministic feedback signals. Therefore, two of the most important concerns here is the accuracy of the on-line HTR subsystem and the determination of how much this accuracy can be boosted by taking into account informations derived from the interaction process. Table 1 reports the writer average feedback on-line recognition error rate of characters considering the different scenarios studied in section 3. As observed, feedback decoding accuracy increases significantly as more interaction derived constraints are taken into account. In addition, Table 1 also shows the relative accuracy improvements obtained respect to the baseline case.

Table 1. On-line HTR subsystem error rates for the four language models: plain character uni-gram (CU, *baseline*), error conditioned character uni-gram (CU_e), prefix-and-error conditioned character bi-gram (CB_e) and prefix-and-error conditioned word uni-gram (WU_e). The relative accuracy improvements for CU_e, CB_e and WU_e are shown in the last three columns. The same GSF value (15) is used for all the cases. All values are in percentages.

Error Rate				Relative Improv.		
CU	CU_e	CB_e	WU_e	CU_e	CB_e	WU_e
7.0	6.9	6.7	5.0	1.4	4.3	28.6

Table 2. From left to right: PKSR obtained with the post-editing autocompleting approach, KSR achieved with CATTI at character level and KSR obtained with the *baseline* and best scenarios for MM-CATTI approach. EFR for KSR of CATTI with respect to PKSR and for KSR for the two scenarios of MM-CATTI with respect to PKSR. All results are in percentages.

	CATTI	MM-CATTI		EFR		
PKSR	KSR	CU-KSR	WU_e-KSR	CATTI	MM-CATTI (CU)	MM-CATTI (WU_e)
15.8	12.5	13.4	13.1	20.9	15.2	17.1

As a final overview, Table 2 summarizes all the CATTI and MM-CATTI results obtained in this work. The third and fourth columns show the MM-CATTI KSR for the baseline as well as the best scenarios. This values are calculated under the simplifying assumption that the cost of keyboard-correcting a feedback on-line decoding error is similar to that of another on-line touchscreen interaction step. That is, each correction is counted twice: one for the failed touchscreen attempt and another for the keyboard correction itself. According to these results, the expected user effort for the best MM-CATTI approach is only barely higher than that of CATTI.

6 Conclusions

In this paper, we have studied the character level interaction in the CATTI system presented in previous works using pen strokes handwritten on a touchscreen as a complementary means to introduce the required CATTI correction feedback. From the results, we observe that the use of this more ergonomic feedback modality comes at the cost of a reasonably small number of additional interaction steps needed to correct the few feedback decoding errors. The number of these extra steps is kept very small thanks to the ability to use interaction-derived constraints to considerably improve the on-line HTR feedback decoding accuracy.

References

1. Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S.: UNIPEN Project of On-Line Data Exchange and Recognizer Benchmarks. In: Proc. of the 14th International Conference on Pattern Recognition, Jerusalem, Israel, pp. 29–33 (1994)
2. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1998)
3. Marti, U.V., Bunke, H.: A full English sentence database for off-line handwriting recognition. In: Proc. of the ICDAR 1999, Bangalore, India, pp. 705–708 (1999)
4. Rabiner, L.: A Tutorial of Hidden Markov Models and Selected Application in Speech Recognition. Proc. IEEE 77, 257–286 (1989)
5. Romero, V., Toselli, A.H., Vidal, E.: Character-level interaction in computer-assisted transcription of text images. In: Proc. ICFHR 2010, Kolkata, India, pp. 539–544 (November 2010)
6. Toselli, A.H., Pastor, M., Vidal, E.: On-Line Handwriting Recognition System for Tamil Handwritten Characters. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4477, pp. 370–377. Springer, Heidelberg (2007)
7. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. Pattern Recognition 43(5), 1814–1825 (2010)
8. Toselli, A.H., Romero, V., i Gadea, M.P., Vidal, E.: Preprocessing and feature extraction technics for multimodal interactive transcription of text images. Tech. rep., Instituto Tecnológico de Informática (2008), `http://prhlt.iti.es`

Simultaneous Lesion Segmentation and Bias Correction in Breast Ultrasound Images

Gerard Pons[1], Joan Martí[1], Robert Martí[1], and J. Alison Noble[2]

[1] Department of Computer Architecture and Technology,
University of Girona, Girona, Spain
{gpons,joanm,marly}@eia.udg.edu
[2] Department of Engineering Science,
University of Oxford, Oxford, United Kingdom
noble@robots.ox.ac.uk

Abstract. Ultrasound (US) B-mode images often show intensity inhomogeneities caused by an ultrasonic beam attenuation within the body. Due to this artifact, the conventional segmentation approaches based on intensity or intensity-statistics often do not obtain accurate results. In this paper, Markov Random Fields (MRF) and a maximum *a posteriori* (MAP) framework in combination with US image spatial information is used to estimate the distortion field in order to correct the image while segmenting regions of similar intensity inhomogeneity. The proposed approach has been evaluated using a set of 56 breast B-mode US images and compared to a radiologist segmentation.

Keywords: Breast lesion segmentation, ultrasound, inhomogeneity correction, Markov Random Fields, unsupervised initialization.

1 Introduction

Breast cancer is one of the leading causes of death for women all over the world and more than 8% women will suffer this disease during their life time [4]. Detection in early stages of the cancer development is the key to reduce the death rate (40% or more) [3]. The earlier the cancers are detected, the better treatment can be provided. Currently, the most effective and used modality for breast cancer detection and diagnostic is digital mammography (DM) [3]. However, there are some limitations of DM imaging in breast cancer detection, mostly in dense breasts where lesions can not be easily detected. Nowadays, an important alternative to DM is ultrasound (US) imaging, which is also commonly used as a complementary technique for breast cancer detection. One of the advantages of US against DM is its higher sensitivity when detecting lesions in dense breasts, reducing considerably the rate of false positives and thus the number of unnecessary biopsies. However, reading and understanding ultrasound images requires well-trained and experienced radiologists because of its complexity. Here lies the significance of the segmentation methods. An accurate and solid segmentation method is needed to, for instance, check if the lesion has grown through time.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 692–699, 2011.

Several segmentation methods have been applied to breast US images, from the most traditional histogram thresholding methods [5,11] to novel approaches based on graph-cuts [7,12]. Although active-contours methodologies are widely used [6,8] to determine the outline of the object of interest, they fail when dealing with blurred boundary lesions. Ultrasound image segmentation can also be considered as a labeling problem where the solution is to assign a set of labels to pixels, which is a natural representation for Markov Random Fields [2,10]. Within this group we want to remark the work of Xiao et al. [10] who adapted a bias field removal method in magnetic resonance imaging (MRI) to segment abnormalities in breast B-mode US images. Precise and accurate results were obtained but the method required expert supervision to initialize the process. In this paper, we present different approaches to reduce the human interaction to one-click process and we evaluate these approaches and compare them to the original method using a data-set composed by breast US images with different lesion typologies.

2 Methodology

The method proposed by Xiao et al. considered the bias field as an additive artifact of the logarithmic ideal image. They estimated this field to restore the ideal image while at the same time identifying regions of similar intensity inhomogeneity using an MRF-MAP framework (see [10] for detail). Our proposal consists of combining this MRF-MAP framework with spatial information (see Fig. 1) in order to overcome the limitations of the method in terms of user interaction.

2.1 Image Model

The Xiao et al. proposal assumed that an attenuation-related intensity inhomogeneities is described as a multiplicative field with low-frequency. A logarithmic transformation of such multiplicative model yields an addition. Let y and y^* denote, respectively, the observed and ideal log-transformed intensities respectively, then $y = y^* + d$, where d denotes the log-transformed intensity distortion field and x_i denotes the corresponding class label of pixel i. A Gaussian distribution model is used to describe the intensity distribution in an image

$$p(y_i^*|x_i) = g(y_i^*; \theta(x_i)) \tag{1}$$

where $g(y_i^*; \theta(x_i))$ is a Gaussian function and $\theta = \{\mu, \sigma\}$

Taking the distortion field into account and reformulating as a class-independent the intensity distribution (see Eq.1) can be modeled as a Gaussian mixture

$$p(y|d) = \prod_{i \in \mathcal{S}} \sum_{j \in \mathcal{L}} g(y_i - d_i; \theta(x_i)) p(x_i = j) \tag{2}$$

where $\mathcal{S}$ denotes the image pixels set and d the distortion field.

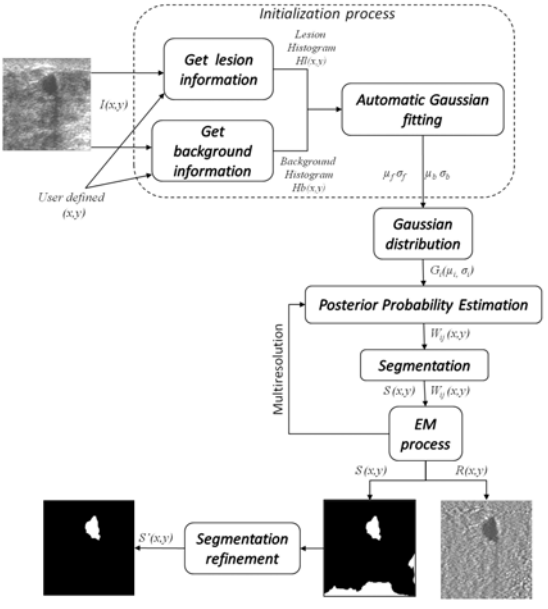


Fig. 1. Block diagram of our proposal, based on Xiao et al. [10] method

2.2 Expectation-Maximization (EM) Algorithm

As in Xiao et al. [10], Baye's rule can be used to obtain the posterior probability of the distortion field ($p(d|y)$), given the observed intensity values

$$p(d|y) = \frac{p(y|d)p(d)}{p(y)} \tag{3}$$

where $p(y)$ is a normalization factor and $p(d)$ is modeled as a Gaussian distribution with zero mean to capture its smoothness property. The maximum a posteriori principle is employed to obtain the optimal estimate of the distortion field. A zero-gradient condition is then used to assess this maximum, which leads to (see [9] for detail)

E step:

$$W_{ij} = \frac{p(y_i|x_i, d_i)p(x_i = j)}{p(y_i|d_i)} \tag{4}$$

M step:

$$d_i = \frac{[FR]_i}{[F\psi^{-1}E]_i}, \text{ with } E = (1, 1, ...)^T \tag{5}$$

where W_{ij} is the posterior probability that pixel i belongs to class j given the distortion field estimate. F is a low-pass filter and R is the mean residual in which for pixel i

$$R_i = \sum_{j \in \mathcal{L}} \frac{W_{ij}(y_i - \mu_j)}{\sigma_j^2} \tag{6}$$

and ψ is the mean inverse covariance

$$\psi_{ik}^{-1} = \begin{cases} \sum_{j \in \mathcal{L}} W_{ij} \sigma_j^{-2} & \text{if } i = k \\ 0 & \text{if } otherwise \end{cases} \tag{7}$$

As we said before, W_{ij} is the posterior probability that pixel i belongs to class j, and it is updated to the following form after using the MRF prior model

$$W_{ij} = \frac{p(y_i|x_i, d_i)p(x_i = j|x_{\mathcal{N}_i})}{p(y_i|d_i)} \tag{8}$$

where $p(x_i = j|x_{\mathcal{N}_i})$ has the form of Eq. 2.

To obtain the estimation of the low-frequency distortion field, the EM algorithm is used to update one label image and intensity inhomogeneity field iteratively. Such an updating process converges rapidly in a few iterations. An ICM (Iterated Conditional Modes) algorithm [1] is used for this purpose.

2.3 Automatic Initialization Process

A major improvement on the original initialization process has been carried out. Different alternatives to obtain the Gaussian distribution parameters (mean and standard deviation) as much unsupervised as possible have been proposed, thus avoiding to empirically check the background and lesion histograms in order to manually fit the best Gaussian distribution to them. Actually, this lack of automatization in the original proposal of Xiao et al. prevented to properly evaluate that proposal since only one patient and one synthetic image were used, while the current proposal has been evaluated on several images from different patients. The initialization process starts by analyzing the intensity histogram of thee image and how the different Gaussian distributions fit to the lesion and background. Fig. 2 shows the histogram of a whole US image which almost fits perfectly to a Gaussian distribution. As far as we have experimented, always the histogram of the lesion is overlapped with the histogram of the background information. It is clear that the automatic initialization of the lesion distribution is not an easy task to do using only intensity information. Hence, we decided to include local spatial information of the lesion.

Our proposal reduces the user interaction from the empirical choice of parameters to one-click. Thus, the user marks the approximate location of the lesion and the method opens a small window to obtain lesion information and a larger one for the background information. Once the lesion and background information are extracted, the histograms of both regions are calculated and

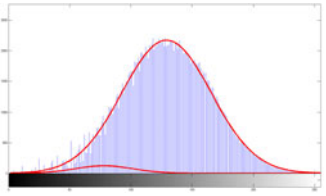


Fig. 2. Gaussian distributions of the US images. The larger Gaussian corresponds to the background and the smaller to the lesion. Both plotted over the image histogram.

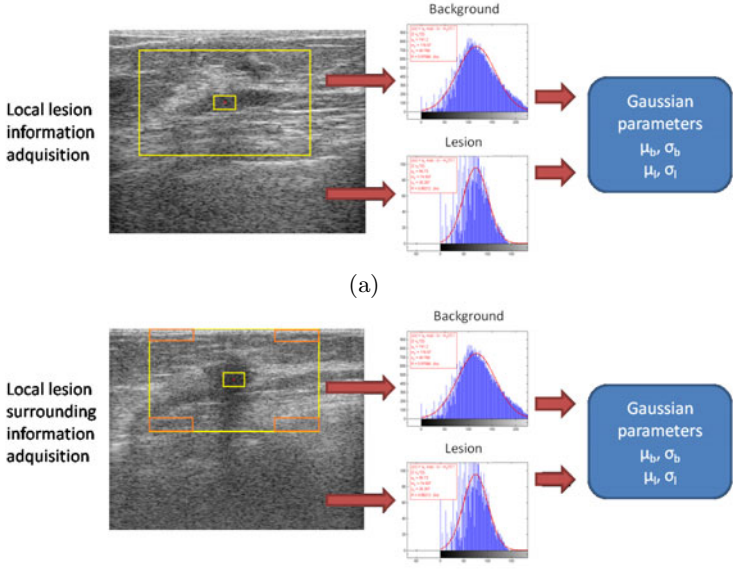


Fig. 3. (a) Local lesion information acquisition. The central rectangle contains the pixel values for the lesion description and the exterior rectangle contains the background information. (b) Local lesion surrounding information acquisition. The exterior orange rectangles contain the background information.

the best Gaussian distributions are properly fitted to them (see Fig.3(a)). Two different approaches have been extracted from this initialization proposal: Local Lesion Information Global (LLIG) which applies the method to the entire image and Local Lesion Information Partial (LLIP) which crop the image using the background window and applying the segmentation process to this partial image. By cropping the image, the dark regions near image limits are avoided. However, in these proposals the background window could include part of the lesion information so, in order to properly fit the Gaussian distribution to the background histogram, we also propose Local Lesion Surrounding Information (LLSI) which gets histogram information from four small windows surrounding the lesion (see Fig.3(b)). In total, we have evaluated 3 different initialization proposals and compared them to the original method initialization.

Regarding the final segmentation results, in some cases the method incorrectly classifies parts of the background as a lesion tissue, mostly in the image borders. In contrast with Xiao et al. proposal, a post-processing step is performed, consisting in deleting such regions incorporating information from radiologist knowledge such as the fact that usually the lesion is located centered on the top part of the US image as suggested by Madabhushi and Metaxas in [8](see Fig. 4(e)).

3 Experiments and Results

In order to perform the experiments for this work, a data-set of breast B-mode US images has been collected from the Churchill Hospital of Oxford Radcliffe Hospitals NHS Trust composed by 56 images acquired from 24 different patients, formed by 31 DCIS+IDC (Ductal Carcinoma In Situ + Invasive Ductal Carcinoma), 5 IDC, 6 DCIS, 6 Fibroadenoma, 2 Fibrosis, 1 Cyst, and 5 Mucinous Carcinoma.

3.1 Ground Truth

The limited availability of the radiologists made us impossible to have the entire data-set segmented by an expert. Because of that, an extra experiment was performed comparing a subset of 15 segmentations between a radiologist and a non-expert (i.e. biomedical engineer). The experiment obtained a result of 0.7426 for Dice Similarity Coefficient (DSC) and 0.8462 for Area Overlap measures which indicate a high degree of agreement between the non-expert and the radiologist.

3.2 Results

Fig. 4 shows an example of graphical segmentation results for different images from the data-set. Fig. 4(a) shows the original US breast image. The estimated distortion field and the restored image are shown in Fig. 4(b) and (c) respectively. Fig. 4(d) shows the segmentation result and Fig. 4(e) the final result after the post-processing step. Fig. 5 shows the segmentation results using Dice Similarity Coefficient (DSC) and Area Overlap (AO) values for all the images. Analyzing these diagrams for both measures we can see that all the median values of the results for each method are within the confidence interval of the other methods. This means that there is not significative differences between methods and it is not possible to assure that one method obtained better results than another. Note that missegmented images are represented as outliers in the diagram.

Table 1 shows the mean values for sensitivity, specificity, PPV, area overlap and DSC measures. To compare properly all methodologies, all measures were obtained descanting the missegmented images by the original method. Again, there are not significative differences which shows the validity of the proposed initialization. Note that LLIG proposal obtained similar results compared the original implementation while there is not a significant increase of the number of missegmented images.

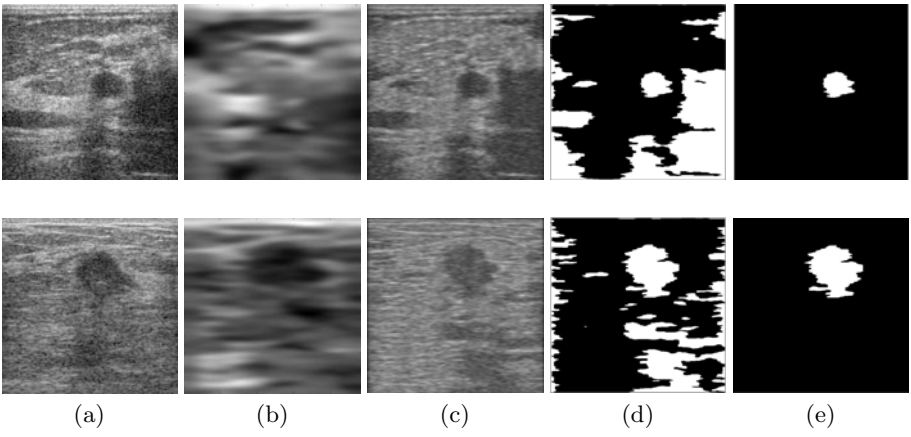


Fig. 4. Lesion segmentation results. (a) original image, (b) estimated distortion field, (c) corrected image, (d) segmentation result applying LLIG and (e) final segmentation after the segmentation refinement.

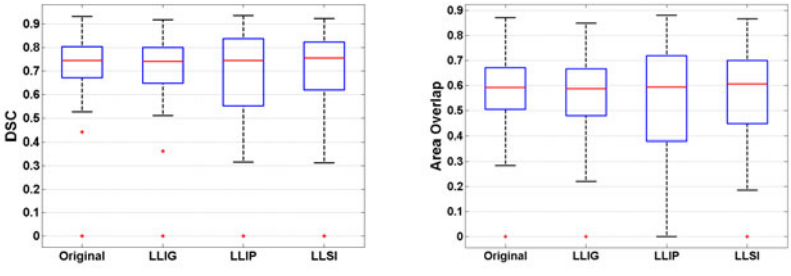


Fig. 5. Box diagrams of DSC and Area Overlap measures for all the methods

Table 1. Comparison between measure means obtained for each method

Method	Sensitivity	Specificity	PPV	Area Overlap	DSC	Missegmented Images
Original	0.8978	0.7577	0.6723	0.6110	0.7544	8/56
LLIG	0.8701	0.7589	0.6757	0.5996	0.7437	11/56
LLIP	0.8035	0.6706	0.5946	0.5538	0.6696	15/56
LLSI	0.7889	0.7016	0.6432	0.5848	0.6912	13/56

4 Conclusion

After this study, we have concluded that fully automatic initialization is a hard task using intensity values only. For that reason, we have presented three different approaches where only a one-click interaction is needed. This supposes a high reduction of the user interaction compared to the original approach where a manual selection of the Gaussian parameters was needed. We have compared the obtained results with the original method and we have concluded that all the

three methods are comparable to the original proposal although LLIG method seems to be more robust and accurate taking into account all the measures used. This lead us to remark that we have improved the original method by means of simplifying the initialization process with no significant loss of correctness.

In order to improve the initialization of the method we propose as a future work to include additional information such as classical features (i.e. SIFT, textons, etc.) or information extracted from the elastography.

References

1. Besag, J.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society B 48, 259–302 (1986)
2. Boukerroui, D., Baskurt, A., Noble, J.A., Basset, O.: Segmentation of ultrasound images: multiresolution 2d and 3d algorithm based on global and local statistics. Pattern Recognition 24(4-5), 779–790 (2003)
3. Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., Du, H.: Approaches for automated detection and classification of masses in mammograms. Pattern Recognition 39(4), 646–668 (2006)
4. Gil, F., Méndez, I., Sirgo, A., Llort, G., Blanco, I., Cortés-Funes, H.: Perception of breast cancer risk and surveillance behaviours of women with family history of breast cancer: a brief report on a spanish cohort. Psycho-Oncology 12, 821–827 (2003)
5. Horsch, K., Giger, M.L., Venta, L.A., Vyborny, C.J.: Automatic segmentation of breast lesions on ultrasound. Medical Physics 28(8), 1652–1659 (2001)
6. Huang, Y.L., Jiang, Y.R., Chen, D.R., Moon, W.K.: Level set contouring for breast tumor in sonography. J. Digital Imaging 20(3), 238–247 (2007)
7. von Lavante, E., Noble, J.: Segmentation of breast cancer masses in ultrasound using radio-frequency signal derived parameters and strain estimates. In: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2008, pp. 536–539 (May 14-17, 2008)
8. Madabhushi, A., Metaxas, D.: Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions. IEEE Transactions on Medical Imaging 22(2), 155–169 (2003)
9. Wells III, W.M., Grimson, W., Kikinis, R., Jolesz, F.: Adaptive segmentation of mri data. IEEE Transactions on Medical Imaging 15(4), 429–442 (1996)
10. Xiao, G., Brady, M., Noble, J., Zhang, Y.: Segmentation of ultrasound b-mode images with intensity inhomogeneity correction. IEEE Transactions on Medical Imaging 21(1), 48–57 (2002)
11. Yeh, C.K., Chen, Y.S., Fan, W.C., Liao, Y.Y.: A disk expansion segmentation method for ultrasonic breast lesions. Pattern Recognition 42(5), 596–606 (2009)
12. Zouqi, M., Samarabandu, J.: 2d ultrasound image segmentation using graph cuts and local image features. In: IEEE Symposium on Computational Intelligence for Image Processing, CIIP 2009, pp. 33–40 (March 2009)

Music Score Binarization Based on Domain Knowledge

Telmo Pinto[1], Ana Rebelo[1], Gilson Giraldi[2], and Jaime S. Cardoso[1]

[1] INESC Porto, Faculdade de Engenharia, Universidade do Porto, Portugal
[2] National Laboratory for Scientific Computing, Petrópolis, Brazil
telmotpinto@gmail.com, {arebelo,jaime.cardoso}@inescporto.pt, gilson@lncc.br

Abstract. Image binarization is a common operation in the preprocessing stage in most Optical Music Recognition (OMR) systems. The choice of an appropriate binarization method for handwritten music scores is a difficult problem. Several works have already evaluated the performance of existing binarization processes in diverse applications. However, no goal-directed studies for music sheets documents were carried out. This paper presents a novel binarization method based in the content knowledge of the image. The method only needs the estimation of the staffline thickness and the vertical distance between two stafflines. This information is extracted directly from the gray level music score. The proposed binarization procedure is experimentally compared with several state of the art methods.

Keywords: Computer Vision, Image Processing, Optical Music Recognition.

1 Introduction

Printed documents and handwritten manuscripts deteriorate over time, causing a significant amount of information to be permanently lost. Among such perishable documents, musical scores are especially problematic. Digitization has been commonly used as a possible tool for preservation, offering easy duplications, distribution, and digital processing. However, to transform the paper-based music scores and manuscripts into a machine-readable symbolic format, an Optical Music Recognition (OMR) system is needed.

After the image preprocessing (application of several techniques, e.g. binarization, noise removal, among others, to make the recognition process more robust and efficient), an OMR system can be broadly divided in three principal modules: recognition of musical symbols from a music sheet; reconstruction of the musical information to build a logical description of musical notation; construction of a musical notation model for its representation as a symbolic description of the musical sheet.

The binarization of the music score seldom justifies significant attention, with researchers invariably using some standard binarization procedure, such as the Otsu's method (e.g. [1]). Nonetheless, the development of binarization methods

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 700–708, 2011.

specific to music scores has the potential of showing better performance than the generic counterparts and of leveraging the performance of subsequent operations.

Effective binarization of images should not only use the raw pixel information, but consider the image content as well. Unfortunately, since the binarization procedure is usually the first step of the processing system, there is usually no information available about the image content to assist the binarization procedure. A possible workaround is when we are able to extract content related information from the gray-scale image to guide the binarization procedure.

The recent work [2] on the estimation of the staffline thickness and distance without binarizing the music score, working directly in the gray-scale image, opens the door to such content aware image binarization applied to music scores, which we explore in this work.

1.1 Related Work

Different methods for image binarization have been developed and proposed in the literature. The categorization of existing techniques adopted in this work follows the survey presented in [3].

Global thresholding methods apply one threshold to the entire image. Among these techniques, the Otsu threshold selection is ranked as the best and the fastest global thresholding method [4]. It considers that the image contains two classes of pixels – foreground and background. The algorithm computes the parameter of intensity level T by maximizing the variance between the classes. This procedure considers that any point that presents an intensity equal or greater than T belongs to one class, and all the others are considered part of the other class [5]. Entropy-based methods are also very common in the image segmentation area. In [6], an image thresholding based in Tsallis entropy is proposed. The authors claim that by using Tsallis entropy they can avoid the presence of nonadditive information in some type of images that influence the segmentation operation. Edge information has also been used by several image binarization methods. In [7] the Canny edge detector was adopted: if the boundaries of actual objects in the edge image are considerably complete and closed, the binarized image can be obtained with seed filling inside the boundaries of the objects. Other works encompass similarity measures between the gray-level and the binarized images such as the minimization of fuzziness shapes [8] or the smoothed histogram via Gaussians to detect peaks and valleys [9].

In adaptive binarization methods a threshold is assigned to each pixel using local information from the image. The Bernsen's local thresholding method [10] computes the local minimum and maximum for a neighborhood around each pixel level, and uses the mean of the two as the threshold for the pixel in consideration. Niblack [11] suggested using as local information for the threshold decision the mean and the standard deviation of the pixel's neighbourhood. The works of [12,13] applied this technique to their OMR procedures.

Although there is goal directed evaluation of binarization methods, there is no goal-directed design of binarization methods specific for certain class of images. Existing methods are generic in the sense that are agnostic of the content of the image.

2 Robust Estimation of Staffline Thickness and Spacing in the Gray-Scale Domain

The conventional estimation of the staffline thickness and spacing assumes the run-length encoding (RLE) of each column of the *binary* music score. In this representation, the most common black-run is likely to represent the staffline thickness and the most common white-run is likely to represent the staffline spacing. Even in music scores with different staff sizes, there will be prominent peaks at the most frequent thickness and spacing. These estimates are also immune to severe rotation of the image.

In [2] the authors suggest to estimate directly the *sum* of the staffline thickness and spacing, hereafter termed `line_thickness+spacing`, since this can be robustly estimated by finding the most common sum of two consecutive vertical runs (either black run followed by white run or the reverse).

Moreover, instead of computing the most frequent peak in the histogram of the runs for a binarized image (binarized with a state-of-the-art binarization method), the authors propose to compute the histogram of the runs for 'every' possible binary image, by accumulating the runs' frequency when varying the binarization threshold from the lowest to the highest possible values. This procedure of computing the reference length `line_thickness+spacing` without assuming any binarization threshold, allows the extraction of important information directly from the gray-scale image. We propose now to use this information to guide the binarization procedure.

3 Content Aware Music Score Binarization

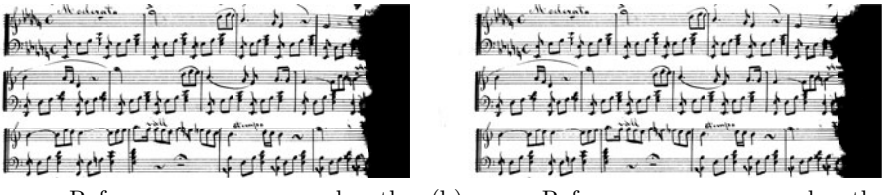
As stated in the introduction, an OMR system typically encompasses, in one of its first steps, the detection of the stafflines to facilitate the subsequent operations. A binarization method designed to maximize the number of the pairs of consecutive runs summing `line_thickness+spacing` (the peak computed over the gray-level image) will likely maximize the quality of the binarized lines. However, the direct maximization of the count of pairs of consecutive runs summing `line_thickness+spacing` could lead to a threshold value producing many, 'noisy', runs, and as a side effect, many runs at `line_thickness+spacing`. The use of relative histograms is also prone to problems since now one may end up choosing a threshold with a very low absolute count of runs in `line_thickness+spacing` but that, by chance, is the highest relative count.

Therefore, we restrict the candidate thresholds to those producing a histogram of runs with the mode at `line_thickness+spacing`. If no threshold is found with this condition (note that even if the integration over all thresholds does have a mode at `line_thickness+spacing`, it is possible that no individual threshold produces a histogram with mode at `line_thickness+spacing`), we consider the minimum integer i for which there are threshold values with histogram mode at `line_thickness+spacing` $\pm i$. From the set of candidate thresholds, the proposed binarization method for music scores simply selects the threshold that maximizes the count of pairs of consecutive runs on the mode.

3.1 Using Other Reference Lengths to Guide the Binarization

The same rationale used to motivate the estimation of the sum of pair of consecutive lengths, can be used to work with sets of three or more consecutive runs. However, two problems arise when proceeding that way: there is the underlying assumption that each staff have enough lines to give meaning to the consecutive runs and one starts getting less and less values to accumulate in the histogram, potentially leading to less accurate estimations.

A potentially interesting balance is estimating the sum of two times the line thickness plus the spacing, `line_2thickness+spacing`, by working with the frequencies of triplets (black run, white run, black run). This only assumes that each staff has at least two lines, but does impact the number of accumulated values, roughly halving it. The proposed content aware binarization method does not suffer any adaptation, besides the change of the reference length, `line_thickness+spacing` by `line_2thickness+spacing`. Further on in this paper we will compare the two options. In Fig. 1 we illustrate the results obtained with the proposed approach, using the two aforementioned reference lengths. One can observe that the resulting stafflines have good quality, with minor differences between the two results. Nevertheless, the original music score in this



(a) Reference length: `line_thickness+spacing` (b) Reference length: `line_2thickness+spacing`

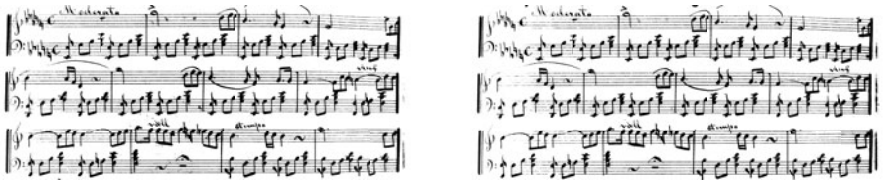
Fig. 1. Result of binarizing an example of a music score

particular example is not correctly binarized with a global threshold. The digitalization of bound documents, such as books, either performed by flatbed scanners or digital cameras often yields images that exhibit a gradient-like distortion in the average colour in the region close to the book spine. In these cases, adaptive methods can show better performance.

3.2 Adaptive Content Aware Music Score Binarization

Despite having been presented as a global thresholding method and having been applied it to the whole image, nothing prevents the application of the just developed ideas to a sampling window around a pixel p, effectively converting the proposed method to a local method.

As with other adaptive methods, the size of the sampling window is a key parameter. With our approach, the sampling window should be big enough to accumulate enough information (runs) to provide a proper solution. Since the



(a) Reference length: `line_thickness+spacing` (b) Reference length: `line_2thickness+spacing`

Fig. 2. Result of binarizing an example of a music score with the adaptive method

typical distortions in this kind of documents are vertically oriented, the local threshold should be constant along a column of the image. Therefore we suggest computing a single threshold per column, using as window a vertical strip with height equal to the height of the image and width defined by the user. In order to reduce the computational cost, the threshold value is calculated by interpolating the values on a set of sampled columns. In Fig. 2 we illustrate the results obtained with the proposed approach, using a window width and step size of 2% of the width of the image, and cubic polynomial interpolation. In this example, the adaptive method using the `line_thickness+spacing` reference length provided the best results, with a better staffline definition.

4 Experimental Evaluation

In order to support the comparison between different binarization procedures, quantitative evaluation methods were run on a dataset of music scores. This dataset is composed of 65 handwritten scores, from 6 different authors. All the scores in the dataset were reduced to gray level information. The methods chosen try to encompass different categories of thresholding operations. Some of the algorithms tested required the input of different parameters that were obtained by experimental testing. For Sahoo's Correlation method: $Q_1 = 0.4$, $Q_2 = 1$, $Q_3 = 3$; for Tsallis entropy method: $\alpha = 2$.

For global thresholding processes, three different approaches were taken for evaluation: Difference from Reference Threshold (DRT); Misclassification Error (ME); comparison between results of staff finder algorithms applied to each binarized image. For the first method (DRT), five people were asked to choose the best possible threshold for each image. The average value of these chosen thresholds was compared to the resulting threshold value of each binarization. Ground truth versions of some scores from the dataset were also used as a comparison procedure. The Misclassification Error was defined as the difference rate between these ground truth images and the resulting images from each binarization as:

$$ME = 1 - \frac{\#(B_{bin} \cap B_{gt}) + \#(F_{bin} \cap F_{gt})}{\#B_{bin} + \#F_{bin}} \quad (1)$$

Table 1. Test results for various global thresholding methods, using different evaluations: difference from reference thresholds values, misclassification error (in percentage), staff detection error rates for missed and false staffs (in percentage) with Stable Path and Dalitz

	Huang [8]	Khashman [15]	Kapur[16]	Sahoo [17]	Tsai [9]	Tsallis [6]	Otsu [5]	BLIST pairs	BLIST triplets
DRT: avg	48	33	50	50	29	50	19	19	29
ME: avg %	6.2	3.8	4.9	7.6	4.7	5.7	4.6	4.8	5.1
SP False: avg(std) %	2.6(5.5)	2.1(4.0)	1.4(3.4)	3.5(10.2)	2.1(4.1)	3.3(7.4)	2.0(3.4)	1.3(2.7)	1.7(3.7)
SP Missed: avg(std) %	18.0(34.5)	30.2(42.3)	27.1(42.3)	25.7(40.1)	17.0(30.3)	21.0(36.4)	8.6(20.5)	1.5(2.8)	2.8(6.3)
Dal False: avg(std) %	21.6(41.1)	3.2(7.8)	1.8(4.2)	5.4(25.6)	4.4(8.1)	2.4(6.0)	3.6(5.4)	3.2(5.0)	3.8(6.5)
Dal Missed: avg(std) %	39.6(36.9)	32.7(41.4)	31.2(42.0)	35.4(42.5)	25.4(35.0)	31.5(41.8)	18.8(31.0)	14.8(28.6)	14.9(27.4)

In Eq. (1) B_{bin} and F_{bin} represent the background and foreground pixels of the binarization being tested, and B_{gt} and F_{gt} the background and foreground pixels in the reference ground truth image, respectivelly. # is the cardinality, or more precisely, the number of elements in a specific set. Since the manual binarization of the images is very time consuming, this evaluation method was applied only to ten scores chosen randomly from the complete dataset. The third technique is based on the results of staff finding algorithms applied to the binarized scores. Comparing these results, one can detect the method that will most likely produce the best outputs in the next image processing steps. The staff finding algorithms applied were Stable Path [1] and Dalitz [14]. Table 1 summarizes all the results. Both versions of the Binarization based in LIne Spacing and Thickness (BLIST) method, proposed in this article, performed above average. Even so, the version that uses the pairs of runs instead of the triplets did better in the tests. This version will be considered on all the following comparisons. Entropy based binarizations and Khashman's algorithms got fairly similar results to each other. Huang and Tsai managed to top these results, with acceptable line detection rates and Misclassification Error. There are, however, two binarization techniques that get consistently better results than the others: Otsu's Method and BLIST method. The only major difference is the higher missed staff detection rate for the Otsu's algorithm.

Global methods can generally produce good outputs. Even so, for some of the scores, like those with heterogenous light distribution resulting from the digitalization process, there is no perfect threshold. In these scores, it is not possible to find a single threshold value that produces both the presence of perfectly connected staves and no occlusion of data with noise. Although staves can be correctly found in global thresholding procedures, adaptive methods can produce results with little or no loss of information.

The adaptive version of the BLIST method was implemented as described previously. The window width used was a fixed percentage of the total image width. The interpolation of the threshold values obtained was generated with a third degree polynomial regression. Otsu's method, having good results among global methods was also implemented as adaptive, using the same reasoning. Most adaptive algorithms tested required the input of some parameters, determined experimentally. For Bernsen: window size= 10x10 px, minimum difference

in contrast = 20; for Niblack: window size = 200x200 px, k = -1; for Otsu Adaptive: window width= 2% image width; for Adaptive BLIST: window width = 2% of image width.

For the adaptive binarizations, the Misclassification Errors were all very similar. A further analysis was conducted, still based on ground truths of ten scores. Two new error rates are presented: the Missed Object Pixel rate and the False Object Pixel, dealing with loss in object pixels and excess noise, respectively.

$$MOPx = \frac{\#F_{gt} - \#(F_{bin} \cap F_{gt})}{\#F_{gt}}, \; FOPx = \frac{\#F_{bin} - \#(F_{bin} \cap F_{gt})}{\#F_{bin}}$$

Table 2. Test results for various local thresholding methods, using different evaluations (in percentage): misclassification error, Missed Object Pixels, False Object Pixels, staff detection error rates for missed and false staffs with Stable Path and Dalitz.

	Bernsen [10]	**Chen [7]**	**Ad BLIST**	**Niblack [11]**	**Ad Otsu**	**YB [18]**
ME: avg %	4.3	3.2	4.2	4.3	4.2	3.5
MOPx: avg %	24.6	22.5	15.6	22.5	21.7	12.4
FOPx: avg %	13.2	4.3	18.5	13.8	16.5	14.7
SP False: avg(std) %	1.3(3.0)	9.9(9.7)	2.1(5.6)	3.2(4.6)	2.7(5.9)	4.2(7.6)
SP Missed: avg(std) %	1.9(4.4)	33.0(32.8)	2.3(5.5)	14.2(23.2)	10.7(23.6)	7.9(13.8)
Dal False: avg(std) %	3.9(12.9)	3.4(5.6)	3.8(6.2)	3.1(5.1)	3.2(4.9)	3.5(6.1)
Dal Missed: avg(std) %	9.0(16.2)	17.3(27.0)	8.4(14.6)	10.2(14.1)	10.7(18.9)	7.7(10.1)

Ad BLIST and YB show the lowest MOPx, meaning these are the methods that find most of the correct pixels, which translates into lower missed staves rates. Even so, Ad BLIST also has a FOPx rate slightly higher than the other methods. This higher noise also translates into a slightly higher false staves rate with Dalitz method. Bernsen's binarizations, although presenting the highest missed pixel rate, seem to perform well in the staff finding steps, having both the lowest missed and false staves rates.

5 Conclusion

Many binarization techniques have been proposed for digital images in the past. These methods can be applied to music scores with different rates of success although none is based on the knowledge of the content of a music score. The main contribution of this work is the introduction of a content aware binarization method for music scores. The method, based on the knowledge of the staff line thickness and spacing, extracted directly from the gray-level image, tries to find the threshold that maximizes the information content of the image, as measured by these values. We then introduced an adaptive version of our method. The basic idea of using knowledge from the image to improve the binarization operation, may apply in other areas of document image analysis, or in general image analysis.

Acknowledgments. This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) - Portugal through projects PTDC/EIA/71225/2006 and SFRH/BD/60359/2009. The authors thank Prof. Paulo S. S. Rodrigues for providing the Matlab implementation of the method based in Tsallis entropy.

References

1. Cardoso, J.S., Capela, A., Rebelo, A., Guedes, C., da Costa, J.P.: Staff detection with stable paths. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 1134–1139 (2009)
2. Cardoso, J.S., Rebelo, A.: Robust staffline thickness and distance estimation in binary and gray-level music scores. In: International Conference on Pattern Recognition, pp. 1856–1859 (2010)
3. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13(1), 146–165 (2004)
4. Trier, O.D., Taxt, T.: Evaluation of binarization methods for document images. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(3), 312–315 (1995)
5. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics 9(1), 62–66 (1979)
6. de Albuquerque, M.P., Esquef, I.A., Mello, A.R.G.: Image thresholding using tsallis entropy. Pattern Recognition Letters 25(9), 1059–1065 (2004)
7. Chen, Q., sen Sun, Q., Heng, P.A., shen Xia, D.: A double-threshold image binarization method based on edge detector. Pattern Recognition 41(4), 1254–1267 (2008)
8. Huang, L.K., Wang, M.J.J.: Image thresholding by minimizing the measures of fuzziness. Pattern Recognition 28(1), 41–51 (1995)
9. Tsai, D.M.: A fast thresholding selection procedure for multimodal and unimodal histograms. Pattern Recognition Letters 16(6), 653–666 (1995)
10. Bernsen, J.: Dynamic thresholding of grey-level images. In: Bieniecki, W., Grabowski, S. (eds.) Multi-pass approach to adaptive thresholding based image segmentation. Proceedings of the 8th. International IEEE Conference CADSM (2005)
11. Niblack, W.: An introduction to digital image processing (1986). In: Leedham, G., Yan, C., Takru, K., Tan, J.H.N., Mian, L. (eds.) Comparison of Some Thresholding Algorithms for Text/Background Segmentation in Difficult Document Images. Proceedings of the Seventh International Conference on Document Analysis and Recognition (2003)
12. Fornés, A., Lladós, J., Sánchez, G., Bunke, H.: Writer identification in old handwritten music scores. In: DAS 2008: Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems, pp. 347–353. IEEE Computer Society, Washington, DC, USA (2008)
13. Fornés, A., Lladós, J., Sánchez, G., Bunke, H.: On the use of textural features for writer identification in old handwritten music scores. In: ICDAR 2009: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition, pp. 996–1000. IEEE Computer Society, Washington, DC, USA (2009)
14. Dalitz, C., Droettboom, M., Czerwinski, B., Fujigana, I.: A comparative study of staff removal algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 753–766 (2008)

15. Khashman, A., Sekeroglu, B.: A novel thresholding method for text separation and document enhancement. In: Proceedings of the 11th Panhellenic Conference on Informatics (PCI 2007) (May 2007)
16. Kapur, J.N., Sahoo, P.K., Wong, A.K.C.: A new method for gray-level picture thresholding using the entropy of the histogram. Computer Vision, Graphics, and Image Processing 29(3), 273–285 (1985)
17. Sahoo, P.K., Wilkins, C., Yeager, J.: Threshold selection using renyi's entropy. Pattern Recognition 30(1), 71–84 (1997)
18. Yanowitz, S., Bruckstein, A.: A new method for image segmentation. Computer Vision, Graphics, and Image Processing 46, 82–95 (1989)

Identifying Potentially Cancerous Tissues in Chromoendoscopy Images

Farhan Riaz, Fernando Vilarino, Mario Dinis Ribeiro, and Miguel Coimbra

Instituto de Telecomunicacoes, Universidade do Porto, Portugal
Computer Vision Center, Universitate Autnoma de Barcelona, Spain
Instituto Portugues Oncologia, Porto, Portugal
farhan.riaz@dcc.fc.up.pt

Abstract. The dynamics of image acquisition conditions for gastroenterology imaging scenarios pose novel challenges for automatic computer assisted decision systems. Such systems should have the ability to mimic the tissue characterization of the physicians. In this paper, our objective is to compare some feature extraction methods to classify a Chromoendoscopy image into two different classes: Normal and Potentially cancerous. Results show that LoG filters generally give best classification accuracy among the other feature extraction methods considered.

Keywords: Endoscopy, Computer Assisted Diagnosis, Gradient.

1 Introduction

Gastric cancer is a major cause of death worldwide. From a total of 57 million deaths worldwide in 2004, cancer accounts for 7.4 million (or 13%) of all deaths (World Health Organization, www.euro.who.int). Gastroenterology Imaging (GI) is today an essential tool for clinicians to detect cancer effectively. This is a rapidly evolving technological area with novel imaging devices such as Capsule, Narrow-Band Imaging (NBI) or High-Definition Endoscopy. In this paper, we are focused on classifying images from one of the most widely used gastrointestinal imaging modality: Chromoendoscopy (CH). It is based on using the full visible spectrum of light, accompanied by staining of the GI tissue with a dye, such as methylene blue to enhance the gastric mucosa in the images, thus helping in classifying the images as normal, pre-cancerous or cancerous. The clinical support of our work is provided by Dinis-Ribeiro classification proposal [1] which underlines the features which are supportive for classifying the images. Owing to the difficulties in the manual diagnosis systems and the training of clinicians for these novel imaging modalities, Computer Assisted Decision (CAD) systems are increasingly desirable to detect Gastrointestinal cancer effectively.

Our objective in this paper is to classify the CH images as being 'normal' (Group A) or 'potentially cancerous' (Group B). Our previous work shows the dominance of texture features in such images hinting at proficiency of methods based texture feature extraction for classifying such images (paper submitted in a journal for peer review). Many texture recognition methods are available in the

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 709–716, 2011.

literature, which can be mainly divided into four different categories: statistical methods, model based methods, structural methods and filter based methods [2]. Most techniques based on the former methods are more suitable for highly regular, semi-regular or micro-textures. These problems are mitigated by filter-based methods due their diversity provided by their ability to combine micro- and macro- texture features, thus giving a richer description of the images. In this paper, we focus on the extraction of interest points in the images followed by the use of Edge Histograms for finally classifying our images. The outline of the paper is as follows: We discuss the dataset (Section 2), followed by our methodology of feature extraction (Section 3). Afterwards, we describe our experimental setup (Section 4), followed by discussion and future work (Section 5).

2 Materials

The CH images were obtained using an Olympus GIF-H180 endoscope at the Portuguese Institute of Oncology (IPO) Porto, Portugal during routine clinical work. Optical characteristics of this endoscope include 140 field of view and four way angulation (210 up, 90 down and 100 right/left). The endoscopic videos were recorded on tapes using a Digital Video (DV) recorder while performing real endoscopic examinations. Around 4 hours of video (360000 frames) were analyzed and 176 images were initially selected given their clinical relevance. This was first determined by pre-selecting images that were annotated during the procedure by the clinician performing the exam, and later each image was individually selected for this study by an expert clinician. Images were saved as graphics files of type PNG (Portable Network Graphics) with a resolution of 518x481. Two clinicians classified these images into three groups, following Dinis-Ribeiro's classification proposal [1], manually segmenting the image region that led them to this conclusion and labeling their choice with a confidence value (high or low confidence). Our resulting gold-standard not only uses regions where there was high-confidence annotations and agreement between the two specialists (135 images), but a second analysis was carefully performed for images where both doctors were confident and obtained different results. This typically showed that they selected different regions in an image that corresponded to different classifications. The final number of high-confidence image regions used in this study was thus increased by 41 to a total of 176. Finally, a careful analysis of images was carried out to remove the images belonging to the same patients, giving us a dataset of 142 images distributed as 31.6 % (45 images) belonging to Group I, 54.9 % (78 images) belonging to Group II and 13.3 % (19 images) to Group III.

3 Methods

3.1 Interest Point Detection

Conventionally, the spatial characteristics of an image are usually smooth over a particular neighborhood which makes a lot of data redundant. This enforces the

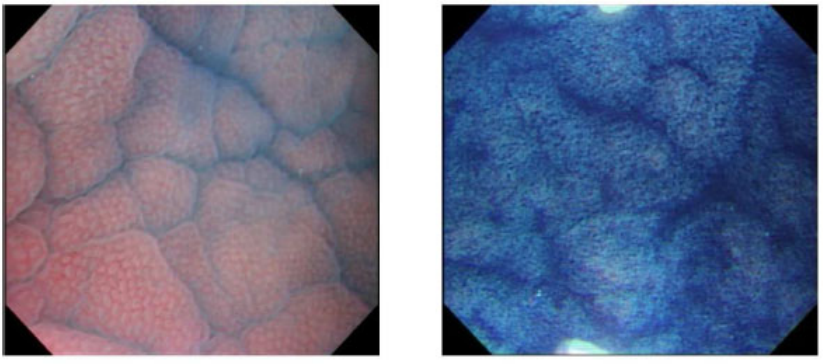


Fig. 1. Typical normal and cancerous images. A distinct clear observation is the richness of texture for cancerous image, resulting from distortion of pattern.

need to discard the redundant data and take into account only a set of points which contain the salient features in the images. Our CH images show distinctive information between normal and potentially cancerous images, especially in terms of the distribution of interest points. For normal images this distribution is expected to be sparse in an area of annotation due to smooth texture whereas for other images we expect their dense distribution due to high texture. Also, we want to mimic the manual annotation of the physician by selecting some interest points in the images, which are representatives of important visual characteristics of the images. This motivates us to extract interest points in the images and process them to obtain salient features in images. We use three different methods for this purpose:

Harris operator: They are widely used for corner detection. For a 2-dimensional image, the Harris matrix is constructed which is a representative of the image derivatives. The magnitude of Eigen values of this matrix is a representative of potential corners in the image [3].
FAST operator: This operator considers the pixels under a Bresenham circle of a particular radius around an interest point [4] . The original detector classifies a point as a corner if there exists a set of contiguous pixels in the circle which are all brighter than the intensity of the candidate pixel plus a threshold or darker than the intensity of candidate pixel minus a threshold.
SIFT operator: This operator identifies scale-invariant features using a filtering approach through different stages [5]. In the first stage, key location in the scale space are selected by looking for location that are maxima or minima of different of Gaussian (DoG) function. The next steps consist of resampling the image and repeating the same procedure. The number of resampling stages conform the scale-space. Maxima and minima of this function are determined by comparing each pixel in a stage to its neighbors. It eventually detects key points which are translation, rotation and scale invariant.

Experiments show that the density of interest points in an annotated region could possibly be an indicator of richness of the texture which in turn can help us in characterizing the image into their respective classes.

3.2 Feature Extraction

The perception of salient points in the images is greatly affected by the amount and intensity of gradients. Even under varying lighting conditions, the most robust way to characterize an image could be the distribution and intensity of gradients in the images. It is one of the most fundamental tasks in computer vision. A variety of filters can be used this purpose, we used the following:

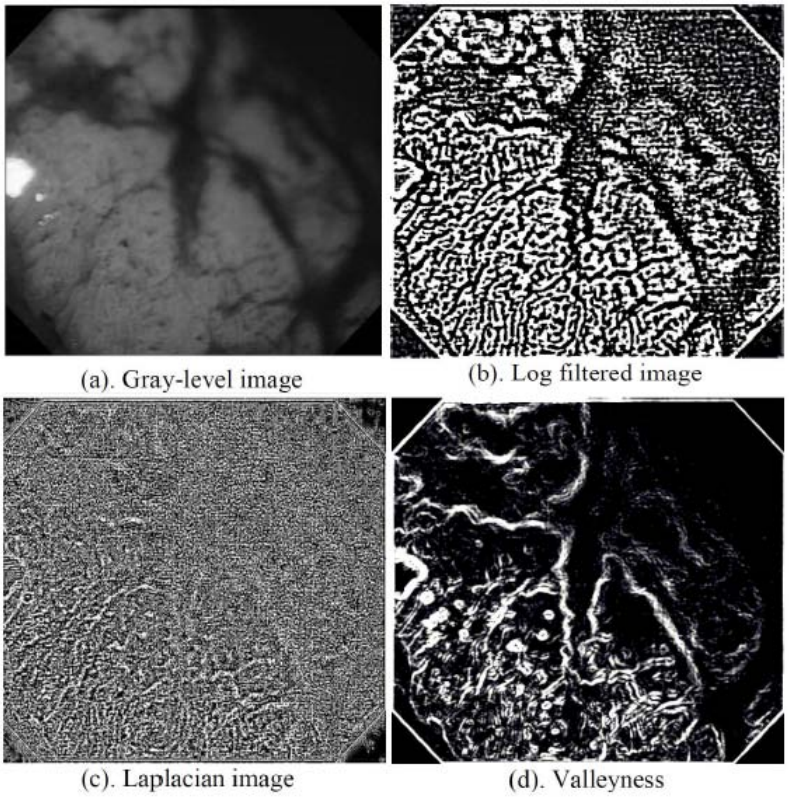


Fig. 2. Image after filtering using log, laplacian and valleyness filters

Laplacian filter: Laplacian is a differential operator given by divergence of the gradient of the function on the Euclidean space [6]. In image processing, this operator is used for various tasks such as blob detection or edge detection.
Log filter: LoG is an acronym for Laplacian of Gaussian. It is one of the most common blob detectors. It combines the use of two operators (Laplacian and

Gaussian) to improve blob detection [6]. For an input image, convolution with a Gaussian filter is done and then the Laplacian operator is applied. This operator usually gives strong positive responses for dark and strong negatives responses for bright blobs.
Valleyness filter: Ridges and valleys are useful geometric features for image analysis [7]. Researchers have characterized the mathematical model of the flow of water on the earth's surface based on the presence of ridges and valleys in the images. These models have lately been used for feature extraction and segmentation of the images by a well known method known as watershed segmentation. We use the valleyness operator to extract the ridge and valley features from the images.

3.3 Classification

The feature extraction is followed by a generation of histograms of underlying features. Those histograms are then used for classification of images into their respective classes. We classify these histograms using Support Vector Machines (linear kernel, one vs. one classification, sequential optimization) [8]. The objective of the classification task is to classify each image into two possible categories: Group A, which are images of patients which have no signs of cancer and Group B, which are images of patients who either are at initial stages of cancer or those who are suffering from cancer.

4 Experimental Setup

4.1 Parameter Setting

The first task while formulating the experimental setup is parameter setting of the methods. For this purpose, we divided the dataset into training and testing set. From a total of 142 images in the dataset, we selected 15 training images, which were used to tune the parameters of the methods used in this paper. Rest of the 127 images were used for testing.
Parameter setting for interest point detectors: The physicians provided us with ground truth data containing clinically relevant manual annotations (Region of Interest - ROI) and the respective labels for each image (manually classified). The objective of this step is to set the parameters of interest point detectors to ensure that all the interest points lie inside the ROI. We used our training set and their corresponding ROIs to tune the parameters such that all (or most) of the points lie inside the ROI. Experiments show that this is an important step as a change in the parameters of the methods can potentially result in a lot of points in clinically uninteresting regions.
Parameter setting for feature extraction methods: In the above feature extraction methods, only two of them (log filters and valleyness) take a parameter as input - the standard deviation of the Gaussian functions used. Experiments show that selection of a smaller value gives higher classification accuracy, we therefore used a small value (sigma = 0.5).

4.2 Data Normalization

To obtain meaningful and comparable results, we need to normalize the data. It involves eliminating the outliers, normalization of histogram and defining the centers of bins, which are to be used for feature extraction. For each of the feature extraction methodology, we sort the features ascendingly and discard 5% of the features (2.5% of the features at the lower extrema and 2.5% at the upper) to cater for the outliers. Mean and standard deviation from the rest of the data are extracted and the remaining data is normalized to zero mean and unit variance. Afterwards, 15 equally spaced bins between extrema of the rest of the data are created. Our experiments show that using a smaller number of bins depreciates the performance of the system, whereas using a higher number of bins does not have a change on the average output any further. For every novel image, the feature vectors are calculated and the outliers are discarded using the same procedure and the remaining data is normalized by the mean and standard deviation of the training set. Afterwards, the representative histogram of an image is generated using the bins which were calculated using the training set and the final feature vectors are obtained.

4.3 Classification Results

For analyzing the overall performance of the system, leave-one-out cross validation is used on 127 images which were saved for testing. This is because due to the lack of sufficient amount of data, we should do adequate training of the classifier to get consistent results. A visual illustration of our classification results is shown in Figure 3. The most notable observation in the graph is a higher discrimination power of gray level histograms for classifying CH images. This effectively means the classification is a function of image brightness. A deeper analysis (Fig. 4a) of histogram distribution using gray level values with the application of manual annotation mask shows that Group A images have an overall brighter intensity as compared with Group B images. We suspect that this happens because of very high texture generally for Group B images which tends to make the images to appear darker as compared with normal images. The need of feature detection is emphasized by the fact that when using full images, we have higher error rates for classification of both, Group A and Group B images (Fig. 4b). When feature detection is used, low error rates are obtained thus emphasizing the need to pre-process the image using one of the feature detection methods.

Relative conclusion does not change by altering feature detection methods, therefore results with only Harris corners is presented (Fig. 4b). Another important observation is that although gray level histogram performs reasonably well for CH images, LoG filter tend to perform better which can be observed by visualizing lower error rates for LoG filters for both Group A and Group B images.

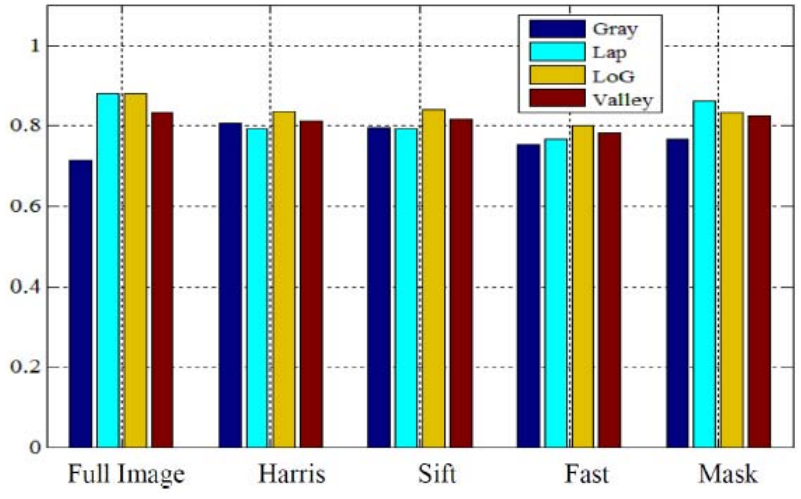


Fig. 3. Classification rates achieved for different feature extraction methods

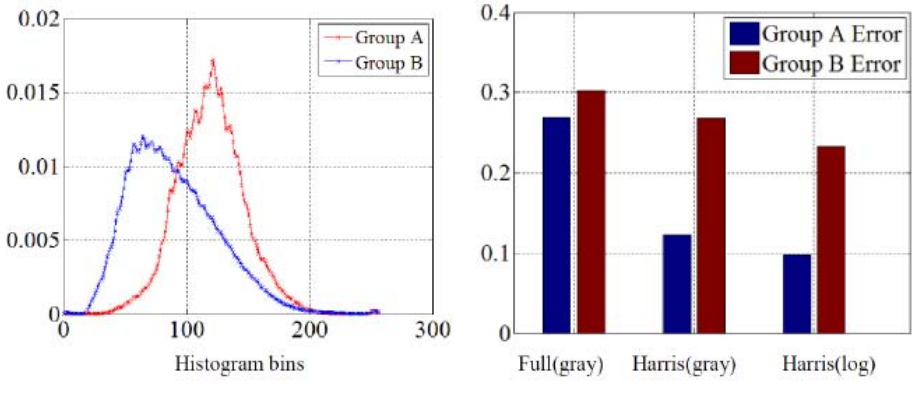


Fig. 4. Detailed analysis of classification performance. (a). Significance of gray level histograms (b). Advantages of feature detection.

5 Discussion

In this paper, we have used gradient based features to characterize Chromoendoscopy images into two different classes: Normal (Group A) and potentially cancerous (Group B). An important observation from our experiments was that the potential of gray level histograms of the images give unexpectedly good results. This is attributed to the richness of texture of images, which tends to darken the image for Group B patterns. Importance of feature detection is highlighted by an analysis of classification errors for Group A and Group B images.

When using full image for feature extraction, higher error rates are achieved however using any on the feature detection methods comparatively improved the results. This emphasizes the need to pre-process the images to select a few interest points, which improve the final classification of the images. Future work hints at studying the anatomy of the images in detail and trying to find ways to extract invariant features, which are expected to give better performance for our dataset. An interesting study would be the effects of reduction of simplification by classifying the images into 3 classes (normal, pre-cancerous and cancerous).

Acknowledgments. We would like to thank Portuguese Institute of Oncology, Porto for providing us clinical support for this work. This work was financially supported in parts by FCT (Fundao para Cincia e a Tecnologia) individual grant SFRH / BD / 45066 / 2008, project of Portuguese Government PTDC/EIA-CCO/109982/2009 and the projects of the Spanish Government COLON-QA (TIN2009-10435) and MIPRCV (CSD2007-00018). I am also very thankful to Jorge Bernal from Computer Vision Center, Barcelona for his technical support in carrying out this research. I would also like to thank the Computer Vision Center, Barcelona for providing me with logistic support for this research.

References

1. Ribeiro, M.D.: Clinical, Endoscopic and Laboratorial Assessment of Patients with Associated Lesions to Gastric Adenocarcinoma. Faculdade de Medicina, Universidade do Porto, PhD Thesis (2005)
2. Selvan, S., Ramakrishnan, S.: SVD-based modeling for texture classification using wavelets transformation. IEEE Trans. Image Process. 16(11), 2688–2696 (2007)
3. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: Avley Vision Conference, pp. 147–152 (1988)
4. Rosten, E., et al.: A machine learning approach to corner detection. IEEE Trans. Pattern Analysis and Machine Intelligence 32(1), 105–119 (2009)
5. Lowe, D. G.: Object recognition from local scale-invariant features. In: International Conference on Computer Vision, Greece. Published by IEEE Signal Processing Society
6. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. International Journal of Computer Vision, 63–86 (2004)
7. Lopez, A.M., Lumbreras, F., Serrat, J., Villanueva, J.: Evaluation of methods for Ridge and Valley detection. IEEE Trans. Pattern Analysis and Machine Intelligence 21(4), 327–335 (1999)
8. Herbrich, R.: Learning Kernel Classifiers: Theory and Algorithms. MIT Press, Cambridge (2001)

Myocardial Perfusion Analysis from Adenosine-Induced Stress MDCT

Samuel Silva[1], Nuno Bettencourt[2], Daniel Leite[2], João Rocha[2], Mónica Carvalho[2], Joaquim Madeira[1], and Beatriz Sousa Santos[1]

[1] DETI / IEETA — University of Aveiro, Aveiro, Portugal
[2] Cardiology Department, CHVNG/E, EPE, Vila Nova de Gaia, Portugal
sss@ua.pt

Abstract. Myocardial perfusion assessment is of paramount importance for the diagnosis of coronary artery disease. This can be performed using different image modalities such as single-photon emission computed tomography (SPECT) or magnetic resonance imaging (MRI). Recently, cardiac multiple-detector computed tomography (MDCT) has shown promising results with the benefit of gathering data regarding coronary anatomy, ventricular function and myocardial perfusion in a single study. Preliminary results for three different methods for automatic assessment of myocardial perfusion from adenosine-induced stress MDCT are presented.

1 Introduction

Multiple-detector row computed tomography (MDCT) cardiac angiography is performed by injecting the patient with a contrast agent which will allow improved visualization of the coronaries and heart chambers (ventricles and atria).

The classical MDCT acquisition is optimized for coronary imaging. However, recently it has been proposed that the same acquisitions could be used for myocardial perfusion imaging.

The MDCT acquisition is obtained during the first pass perfusion of contrast agent. During this period, perfused myocardium appears slightly brighter (since it is being perfused by blood containing contrast) [1]. If, due to coronary occlusion, a region of the myocardium is hypoperfused, it will appear darker than the remaining regions.

Under pharmacological stress, ischemic myocardial areas, which were normally perfused at rest will appear darker, allowing detection of reversible ischemia. The detection of these areas, usually related to the presence of significant epicardial coronary stenosis, is one of the main goals of cardiac imaging. Ischemia is related not only to symptoms but also to patient prognosis and its presence (more than detection of coronary plaques or stenosis) should guide treatment.

Perfusion assessment may be done in 2D using only a few slices along the left ventricle long axis. However, this is not a simple task since there is significant inter-patient variability in the way the myocardium is enhanced. Even for a particular patient, significant inter-segment enhancement variation may exist.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 717–725, 2011.

Furthermore, perfusion analysis is usually performed visually and depends on manual window/level adjustments which might vary among operators [2].

Another important issue is that perfusion analysis is usually performed using qualitative criteria, which may increase inter-observer variability and impair comparison between studies. Semi-automatic quantification might help in the development of a systematic approach, contributing to a decreased variability among clinicians.

Following on previous work concerning left-ventricle functional analysis [3] this article presents some preliminary work concerning automatic myocardial perfusion assessment from adenosine-stress MDCT images. It starts with a brief description of the image acquisition protocol and segmentation method, and then presents three different approaches to myocardial perfusion assessment: myocardium threshold, myocardial radial mean attenuation and segmental attenuation histogram analysis. The article ends with some conclusions, including a brief discussion regarding the presented methods, and ideas for future work.

2 Image Acquisition Protocol and Myocardium Segmentation

Images were obtained by multiple-detector computed tomography (MDCT) using a 64-Slice CT scanner (Somaton Sensation 64, Siemens Medical Solutions, Forchheim, Germany) with the following scan parameters: gantry rotation time of 330ms; tube voltage of 100kV; tube current of 500 – 700 mAs; electrocardiographic pulsing for decreasing radiation dose with full tube current applied at 60–65% of the cardiac cycle. Patients were under pharmacological stress by a bolus of adenosine delivered according to patient body weight (140 μg/kg/min.).

Given that image acquisition was performed using a low radiation dose protocol – and due to the moving nature of the heart – images tend to present a considerable amount of noise and movement artifacts. Overcoming these limitations is essential to allow perfusion imaging to be used as a clinical important tool. Powerful and specifically designed post-processing software is, therefore, needed.

Before performing perfusion analysis, the myocardium is segmented. This is performed with an improved version of the segmentation method proposed in Silva et al. [4] and implemented in CardioAnalyzer [5]. This segmentation method starts by proposing segmentations for the endocardium and epicardium and allows users to correct them using a 3D editing tool [3].

After segmentation, the myocardium is processed in order to identify the 17 segments used for analysis [6], by dividing the myocardium in three parts along its long axis (basal, mid-cavity and apical) and then dividing each of those parts in six, six and four angular segments respectively. A segment is included for the apical cap. Each of these segments has a direct correspondence with the different regions of the polar map (also known as bull's-eye diagram) which is often used to depict myocardium analysis data (see figure 1).

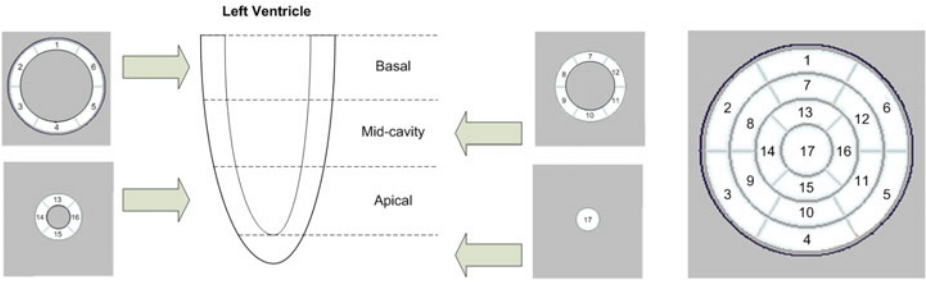


Fig. 1. The different segments considered when analysing the left ventricle and their placement in the polar map

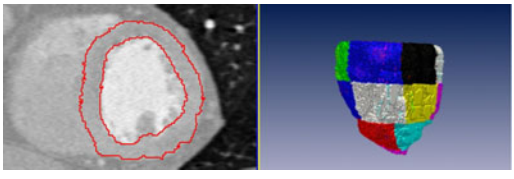


Fig. 2. Short-axis view of the left ventricle with the segmented myocardium identified by red contour and a side-view of the myocardium partially showing 11 segments

Figure 2 shows, on the left, a short axis view of the left ventricle with the segmented myocardium delimited in red and, on the right, a side view of the myocardium showing some of its segments.

3 Threshold Based Method

The first method uses global myocardium statistics. The mean ($\overline{x}$) and standard deviation (σ) for myocardium attenuation are computed. Then, each myocardium voxel which has an attenuation lower than ($\overline{x} - \sigma$) [7] is labeled as part of a possible hypoperfused region. A representation of this labeling is performed by superimposing a colored mask over the myocardium. Due to the noise present in the images, in order to improve the mask by removing isolated hypoperfused voxels or to fill small holes inside hypoperfused regions a smoothing filter is applied.

Figure 3 shows two short-axis slices and their respective colored masks where red/orange represent possible hypoperfused regions. This method of visualizing the hypoperfused regions has the advantage of being window/level independent.

The colored mask described above highlights all possible hypoperfused regions. Nevertheless, some of those regions might have resulted from an image artifact. One important aspect to consider is that – due to the terminal circulation pattern of the heart, myocardial perfusion develops from the epicardium to the endocardium – all hypoperfused regions start near the endocardium and may or may not expand to the subepicardium – but never the opposite. Thus,

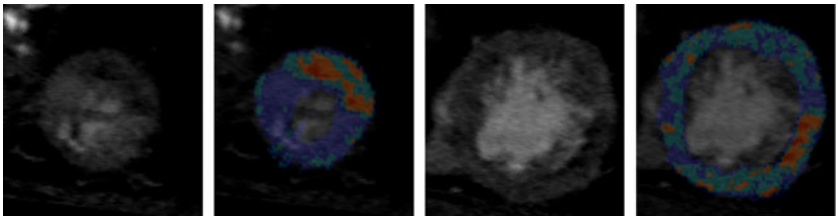


Fig. 3. Mask showing areas which might correspond to hypoperfusion since they have an attenuation bellow $\overline{x} - 1.5\sigma$

previously identified regions can be discarded if they do not involve the subendocardium. After removing these regions the mask can be processed in order to provide a polar map which conveys a summary of the obtained data.

For each slice, along its long axis, the myocardium is analysed radially and the total number of hypoperfused voxels is counted and divided by the total number of voxels found (see figure 4). The obtained data varies between zero (no hypoperfusion) to one (transmural lesion, i.e, from the endocardium to the epicardium) providing local perfusion data, as depicted in polar map (a) in figure 8.

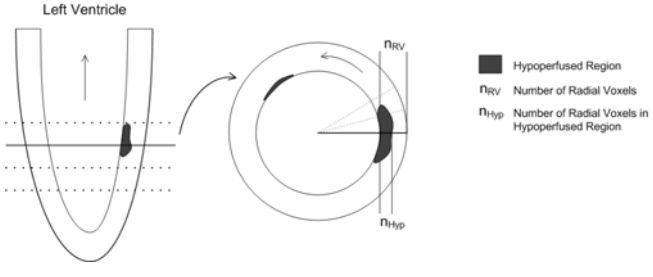


Fig. 4. The myocardium is analysed radially to compute the number of voxels corresponding to hypoperfusion

Then, the ratio between the total number of hypoperfused voxels and the total number of voxels in each myocardial segment is computed resulting in the regional (i.e., one value for each myocardial segment) polar map presented as (b) in figure 8.

4 Myocardium Radial Mean Attenuation

For all available slices along the left ventricle long axis, myocardium mean radial attenuation is computed. The highest radial mean attenuation value obtained is considered to correspond to a region presenting the best possible perfusion and the remaining values represented accordingly in a polar map.

Polar map (c) in figure 8, depicts the mean radial myocardial attenuation using a color scale typically used in workstations for myocardial perfusion assessment from SPECT. Using this color scale the hypoperfused regions appear darker than the remaining regions.

5 Segmental Attenuation Histogram

A different approach to perfusion analysis tested (as proposed by Kachenoura et al. [2]), consists in performing the analysis using the statistics of each myocardial segment instead of the statistics for the whole myocardium. For each segment the attenuation histogram is computed. Then, the histogram of the segment with the highest mean is used as a reference and subtracted from the histograms of the remaining segments. Figure 5 illustrates this procedure.

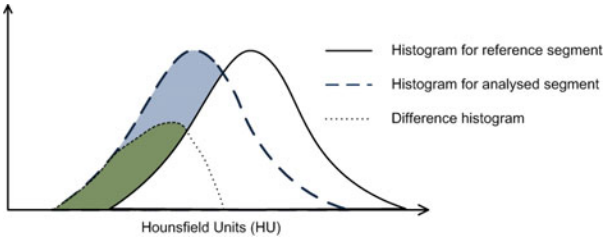


Fig. 5. The reference histogram is subtracted from the analysed segment histogram and the area below the difference histogram which does not overlap the reference histogram is used to compute the hypoperfused volume

Based on the obtained difference histogram three methods to assess perfusion have been tested. The first method considers the area bellow the difference histogram and divides it by the total number of voxels in the analysed segment. The second method considers the area below the difference histogram which does not overlap with the reference histogram also dividing it by the total number of voxels in the analysed segment. The third method uses the product of the result obtained using the second method with the distance between the maxima of the reference histogram and the difference histogram.

Figure 6 shows the polar maps obtained for a patient using the three methods. Using the third method results in a clearer representation, better highlighting the segments with a higher level of hypoperfusion.

As previously stated, there is a natural occurring difference between the mean attenuation values for each myocardial segment due to the image acquisition method characteristics. Therefore, in normal (well perfused myocardium) conditions, the histograms for different segments might not completely overlap and, eventhough this difference is usually lower than that observed between healthy and hypoperfused segments, it still affects the detection of borderline situations

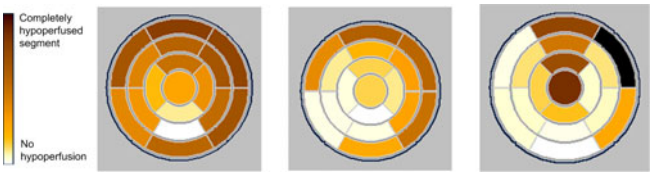


Fig. 6. Polar maps obtained for the three segmental histogram based methods tested

and hypoperfusion quantification. To cope with this issue it is important to determine the mean attenuation for each segment, for well perfused myocardia, and then account for this naturally occurring deviation when computing histogram differences.

Mean attenuation values for all myocardial segments (except the apex) have been computed from a set of 7 patients with no coronary artery disease, as assessed by coronary catheterization and a normal myocardium perfusion, as assessed by adenosine stress perfusion CMR. Figure 7 shows the mean attenuations obtained for all segments. Notice the wave-like variation among segments, consistent with analogous results obtained by Kachenoura et al. [2]. Considering the small number of patients involved, these are preliminary results and we aim to extend this study to a larger population.

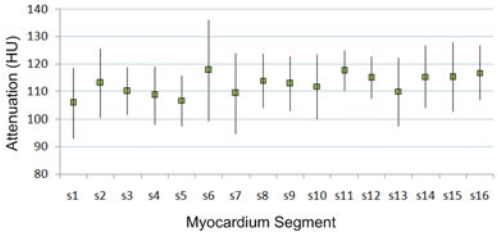


Fig. 7. Mean attenuation for all myocardial segments (except the apex) and respective standard deviation

With these segmental mean attenuations it is possible to adjust the reference histogram, before performing the subtraction, by shifting it left or right according to the naturally occurring mean attenuation difference between the two segments. The polar maps presented in figure 6 have been computed considering such shifts.

6 Conclusions and Future Work

The presented methods for perfusion assessment provide promising results allowing the detection and quantification of hypoperfused regions.

The use of myocardium threshold allows detection of possible hypoperfused regions and, with a postprocessing step, elimination of those with no contact with the endocardium. This is important since those hypoenhanced regions probably represent acquisition artifacts. It is also possible to measure the transmurality of the lesions which might be important for prognostic assessment and management decision.

Computing the mean radial attenuation also provides interesting results but using the highest mean radial attenuation as a reference might strongly affect the representation in the polar map if, for example, an image artifact results in an unusually high mean attenuation for a section of the myocardium. Therefore, a more robust method to find a reference is needed, e.g., removing outlier voxels from mean computation.

Perfusion assessment based on segmental myocardial histograms provides good results. As proposed by Kachenoura et al. [2] using a mixed index accounting for severity (how low the attenuation goes) and volume (how much voxels are below the reference) of the lesions leads to a good distinction between healthy and hypoperfused segments. Furthermore, this method also allows proper adjustments considering the naturally occurring mean attenuation differences among myocardial segments.

Figure 8 shows three representative slices for a patient and polar maps obtained using the different methods described in this article. Notice how the different polar maps present coherent results.

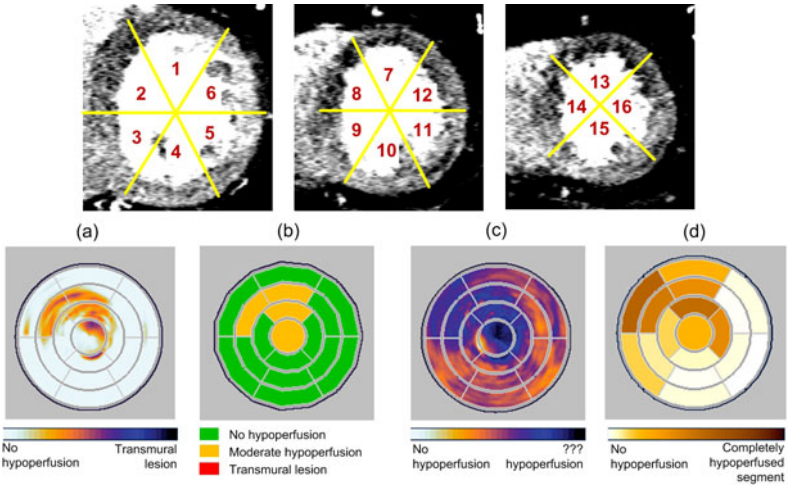


Fig. 8. Top, representative basal, mid-cavity and apical slices; bottom, from left to right, corresponding polar maps obtained using local transmurality, regional hypoperfused volume, radial myocardial mean attenuation and segmental attenuation histograms

There is still work which can be done to improve the obtained results. Given the possible existence of artifacts in the images it might be helpful to perform perfusion analysis in both systole and diastole. This would allow the detection and elimination of false positive regions by discarding those which are not common to both phases.

Eventhough the obtained perfusion analysis results make sense (i.e., the highlighted regions in the polar maps correspond to hypoperfused areas in the myocardium and the values obtained for the hypoperfused myocardium volumes have a clear meaning), it is important to compare them against perfusion assessment results provided by clinicians. This will allow a calibration of the color mappings used on the polar maps in order to, for example, attribute a precise meaning to each color in terms of hypoperfusion severity/extent. Although the presented methods allow a reasonable detection of moderate/serious hypoperfused regions, one of the main difficulties found is to detect small lesions. A comparison with perfusion assessment results provided by clinicians might help establish proper thresholds (e.g., resulting in a non-linear mapping between the computed data and the color scale used to represent it) for improved representation and detection.

As each of the presented methods provides slightly different data (e.g., transmurality, local vs regional, etc.) it would be interesting to explore how they can be blended to enhance the outputs.

Acknowledgments

The first author's work is funded by grant SFRH/BD/38073/2007 awarded by the portuguese Science and Technology Foundation (FCT).

References

1. Blankstein, R., Shturman, L., Rogers, I., Rocha-Filho, J., Okada, D., Sarwar, A., Soni, A., Loureiro, R., Feuchtner, G., Gewirtz, H., Hoffmann, U., Mamuya, W., Brady, T., Cury, R.: Adenosine-induced stress myocadial perfusion imaging using dual-source cardiac computed tomography. Journal of the American College of Cardiology 54(12), 1072–1084 (2009)
2. Kachenoura, N., Veronesi, F., Lodato, J., Corsi, C., Mehta, R., Newby, B., Lang, R., Mor-Avi, V.: Volumetric quantification of myocardial perfusion using analysis of multi-detector computed tomography 3D datasets: Comparison with nuclear perfusion imaging. European Radiology 20, 337–347 (2010)
3. Silva, S., Sousa Santos, B., Madeira, J., Silva, A.: A 3D tool for lefIt ventricle segmentation editing. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, Springer, Heidelberg (2010)
4. Silva, S., Madeira, J., Silva, A., Sousa Santos, B.: Left ventricle segmentation from heart MDCT. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) IbPRIA 2009. LNCS, vol. 5524, pp. 306–313. Springer, Heidelberg (2009)
5. Silva, S., Madeira, J., Sousa Santos, B., Silva, A.: Cardioanalyser: A software tool for segmentation and analysis of the left ventricle from 4D MDCT images of the heart. In: Proc. 7th International Conference on Biomedical Visualisation (MediVis 2010), London, UK, pp. 629–634 (2010)

6. Cerqueira, M.D., Weissmn, N.J., Dilsizian, V., Jacobs, A.K., Kaul, S., Laskey, W.K., Pennel, D.J., Rumberger, J.A., Ryan, T., Verani, M.S.: Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professional from the cardiac imaging comitee of the council on clinical cardiology of the american heart association. Circulation 105, 539–542 (2002)
7. Tamarappoo, B., Dey, D., Nakazato, R., Shmilovich, H., Smith, T., Cheng, V., Thomson, L., Hayes, S., Friedman, J., Germano, G., Slomka, P., Berman, D.: Comparison of the extent and severity of myocardial perfusion defects measured by CT coronary angiography and SPECT myocardial perfusion imaging. Journal of the American College of Cardiology: Cardiovascular Imaging 3(10), 1010–1019 (2010)

Handwritten Digits Recognition Improved by Multiresolution Classifier Fusion

Miroslav Štrba, Adam Herout, and Jiří Havel

Graph@FIT
Brno University of Technology, Brno, Czech Republic
herout@fit.vutbr.cz
http://www.fit.vutbr.cz/research/groups/graph/

Abstract. One common approach to construction of highly accurate classifiers for hadwritten digit recognition is fusion of several weaker classifiers into a compound one, which (when meeting some constraints) outperforms all the individual fused classifiers. This paper studies the possibility of fusing classifiers of different kinds (Self-Organizing Maps, Randomized Trees, and AdaBoost with MB-LBP weak hypotheses) constructed on training sets resampled to different resolutions. While it is common to select one resolution of the input samples as the "ideal one" and fuse classifiers constructed for it, this paper shows that the accuracy of classification can be improved by fusing information from several scales.

Keywords: Digit Recognition, Classifier Fusion, Multiresolution.

1 Introduction

Recognition of handwritten digits is a popular basic task of computer vision and machine learning. This can be explained both by its application potential (automatic processing of various forms etc.) and by its properties from the point of view of machine learning. In contrast to recognition of letters and especially continuous text, in the case of digits, the number of classes is reasonably low and no further syntactic/semantic analysis helps recognition of separate samples, so recognition of handwritten digits is a good (however rather simple) benchmarking task in the field of pattern recognition and computer vision.

Fusion of classifiers is a popular approach to improving performance of a machine learning system [22], [12]. Typically, different classification principles or classifiers using different feature extractors are used (e.g. [7]). Recently, Luís A. Alexandre explored the possibility to improve classification performance on the task of gender recognition [2]. A common approach to face detection, recognition, and related tasks is to identify an "ideal" image resolution, resample the images to it and use it for processing. Alexandre points out that although one of the resolutions can be identified as the best performing, the classification performance can be improved by using several different resolutions. The gain by fusion of classifiers of different sampling resolutions seems to be even higher than the

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 726–733, 2011.

gain of fusing classifiers using different feature sets (histograms of oriented gradients and local binary patterns were used). The best (by far) result is obtained by fusing different feature extractors and different resolutions at the same time. Note that the fusion principle Alexandre used was the simplest possible: decision by majority of classifiers. This paper reports a set of experiments using multiple resolutions to improve recognition of handwritten digits.

2 Handwritten Digits Recognition Background

The problem of handwritten digits recognition is popular [15], [17], [14] and therefore more databases of samples exist ([18], [24], [25], ...), which are used for training and testing the recognition systems. The most popular and most frequently used database is the MNIST database of handwritten digits [18]. This database was constructed by mixing two NIST (National Institute of Standards and Technology) databases. The database includes two datasets: the training dataset comprises 60 000 patterns and the testing dataset 10 000 patterns, 28×28 pixels in greyscale for each sample.

This paper evaluates several recognition systems based on three different methods described below. The evaluations use the MNIST database for both training and testing. In the case of the randomized trees and the SOM, the database is binarized by thresholding in the same manner as in [10].

Self-Organizing Maps. Self-Organizing Maps (SOM) [13] were chosen as a method based on neural networks, where the arrangement of neurons is defined exactly and it does not have any hidden layer. Usage of SOM to recognition of handwritten digits on the MNIST database was based on the work of Appiah et al. [3].

The essence of the training process is iterative: a new neural network is created based on the previous one, with better recognition ability. During training, each neuron adapts its weight and collects information about the count of digits for which it is winning over the other neurons. Experiments show that the best SOM size for MNIST database is 10×10. This constitutes a compromise between accuracy and time/spatial complexity. The training set was divided into smaller parts ($\sim$ 1000 samples) because the SOM are overtraining quickly for a huge set of samples. This dividing together with random weight initialization for each neuron in the net causes variability required by fusion.

Randomized Trees. The algorithm [10,4] is based on finding a relationship between areas of interest that would be unique for each class of digits. A potential area of interest is a neighbourhood of pixels (e.g. 4×4) which contains both values possible in the binarized image. Each neighbourhood is classified into a certain class called a *tag*. The relationship of two particular tags and their mutual direction (out of 8 possible directions: north, north-east, east, ...) is called a rule. The decision tree is constructed with decision rules in each non-leaf node.

The classification of neighbourhoods is carried out by a binary decision tree. This tree is called the *tag tree*. The criterion at each non-leaf node of the tag tree is the value of one pixel whose position is stored in the node.

The second phase of training is assembling the *decision tree* using the tags assigned by the tag tree. The decision tree uses rules for sorting patterns into the subtrees and leafs. Since it is impossible to investigate all rules during training for all nodes, only a subset of possible rules is selected for training each node randomly. This randomness determines the variability among the trees trained on the same training set – similarly to the SOM, this is important for later fusion.

Adaptive Boosting. Adaptive boosting [23,9] is a machine learning algorithm, which can significantly reduce the error of another learning algorithm. Its input is a set *weak classifiers* for learning and it uses a set of previously wrongly classified objects to select and train new classifiers. The result is the robust *strong classifier*. Adaptive boosting was used for recognition of handwritten digits by Carter [6].

The recognition method used in our system is similar to a method previously used for face recognition [19]. In this article, the Multi-Block Local Binary Pattern (MB-LBP) were defined and used instead of the original Local Binary Patterns (LBP) [21] that has been proved to be effective for face recognition by Ahonen et al. [1].

3 Multiresolution Classifier Fusion

Classifier fusion [22,12] – particularly in the case of recognition of handwritten digits [8] – can generate more accurate classification than each of the basic classifiers. Use of fusion is only reasonable if the learning algorithm is not deterministic and therefore it can produce different classifiers or using at least two different classification algorithms. One learning algorithm can produce different classifiers thanks to different algorithm parameters or using randomness or various inputs for training. Various inputs of the training process may be achieved by division of the training set to smaller parts, by bootstrapping the training set, or by rescaling input images to different scales. It is also possible to use identical learning algorithm but with different feature sets extracted from the input samples [2].

Majority Voting. is one of the simplest method of classifier fusion [16,11]. In decision by majority, each classifier k outputs probability decision $p_{k,i}(x)$ about the given pattern x for each class i. Value $\delta_{k,i}(x)$ in Equation (1) identifies the class with the best response (in the case of handwritten digits recognition, ten possible classes i exist):

$$\delta_{k,i}(x) = \begin{cases} 1 \text{ if } a_{k,i}(x) = \max\limits_{i=0}^{9} p_{k,i}(x) \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

Each classifier votes for one class and all classifiers share equal weight. Equation (2) describes the function for majority voting, which uses these representations of classes to evaluate the ensemble response by all N classifiers.

$$M_i(x) = \sum_{k=0}^{N} \delta_{k,i}(x) \tag{2}$$

The pattern is recognized as belonging to the class which received the most votes. If the maximal value is achieved for two or more classes, the decision is made by randomly selecting one of them.

Averaging *Log Likelihood Ratio.* is a fast method, which considers all values for different classes and not only the best selected class from each classifier. Some classification principles (randomized trees, SOM) output directly the probability $p_{k,i}(x)$ of each class. The conversion to LLR $a_{k,i}(x)$ can be done based on Equation (3):

$$p_{k,i}(x) = \frac{1}{1 - e^{a_{k,i}(x)}} \tag{3}$$

Fusion of classifiers using averaging log likelihood ratio can be expressed as

$$M_i(x) = \sum_{k=0}^{N} a_{k,i}(x); \tag{4}$$

note that since the maximal value of $M_i(x)$ is looked for, the right side of Equation (4) is not divided by the number of classifiers. The final decision is found as the class i with maximal response of $M_i(x)$.

Fusion by Linear Logistic Regression. As in the previous paragraphs, the log likelihood ratio vector

$$\boldsymbol{l}_k(x) = (a_{k,0}(x), a_{k,1}(x), \ldots, a_{k,9}(x)) \tag{5}$$

is obtained from the individual classifiers and its values are fused into one log likelihood ratio used for the final decision $\boldsymbol{l}'$ [5]:

$$\boldsymbol{l}'(x) = \sum_{k=1}^{N} \alpha_k \boldsymbol{l}_k(x) + \beta. \tag{6}$$

The fusion coefficients are found as

$$(\alpha_1, \alpha_2, \ldots, \alpha_N, \beta) = \arg\max C'_{llr}, \tag{7}$$

where C'_{llr} is calculated for the fused $\boldsymbol{l}'()$ over a supervised training database. For this purpose, the training set is split into a part used for training the individual classifiers and a part used for training the fusion by linear logistic regression [20].

4 Experimental Results

This section reports the experiments carried out to find out the benefits of fusion of different classification principles and different resolutions of the input samples. One instance of any classifier is referred to as a *weak classifier* and a fused group of weak classifiers is referred to as a *strong classifier*.

Fusion of Classifiers of Identical Parameters. requires a classification algorithm which produces different classifier instances. We tested two recognition systems: classifiers based on randomized trees and classifiers based on self-organizing maps.

Figure 1 shows how the addition of further weak classifiers has positive effect to the accuracy on the recognition system. In this case, majority voting was used for fusion. AdaBoost was not evaluated in this way because its base idea is building a strong classifier from weak ones (MB-LBP in this case) already, and for a given dataset of reasonable size (as is the case of the MNIST dataset), the classifier construction is deterministic.

Fusion of Different Classifiers. uses classifiers with different properties to evaluate the pattern precisely. Figure 2 (top) reports the fusion of the SOM and randomized tree classifiers. Please, note that fusion of systems with significantly different accuracy results in degradation of the successful system by the "noise" introduced by the worse one.

Degradation of the system by fusion is shown clearly in Figure 2 (bottom), which shows results of fusion of three systems. Fusion accuracy is worse (98.4 %) than the best performing system (AdaBoost, 98.82 %) for all fusion methods tested.

Multiscale Fusion for Different Classification Methods. is expected to be useful when the learning algorithm is working locally. In such case, the output classifier for a system with subsampled input will be less accurate than for full resolution, but using multiscale fusion, the final result will be improved. If the learning algorithm is deterministic and therefore produces only one classifier for one given training set, the multiscale fusion could increase the accuracy even if the learning algorithm operates globally over the sample area.

Figure 3 shows the gain achieved by this mode of fusion. AdaBoost and SOM learning methods are working (more-or-less) globally, so the gain is not so evident. SOM with a constant number of weak classifier ("SOM" curve) does not have any effect on the recognition accuracy because there was enough classifiers working globally, but using only one classifier ("SOM – One") for each resolution caused accuracy increase. On the randomized tree recognition system ("Rand Tree"), which has inverse properties (many classifiers work in local area) compared to the SOM, the fusion worked well. The area of interest was enlarged and the system could incorporate new information.

The following table reports the results of fusing of AdaBoost classifiers with different minimal size (all sizes from the minimal one to 28 are fused, one

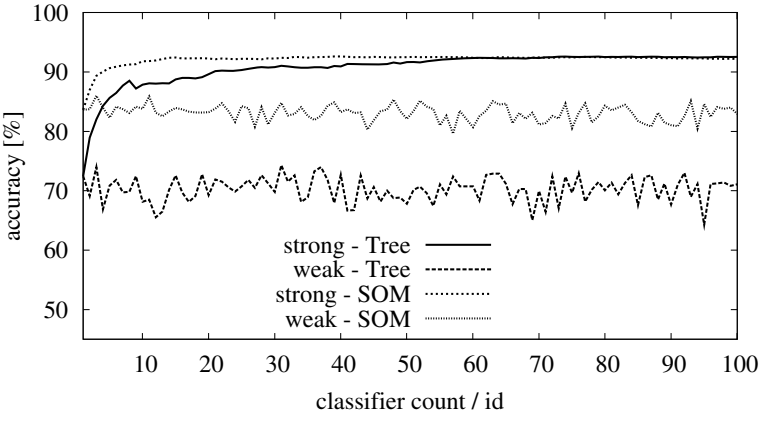


Fig. 1. Fusion of randomized tree classifiers and self-organizing maps trained on the same resolution. For "weak" classifiers, the horizontal axis is the ID of each classifier realization. For "strong", the horizontal axis is the number of weak classifiers (with ID's $1 \ldots x$) fused into the strong classifier.

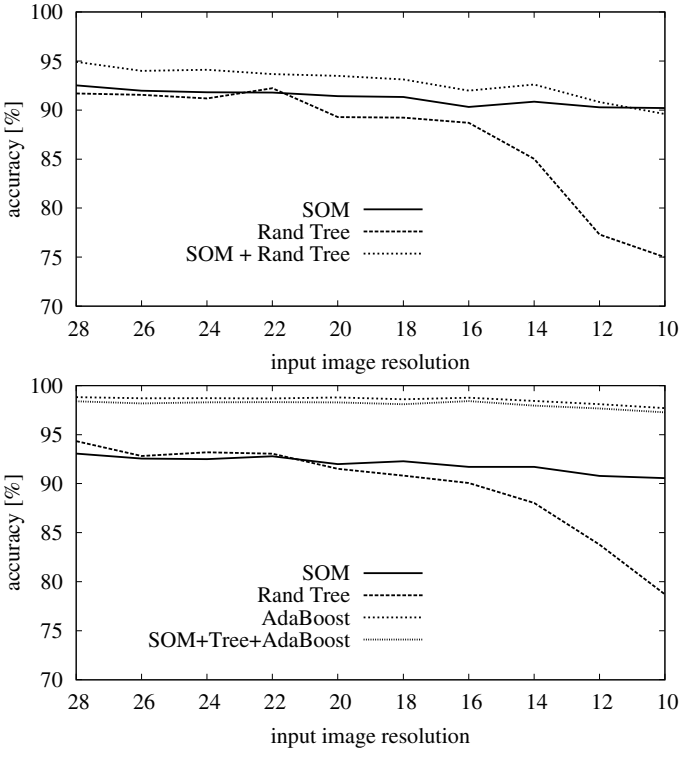


Fig. 2. (top) Fusion for different methods (SOM and Randomized Trees) with the same input resolution. (bottom) Fusion for different methods with the same input resolution.

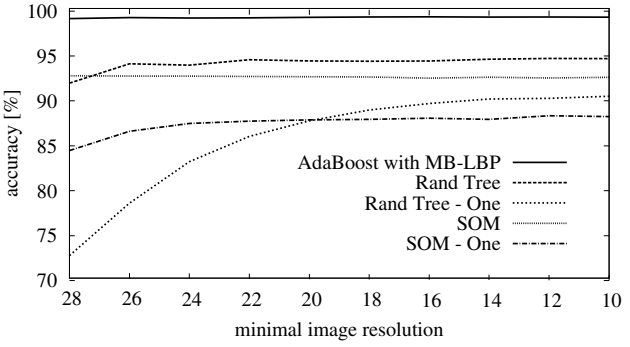


Fig. 3. Multiscale fusion for different methods. "Rand Tree - One" and "SOM - One" stand for fusion of single instances of the classifier on a given resolution, while "Rand Tree" and "SOM" are compound classifiers constructed from a number of weak ones for each resolution.

classifier per resolution). The original sample size is 28×28px, i.e. the images were downsampled. Reported are the numbers of correctly recognized samples:

patterns count	28	26	24	22	20	18	16	14	12	10
10 000	9 887	9 898	9 894	9 896	9 901	9 905	9 907	9 904	9 905	9 903

AdaBoost with input resolution 28×28 has accuracy 98.8 %. Using multiscale fusion, the accuracy was increased over 99 %. Though the curve in the graph appears flat, the accuracy improvement means elimination of over 16 % of wrongly classified samples. Multiresolution fusion therefore appears as a relatively simple method of improving accuracy of an already well-performing classifier.

5 Conclusions

The purpose of the research described in this paper is to consider and evaluate the possibility of fusing classifiers trained on different image resolutions to obtain a superior performance. The experiments carried out show that fusing different scales helps the accuracy of classification; this means that the commonly used approach – to select one best resolution and use it – is not optimal. However, the gain of this method is not as high as in the case of gender recognition reported by Alexandre [2].

We ackwnoledge the support from the BUT FIT grant FIT-10-S-2 and the research plan MSM0021630528.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
2. Alexandre, L.A.: Gender recognition: a multiscale decision fusion approach. Pattern Recognition Letters (2010), doi:10.1016/j.patrec.2010.02.010

3. Appiah, K., Hunter, A., Meng, H., Yue, S., Hobden, M., Priestley, N., Hobden, P., Pettit, C.: A binary self-organizing map and its FPGA implementation. In: IEEE International Joint Conference on Neural Networks (2009)
4. Breiman, L., Schapire, E.: Random forests. In: Machine Learning, pp. 5–32 (2001)
5. Brummer, N.: Focal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scores. Manual to SW Package (June 2007)
6. Carter, M.P.: Boosting a simple weak learner for classifying handwritten digits. Tech. Rep. PCS-TR98-341, Dartmouth College (1998)
7. Chang, S.F., He, J., Jiang, Y.G., Khoury, E.E., Ngo, C.W., Yanagawa, A., Zavesky, E.: Columbia university/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search. In: NIST TRECVID Workshop (2008)
8. Ebrahimpour, R., Hamedi, S.: Hand written digit recognition by multiple classifier fusion based on decision templates approach. In: The International Conference on Computer, Electrical, and Systems Science, and Engineering, pp. 245–250 (2009)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
10. Geman, D., Amit, Y., Wilder, K.: Joint induction of shape features and tree classifiers. IEEE Trans. PAMI 19 (1997)
11. Kant, S., Sharma, V., Dass, B.K.: On recognition of cipher bit stream from different sources using majority voting fusion rule. Sci. Analysis Group, 90–111 (2006)
12. Kittler, J.: A framework for classifier fusion: Is it still needed? In: Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, pp. 45–56. Springer, London (2000)
13. Kohonen, T.: The self-organizing map. Neurocomputing 21, 1–6 (1998)
14. Kussul, E.M., Baidyk, T.: Improved method of handwritten digit recognition tested on mnist database. Image Vision Comput. 22(12), 971–981 (2004)
15. Labusch, K., Barth, E., Martinetz, T.: Simple method for high-performance digit recognition based on sparse coding. IEEE T. Neural Networks 19, 1985–1989 (2008)
16. Lam, L., Suen, C.: Application of majority voting to pattern recognition: An analysis of its behavior and performance. SMC 27(5), 553–568 (1997)
17. Lauer, F., Suen, C.Y., Bloch, G.: A trainable feature extractor for handwritten digit recognition. Pattern Recognition 40(6), 1816–1824 (2007)
18. LeCun, Y., Cortes, C.: The mnist database of handwritten digits (2007), `http://yann.lecun.com/exdb/mnist/` (cit. December 21, 2009)
19. Liao, S., Zhu, X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007)
20. Minka, T.P.: A comparison of numerical optimizers for logistic regression. Tech. rep. (2003), `http://research.microsoft.com/~minka/papers/logreg/`
21. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
22. Ruta, D., Gabrys, B.: An overview of classifier fusion methods. Computing and Information Systems 7, 1–10 (2000)
23. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Mach. Learn. 37, 297–336 (1999)
24. Seewald, A.K.: Digits - a dataset for handwritten digit recognition. Tech. Rep. TR-2005-27, Austrian Research Institut for Artificial Intelligence (2005)
25. Vaquero, D.A., Barrera, J., Hirata Jr., R.: A maximum-likelihood approach for multiresolution W-operator design. In: SIBGRAPI, pp. 71–78 (2005)

A Comparison of Spectrum Kernel Machines for Protein Subnuclear Localization

Esteban Vegas, Ferran Reverter, Josep M. Oller, and José M. Elías

University of Barcelona, Department of Statistics,
Diagonal 645, 08028 Barcelona, Spain
{evegas,freverter,joller}@ub.edu, haliesfrik@gmail.com

Abstract. In this article, we compare the performance of a new kernel machine with respect to support vector machines (SVM) for prediction of the subnuclear localization of a protein from the primary sequence information. Both machines use the same type of kernel but differ in the criteria to build the classifier. To measure the similarity between protein sequences we employ a k-spectrum kernel to exploit the contextual information around an amino acid and the conserved motif information. We choose Nuc-PLoc benchmark datasets to evaluate both methods. In most subnuclear locations our classifier has better overall accuracy than SVM. Moreover, our method shows less computational cost than SVM.

Keywords: kernel classifier, spectrum kernel, protein subnuclear localization.

1 Introduction

The life processes of an eukaryotic cell are guided by its nucleus. The cell nucleus is a highly complex organelle that controls cell reproduction, differentiation and regulation of the metabolic activities. Cell nucleus is organized into several sub-compartments, called subnuclear locations, where proteins are located to function properly (Fig. 1). Information of their localization in a nucleus is indispensable for the in-depth study of system biology because, in addition to helping determine their functions, it can provide illuminative insights of how and in what kind of microenvironments these subnuclear proteins are interacting with each other and with other molecules. As compared to the general subcelluar localization, subnuclear localization is more challenging from biological viewpoints [1]. From computational viewpoints, the characteristic difference (e.g. amino acid composition, phylogenetic history, etc.) among the proteins in nucleus is far less distinct than that among proteins from different macro cell compartments, thus making it hard to achieve satisfactory predictive performance.

Automated techniques for high throughout protein sequencing are cheaply available. In computational proteomics, many computational models are based on protein primary sequence. Besides amino acid occurrence, pair-wise residue correlation and amino acid physicochemical properties are also incorporated to encode protein sequence.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 734–741, 2011.

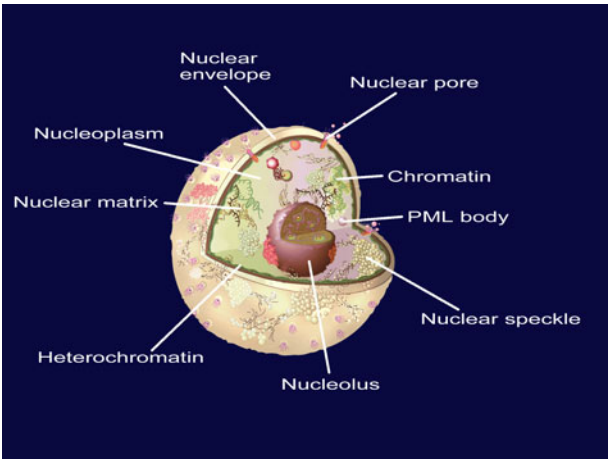


Fig. 1. Schematic drawing to show the nine sub-compartments of the cell nucleus. Image taken from Nuc-PLoc web-server.

The aim of this paper, is to evaluate the performance of two kernel machines for protein subnuclear localization. We compare our kernel classifier, based on the minimum distance probability algorithm (MDP) [2], with a support vector machine (SVM) [3]. Both machines use the same type of kernel but differ in the criteria to construct the classifier. We use k-spectrum kernel [4] to exploit the contextual information around an amino acid and the conserved motif information. Thus, we only use the amino acid information of protein sequence without any other information to train the classifiers.

2 The Spectrum Kernel

Kernel methods ([5],[6]) encompass a variety of algorithms for data analysis and machine learning, including the popular support vector machine, that share in common the use of positive definite (p.d.) kernel to represent data. Formally, a p.d. kernel K over a space of data $\mathcal{X}$ (e.g., the set of finite-length strings over an alphabet) is a function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that is symmetric (i.e., $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$) and positive definite in the sense that for any $n \in \mathbb{N}$, any $(a_1, ..., a_n) \in \mathbb{R}^n$ and any $(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{X}^n$, the following holds:

$$\sum_{i,j=1}^{n} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

Corresponding to any such kernel K there is a map $\boldsymbol{\phi}$ from $\mathcal{X}$ to a feature space $\mathcal{F}$ satisfying

$$K(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle.$$

That is, the kernel can be used to evaluate an inner product in the feature space. This is often referred to as the *kernel trick*. One instantiation of such a feature space is the *reproducing kernel Hilbert space (RKHS)* associated with K. Consider the set of functions $\{K(\cdot, \mathbf{x}) | \mathbf{x} \in \mathcal{X}\}$, where the dot represents the argument to a given function and $\mathbf{x}$ indexes the set of functions. Define a linear function space as the span of such functions. Such a function space is unique and can always be completed into a Hilbert space [7]. The crucial property of these Hilbert spaces is the *reproducing property* of the kernel:

$$f(\mathbf{x}) = \langle K(\cdot, \mathbf{x}), f \rangle \qquad \forall f \in \mathcal{F}.$$

Note in particular that if we define $\boldsymbol{\phi}(\mathbf{x}) = K(\cdot, \mathbf{x})$ as a map form the input space into the RKHS, then we have

$$\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle = \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{z}) \rangle = K(\mathbf{x}, \mathbf{z}),$$

and thus $\boldsymbol{\phi}(\mathbf{x}) = K(\cdot, \mathbf{x})$ is indeed an instantiation of the *kernel trick*.

For our application to protein localization, we use a string kernel [8], which is called the spectrum kernel, on the input space $\mathcal{X}$ of all finite length sequences of characters from an alphabet $\mathcal{A}$, $|\mathcal{A}| = \ell$. Recall, that given a number $k \geq 1$, the k-spectrum of a input sequence is the set of all the k-length contiguous subsequences that it contains. In our work, the feature map is indexed by all possible subsequences a of length k from alphabet $\mathcal{A}$. We define a feature map from $\mathcal{X}$ to $\mathbb{R}^{\ell^k}$ by

$$\Phi_k(\mathbf{x}) = \big(\phi_a(\mathbf{x})\big)_{a \in \mathcal{A}^k}$$

where $\phi_a(\mathbf{x})$ is equal to the number of times a occurs in $\mathbf{x}$. Thus the image of a sequence $\mathbf{x}$ under the feature map is weighted representation of its k-spectrum. Then, the k-spectrum kernel is

$$K_k(\mathbf{x}, \mathbf{y}) = \Phi_k(\mathbf{x})^T \Phi_k(\mathbf{y}).$$

3 Kernel Classifiers

The main objective of this paper is to evaluate the performance of two kernel classifiers for protein subnuclear localization. In particular, we compare our kernel classifier, that extent a previous classifier, called Minimum Distance Probability (MDP) [2], with SVM. Abundant references about SVM can be easy found, for instance [3], then we will pay attention on the MDP method.

3.1 Minimum Distance Probability (MDP)

Given a population Ω and a set of subpopulations $(E_1, ..., E_k)$ which constitutes a partition of Ω, in many practical problems it is necessary to allocate an individual, $w \in \Omega$, to one of the subpopulations $E_1, ..., E_k$, according to the observed values of a random vector $\mathbf{x}$.

Nonparametric classification [9] use the nonparametric approach to estimate the class-conditional densities $P(\mathbf{x}|E_i)$. However, care should be applied to using nonparametric estimates in high-dimensional spaces because the *curse of dimensionality*.

MDP avoids this problem by constructing a partition, $\mathcal{R} = \{R_1, ..., R_k\}$ of the input domain, $\mathcal{X}$, with the same number of elements as classes, by means a distance function. An individual w will be assigned to class R_i if the probability that w will be *nearer* to E_i than to any other class E_j is maximum. Once the partition $\mathcal{R}$ has been established, we can apply Bayes' classifier to assign classes of $\mathcal{R}$ to subpopulations.

We will concentrate on the construction of a partition $\mathcal{R} = \{R_1, ..., R_k\}$ of $\mathcal{X}$. Given an element $r \in \mathcal{X}$ and a random sample, $\mathbf{w}_\lambda$, of size $k\lambda$ (where $\lambda = 1, 2, ...$ is arbitrarily chosen) obtained taking λ elements from each classes, that is:

$$\mathbf{w}_\lambda = (w_{11}, ..., w_{1\lambda}, ..., w_{k1}, ..., w_{k\lambda});$$

$$(w_{11}, ..., w_{1\lambda}) \in E_1^\lambda, ..., (w_{k1}, ..., w_{k\lambda}) \in E_k^\lambda \tag{1}$$

For each $r \in \mathcal{X}$ we can define the set, $S_i^\lambda(r)$, composed of all possible samples obtained as in (1) for which the minimum of the distances between r and the $k\lambda$ images obtained from $\mathbf{w}_\lambda$ corresponds to one of the λ elements from E_i included in $\mathbf{w}_\lambda$. That is

$$S_i^\lambda(r) = \Big\{ \mathbf{w}_\lambda \in E_1^\lambda \times \cdots \times E_k^\lambda \,|\, \exists\, \beta \in \{1, ..., \lambda\} \text{ with } \min\{d(r, \mathbf{x}(w_{11})), ..., d(r, \mathbf{x}(w_{k\lambda}))\} = d(r, \mathbf{x}(w_{i\beta})) \Big\}, \tag{2}$$

$i = 1, ..., k$.

Then, the probabilities $P(S_1^\lambda(r)), ..., P(S_k^\lambda(r))$ of the sets $S_1^\lambda(r), ..., S_k^\lambda(r)$ respectively, allows us to determine the partition $\mathcal{R}_\lambda$. Notice that $P(S_i^\lambda(r))$ is the probability that the result $r \in \mathcal{X}$ will be *nearer* E_i than any other class E_j. *Nearer* must be interpreted in the sense that, if we take a random sample $\mathbf{w}_\lambda$ defined as (1), the minimum of all the distances between r and the elements from $\mathbf{w}_\lambda$ would by attained by an element of E_i.

The partition $\mathcal{R}_\lambda = \{R_1^\lambda, ..., R_k^\lambda\}$ is defined in the following way:

$$R_i^\lambda = \{r \in \mathcal{X} \,|\, P(S_i^\lambda(r)) > P(S_j^\lambda(r)), j = 1, ..., k; j \neq i\}, \qquad i = 1, ..., k \tag{3}$$

Then R_i^λ can be interpreted as the set of results for which the probability of being nearer to E_i than to any other class E_j is maximum.

When the class likelihood $P(\mathbf{x}|E_i)$ is unknown the way to define the partition $\mathcal{R}_\lambda$ is not obvious because the probabilities $P(S_i^\lambda(r))$, $i = 1, ..., k$, cannot calculated directly. However, a reasonable estimation can be obtained as follows:

In practical cases we usually have a controlled sample G with N known classified individuals taken from k possible classes $(E_1, ..., E_k)$:

$$\{e_{11}, ..., e_{1n_1}\} \in E_1, ..., \{e_{k1}, ..., e_{kn_k}\} \in E_k, \tag{4}$$

with $N = n_1 + \cdots + n_k$.

To estimate $P(S_i^\lambda(r))$, $i = 1, ..., k$, we must take all the possible samples of G composed by $k\lambda$ individuals, λ from every subpopulation;

$$\mathbf{w}_\lambda^\alpha = \{(e_{11}^\alpha, ..., e_{1\lambda}^\alpha) \in E_1^\lambda, ..., (e_{k1}^\alpha, ..., e_{k\lambda}^\alpha) \in E_k^\lambda\}, \qquad \alpha = 1, ..., B \tag{5}$$

where $B = \prod_{i=1}^k \binom{n_i}{\lambda}$. In practice, the method that we use to find the optimal value of λ is cross-validation, taking into account that $\lambda = \min\{n_1, ..., n_k\}$.

For each individual $w \in G$ with $\mathbf{x}(w) = r$ and for each sample $\mathbf{w}_\lambda^\alpha$, $\alpha = 1, ..., B$, we need to compute the values $d(\mathbf{x}(w), \mathbf{x}(e_{11}^\alpha)), ..., d(\mathbf{x}(w), \mathbf{x}(e_{k\lambda}^\alpha))$ in order to obtain:

$$\hat{p}(s_i^\lambda(w)) = \frac{1}{B}\#\Big\{\mathbf{w}_\lambda^\alpha \in E_1^\lambda \times \cdots \times E_k^\lambda | \exists\, \beta \in \{1, ..., \lambda\} \text{ with } \min\{d(r, \mathbf{x}(e_{11}^\alpha)), ..., d(r, \mathbf{x}(e_{k\lambda}^\alpha))\} = d(r, \mathbf{x}(e_{i\beta}^\alpha))\Big\}, \tag{6}$$

$\alpha = 1, ..., B$, where # symbolizes the cardinal of the set, so that $\hat{p}(s_i^\lambda(w))$ is the proportion of the number of times that the minimum of the distances from w to the elements of $\mathbf{w}_\lambda$ is attained by an element from E_i. From (6) it is clear that $\hat{p}(s_i^\lambda(w))$ is an estimator of $P(S_i^\lambda(r))$.

The sets $\hat{R}_i^\lambda$ of the partition $\hat{\mathcal{R}}_\lambda$ will include all the individuals of G for which:

$$\hat{R}_i^\lambda = \{w \in G \mid \hat{p}(s_i^\lambda(w)) > \hat{p}(s_j^\lambda(w)), j = 1, ..., k; j \neq i\} \qquad i = 1, ..., k.$$

Finally, the individuals included in $\hat{R}_i^\lambda$, will be assigned to the class E_j if and only if:

$$P(E_j|\hat{R}_i^\lambda) = \max\{P(E_1|\hat{R}_i^\lambda), ..., P(E_k|\hat{R}_i^\lambda)\}, \tag{7}$$

where

$$P(E_j|\hat{R}_i^\lambda) = \frac{P(\hat{R}_i^\lambda|E_j) \cdot P(E_j)}{\sum_{h=1}^k P(\hat{R}_i^\lambda|E_h) \cdot P(E_h)}$$

Here $P(\hat{R}_i^\lambda|E_h)$ can be estimated by

$$\hat{p}(\hat{R}_i^\lambda|E_h) = \frac{\#\{w \in G | w \in (\hat{R}_i^\lambda \cap E_h)\}}{n_h}, \qquad i, h = 1, ..., k,$$

and, if we no further information, $P(E_h)$ can be estimated by: $\hat{p}(E_h) = n_h/N$, or if we suppose that all subpopulations are equally probable, by $\hat{p}(E_h) = 1/k$.

This procedure is repeated for all the individuals of the training set in order to obtain the estimations $\hat{p}(\hat{R}_i^\lambda|E_h)$.

To allocate a new individual $\tilde{\omega}$. Firstly, we estimate the probabilities $\hat{p}(s_1^\lambda(\tilde{\omega}))$, ..., $\hat{p}(s_k^\lambda(\tilde{\omega}))$. Second, assign $\tilde{\omega}$ to the set $\hat{R}_i^\lambda$ if only if

$$\hat{p}(s_i^\lambda(\tilde{\omega})) = \max\{\hat{p}(s_1^\lambda(\tilde{\omega})), ..., \hat{p}(s_k^\lambda(\tilde{\omega}))\}$$

Finally, once we have assigned the new individual to the set $\hat{R}_i^\lambda$, then we use the probabilities (7), obtained during the training step, to allocate $\tilde{\omega}$ to the most likely subpopulation.

Efficient Computation. There is an easier procedure to compute $\hat{p}(s_i^{\lambda}(w))$, $i = 1, ..., k$, in (6). In practice, we need to compute the distances from the individual w to all the individuals of G. The resulting vector, $\mathbf{D} = (d_1, ..., d_N)$, is ordered from least to greatest, $\mathbf{D}^* = (d_{(1)}, ..., d_{(N)})$. Really, we do not need to keep the values of $\mathbf{D}^*$, so if $d_{(h)} = d(\mathbf{x}(w), \mathbf{x}(e_{ij}))$, we only need to know that the individual associated with $d_{(h)}$, e_{ij}, comes from E_i. Then:

$$\hat{p}(s_i^{\lambda}(w)) = \frac{1}{B} \sum_{\beta=1}^{n_i} \Big(\prod_{\gamma=1}^{k} \binom{a_{\gamma}^{(\beta)}}{\lambda - \delta_{i\gamma}} \text{ while } a_{\gamma}^{(\beta)} \geq \lambda - \delta_{i\gamma} \Big),$$

where $\delta_{i\gamma}$ are Kronecker deltas, $a_{\gamma}^{(\beta)}$ symbolizes the number of individuals from E_{γ} ($\gamma = 1, ..., k$) placed on the right of the βth individual from E_i ($\beta = 1, ..., n_i$) in the vector $\mathbf{D}^*$.

Kernel MDP. We present an extension of MDP algorithm to the kernel methods. For brevity we refer to our method as the KMDP. An essential tool for MDP algorithm is the problem of computing distances between two points $\mathbf{x}, \mathbf{z} \in \mathcal{X}$. Implicitly embedding $\mathcal{X}$ into a Hilbert space, the distance between two points $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ in the embedding being given by:

$$d(\mathbf{x}, \mathbf{z}) = \sqrt{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{z}, \mathbf{z}) - 2K(\mathbf{x}, \mathbf{z})} \tag{8}$$

Then, by applying (8) in the computation of the distances (6) we can extend the MDP algorithm to a kernel-based methodology for pattern analysis.

4 Results

We choose Nuc-PLoc [10] benchmark datasets to evaluate the performance of the k-spectrum kernel machines that we compare in this study. The Nuc-PLoc dataset is collected from the Swiss-Prot database (version 52.0 released on 6 May 2007) [11] and divides cell nucleus into 9 subnuclear locations and the number of proteins in the locations is unbalanced, the largest Nucleolus has 307 proteins and the smallest Nuclear PML body has only 13 proteins. The dataset has total 714 proteins.

We have used $k = 2$ in our k-spectrum kernel and we have chosen $\lambda = 2$ in KMDP method, after searching between other values. To evaluate the performance of the classifiers we have implemented a 5-fold cross-validation procedure.

Table 1 summarizes the specificity (SP), the sensibility (SE) and the Matthew's correlation coefficient (MCC) achieved by SVM and KMDP methods for each subnuclear location. The measure MCC reveals that KMDP achieves better performance in all subnuclear locations except nucleolus. Both classifiers have high SP but KMDP shows higher SE than SVM.

To give more details about KMDP method we show Tables 2 and 3. During the training step, KMDP estimate the conditional probabilities $P(E_j/R_i)$ for $i, j = 1, ..., 9$. As is expected, highest probability appear in the diagonal of the

Table 1. Performance comparison on Nuc-Plot dataset

		SP		SE		MCC	
Subnuclear location	Size	KMDP	SVM	KMDP	SVM	KMDP	SVM
Chromatin	99	0.82	1	0.59	0.07	0.34	0.25
Heterochromatin	22	0.87	1	1	0	0.41	0
Nuclear envelope	61	0.99	1	0.28	0.05	0.48	0.23
Nuclear matrix	29	0.91	1	0.90	0.13	0.48	0.35
Nuclear pore complex	79	0.93	1	0.66	0.3	0.54	0.53
Nuclear speckle	67	0.98	0.99	0.28	0.2	0.38	0.37
Nucleolus	307	0.99	0.14	0.06	1	0.17	0.25
Nucleoplasm	37	0.89	1	0.78	0	0.42	0
Nuclear PLM body	13	0.91	1	1	0	0.42	0

Table 2. Estimated conditional probabilities $P(E_j|R_i)$

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
Chromatin (E_1)	0.25	0.04	0.00	0.00	0.05	0.00	0.08	0.00	0.04
Heterochromatin (E_2)	0.05	0.72	0.00	0.00	0.07	0.00	0.00	0.09	0.03
N. envelope (E_3)	0.09	0.00	1.00	0.27	0.25	0.00	0.13	0.00	0.07
N. matrix (E_4)	0.15	0.05	0.00	0.58	0.03	0.00	0.00	0.20	0.11
N. pore complex (E_5)	0.04	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.06
N. speckle (E_6)	0.11	0.04	0.00	0.12	0.02	0.56	0.00	0.09	0.17
Nucleolus (E_7)	0.18	0.10	0.00	0.03	0.05	0.04	0.57	0.10	0.05
Nucleoplasm (E_8)	0.13	0.04	0.00	0.00	0.08	0.10	0.21	0.52	0.02
N. PLM body (E_9)	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.46

Table 3. Estimated probabilities of R_i

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9
seq1	0.47	0.13	0.02	0.00	0.12	0.00	0.09	0.16	0.00
seq2	0.44	0.05	0.00	0.00	0.13	0.01	0.06	0.30	0.00
seq3	0.16	0.05	0.00	0.00	0.24	0.00	0.05	0.46	0.03
seq4	0.45	0.03	0.01	0.00	0.08	0.02	0.05	0.35	0.02
seq5	0.23	0.12	0.00	0.00	0.10	0.02	0.07	0.43	0.02

Table 2. Thus, elements belonging to set R_i will be assign to the location E_i with higher probability.

As an example, Table 3 shows the estimated probabilities of each elements of the partition, R_i, $i = 1, ..., 9$, of five proteins located in the Chromatin. These probabilities have been obtained using formula (6). For example, first protein (seq1) will be assign to the set R_1 because this set has highest probability (0.47).

Then, using the conditional probabilities in the R_1 column of the Table 2, this protein will be correctly assigned to the Chromatin (E_1) because the $P(E_1|R_1) = 0.25$ is highest value in this column. On the other hand, third protein (seq3) will be assign to the set R_8 because this set this is more likely (0.46). Noting the column R_8 of the Table 2, this protein will be assign wrong to the Nucleoplasm (E_8).

5 Conclusions

In this paper, we evaluate the performance of two kernel machines for protein subnuclear localization. We compare our kernel classifier (KMDP) with a support vector machine (SVM), both machines use the same type of kernel. For our application to protein localization, we use a string kernel, which is called the spectrum kernel, that compute the similarity between the k-spectrum of two protein sequences. We choose Nuc-PLoc benchmark datasets to evaluate the performance of both machines. Matthew's correlation coefficient (MCC) reveals that KMDP achieves better performance in all subnuclear locations except nucleolus. Both classifiers have high specificity (SP) but KMDP shows higher sensibility (SE) than SVM. The main KMDP properties are: versatility due to the possibility of selecting in each problem a suitable kernel function, lower computational cost than SVM and provides a measure of the confidence in the allocation of each observation.

Acknowledgments. This research was funded by grant MEC-MTM2008-00642.

References

1. Lei, Z., Dai, Y.: An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 6, 29 (2005)
2. Villarroya, A., Ríos, M., Oller, J.M.: Discriminant Analysis Algorithm Based on A Distance Function and on Bayesian Decision. Biometrics 51, 908–919 (1995)
3. Cristianini, N., Shawe-Taylor, J.: An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
4. Scholkopf, B., Tsuda, K., Vert, J.P.: Kernel Methods in Computational Biology. MIT Press, Cambridge (2004)
5. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
6. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
7. Saitoh, S.: Theory of Reproducing Kernels and its Applications. Longman Scientific & Technical, Harlow (1988)
8. Leslie, C., Eskin, E., Noble, W.: The spectrum kernel: a string kernel for SVM protein classification. In: Altman, R.B., Dunker, A.K., Hunter, L., Lauerdale, K., Klein, T.E. (eds.) Proceedings of the Pacific Symposium on Biocomputing 2002, pp. 564–575. World Scientific, Singapore (2002)
9. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, Cambridge (2004)
10. Shen, H., Chou, K.: Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng. Des. Sel. 20, 561–567 (2007)
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., Donovan, C., Phan, I., et al.: The SWISS-PROT protein knowledgebase and its Supplement TrEMBL. Nucleic Acids Research 31, 365–370 (2003)

Complex Wavelet Transform Variants in a Scale Invariant Classification of Celiac Disease

Andreas Uhl[1], Andreas Vécsei[2], and Georg Wimmer[1]

[1] Department of Computer Sciences, University of Salzburg, Salzburg, Austria
[2] St. Anna Children's Hospital, Vienna, Austria

Abstract. In this paper, we present variants of the Dual-Tree Complex Wavelet Transform (DT-CWT) in order to automatically classify endoscopic images with respect to the Marsh classification. The feature vectors either consist of the means and standard deviations of the subbands from a DT-CWT variant or of the Weibull parameter of these subbands. To reduce the effects of different distances and perspectives toward the mucosa, we enhanced the scale invariance by applying the discrete Fourier transform or the discrete cosine transform across the scale dimension of the feature vector.

Keywords: Endoscopic Imagery, Celiac Disease, Automated Classification, Scale Invariance, Dual-Tree Complex Wavelet Transform, Discrete Fourier Transform, Discrete Cosine Transform.

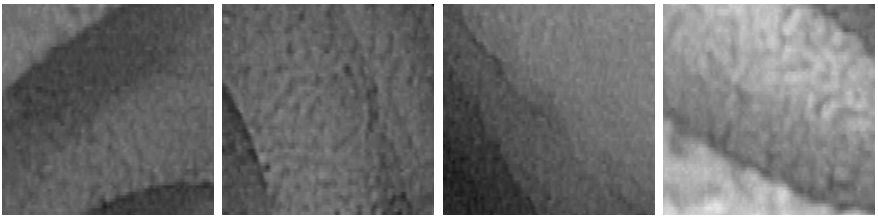
1 Introduction

The celiac state of the duodenum is usually determined by visual inspection during the endoscopy session followed by a biopsy of suspicious areas. The severity of the muscosal state of the extracted tissue is defined according to a modified Marsh scheme, which divides the images in four different classes, Marsh-0 Marsh-3a, Marsh-3b and Marsh-3c (see Figure 1). Marsh-0 represents a healthy duodenum with normal crypts and villi, Marsh-3a, Marsh-3b and Marsh 3c have increased crypts and mild atrophy (3a), marked atrophy (3b) or the villi are entirely absent (3c), respectively. Types Marsh-3a to Marsh-3c span the range of characteristic changes caused by celiac disease, whereat Marsh-3A is the mildest and Marsh-3c the most severe form. We distinguish between two regions of the duodenum, the bulbus duodeni and the pars descendes.

In gastroscopic (and other types of endoscopic) imagery, mucosa texture is usually found with different perspective, zoom (see Figure 1) and distortions (barrel-type distortions of the endoscope [1]). That means that the mucosal textures shows different spatial scales, depending on the camera perspective and distance to the mucosal wall.

Consequently, in order to design reliable computer-aided mucosa texture classification schemes, the scale invariance of the employed feature sets could be essential.

J. Vitrià, J.M. Sanches, and M. Hernández (Eds.): IbPRIA 2011, LNCS 6669, pp. 742–749, 2011.



(a) Marsh-3a (b) Marsh-3b (c) Marsh-3c (d) Marsh-0

Fig. 1. Example images for the respective classes taken from the pars descendes database

We consider feature vectors extracted from subbands of the Dual-Tree Complex Wavelet Transform (DT-CWT), the Double Dyadic Dual-Tree Complex Wavelet Transform (D^3T-CWT) and the Quatro Dyadic Dual-Tree Complex Wavelet Transform (D^4T-CWT) [2], since wavelet transforms in general excel by their respective multiscale properties. A classical way of computing scale invariant features from multi-scale methods like e.g. the DT-CWT [4,5] is to apply the discrete Fourier transform (DFT) to statistical parameters of the subband coefficients' distributions (e.g. mean and standard deviation) and compute the magnitudes of these complex values. In this work, we also use the real part of the DFT or apply the real-valued discrete cosine transform (DCT) to coefficient parameters, which enhanced the results and the scale invariance in magnification-endoscopy image classification [2]. In addition to classical coefficient distribution parameters, we employ shape and scale parameters of the Weibull distribution [3] to model the absolute values of each subband.

This paper is organized as follows. In section 2 we discuss the basics of the DT-CWT, the D^3T-CWT and the D^4T-CWT. Subsequently we describe the feature extraction with focus on achieving scale invariance with the DFT or DCT. In section 3 we describe the experiments and present the results. Section 4 presents the discussion of our work.

2 Cwt Variants and Scale Invariant Features

Kingbury's Dual-Tree Complex Wavelet Transform [6] divides an image into six directional (15°, 45°, 75°, 105°, 135°, 165°) oriented subbands per level of decomposition. The DT-CWT analyzes an image only at dyadic scales. The D^3T-CWT [4] overcomes this issue, by introducing additional levels between dyadic scales. These additional levels between dyadic scales are generated by applying the DT-CWT to a downscaled version of the original image using a factor of $2^{-0.5}$. We use the bicubic interpolation to scale down the image. Instead of the levels 1, 2,..., L in the DT-CWT we get the levels 1, 1.5, 2,..., L+0.5 in the D^3T-CWT, where the integer levels correspond to the levels of the DT-CWT. The D^4T-CWT works similar to the D^3T-CWT, with the difference that the D^4T-CWT has even more additional levels between the scales. These scales are

generated by applying the DT-CWT to the downscaled versions of the original image using the factors $\sqrt{3/8}$, $\sqrt{1/2}$, and $\sqrt{3/4}$. The advantages of these three complex wavelet transforms are their approximately shift-invariance, their directional selectivity and the very efficient implementation scheme. In this paper, we use two ways to generate the feature set from the DT-CWTs. The first and most common approach is to compute the empirical mean ($\mu_{l,d}$) and the empirical standard deviation ($\sigma_{l,d}$) of the absolute values of each subband (decomposition level $l \in \{1,\ldots,L\}$ and direction $d \in \{1,\ldots,6\}$) and concatenate them to one feature vector later denoted as Classic distribution).

The second approach is to model the absolute values of each subband by a two-parameter Weibull distribution [3]. The probability density function of a Weibull distribution with shape parameter c and scale parameter b is given by

$$p(x;c,b) = \frac{c}{b}\left(\frac{x}{b}\right)^{c-1} e^{-(\frac{x}{b})^c}. \tag{1}$$

The moment estimates (c,b) of the Weibull parameters of each subband are then arranged into feature vectors like in the approach before. The feature extraction for the D^{3,4}T-CWTs works the same way, but the feature vector is longer because of the non-dyadic scales.

A common approach to achieve scale-invariance for wavelet-based features is to use the absolute values of a Discrete Fourier Transformation (DFT) applied to extracted statistical moments. We use the method from [4,5] and apply the DFT to the feature vector (of the DT-CWT) as follows

$$U_{n,d} = \frac{1}{\sqrt{L}}\sum_{l=1}^{L} \mu_{l,d}\, e^{\frac{-i\,2\pi(l-1)(n-1)}{L}}, \qquad S_{n,d} = \frac{1}{\sqrt{L}}\sum_{l=1}^{L} \sigma_{l,d}\, e^{\frac{-i\,2\pi(l-1)(n-1)}{L}} \tag{2}$$

for $n \in \{1,\ldots,L\}$ and $d \in \{1,\ldots,6\}$. The feature curve of a feature vector shifts if input texture is scaled. DFT magnitude makes the feature values independent of cyclic shifts of the feature curve. The DFT assumes that there is a periodic input signal; however there is no reason why the statistical features should be periodic. If the statistical features are close to zero at both ends, the approach provides good scale invariance.

For the D^{3}T-CWT, we replace L with $2L$ and $n \in \{1, 1.5, 2, \ldots, L+0.5\}$ and for the D^{4}T-CWT we replace L by $4L$ and $n \in \{1, 1.25, 1.5, 1.75, 2, \ldots, L+0.75\}$. The new feature vector (for the DT-CWT) is

$$f = \{|U_{1,1}|,\ldots|U_{L,1}|,|U_{1,2}|,\ldots|U_{L,2}|,\ldots|U_{L,6}|,|S_{1,1}|,\ldots|S_{L,1}|,\ldots|S_{L,6}|\}.$$

The feature vectors for the D^{3}T-CWT and D^{4}T-CWT are created by analogy.

It turned out that the results of the real values of the $U's$ and $S's$ provide better results than the absolute values [2]. Because of

$$\Re\left(e^{\frac{-i\,2\pi(l-1)(n-1)}{L}}\right) = \cos\left(\frac{-2\pi(l-1)(n-1)}{L}\right), \tag{3}$$

the real values of the DFT are obtained by a cosinus transform. Hence we propose to use the Discrete Cosinus Transform (DCT). The DCT of one of our feature vector is computed by

$$U(n,d) = w(n) \sum_{l=1}^{L} \mu_{l,d} \cos\left(\frac{\pi(2(l-1)(n-1)}{L}\right) \tag{4}$$

for $n \in \{1,\ldots,L\}$ and $d \in \{1,\ldots,6\}$ (and similar for $S(n,d)$) , where $w(n) = 1/\sqrt{L}$ for $n = 1$ and $w(n) = 2/\sqrt{L}$ for $2 \leq n \leq L$. For the Weibull parameter case, $\mu_{l,d}$ and $\sigma_{l,d}$ are simply replaced by $c_{l,d}$ and $b_{l,d}$.

Applying the DCT or DFT for the D^{3}T-CWT works similar, but it turns out, that the transformation leads to better results if we apply the DCT or DFT on $(\mu_{1,d}, \mu_{2,d}, \ldots, \mu_{L,d})$ and $(\mu_{1.5,d}, \mu_{2.5,d}, \ldots, \mu_{L+0.5,d})$ separately, instead of $DCT(\mu_{1,d}, \mu_{1.5,d} \ldots, , \mu_{L+0.5,d})$. The DCT or DFT for the D^{4}T-CWT is done in a similar fashion by applying them four times separately.

Further we have to note, that in case of the DFT, parts of the feature vector will be deleted after the DFT, because the complex conjugates are redundant in the feature vector. If we use RGB-images, than we simply concatenate the feature vectors of each color channel.

3 Experimental Study

We employ two methods to evaluate and compare the feature sets described in the section before: The area under the ROC curve (AUC) [7] and the overall classification accuracy.

To generate the ROC curve for k-NN classifier we used the method described in [7] (for k=20). We consider the 20 nearest neighbors of each image (the 20 feature vectors with the lowest euclidean distance to the feature vector of the considered image). We employ leave–one–out cross–validation (LOOCV) to find these 20 nearest neighbors for each image. We achieve the first point on the ROC curve by classifying the images as positive, if one or more than one nearest neighbor of a considered image is positive. Because for nearly every image, there is at least one of the 20 neighbors positive (positive means that the image belongs to class Marsh-3a, Marsh-3b or Marsh-3c), the true positive rate (TPR) (= sensitivity) and the false positive rate (FPR) (= 1 - specificity) will be 100 % or near to 100 %. The second point on the ROC curve is achieved by classifying an image as positive, if two or more of the 20 nearest neighbors are positive, the third point if three or more of the 20 nearest neighbors are positive and so on till 20. The more positive nearest neighbors of the 20 nearest neighbors are needed to classify an image as positive, the lower are the TPR and FPR. The last point is achieved by classifying an image as positive, if all the 20 nearest neighbors are positive. Because there is hardly ever one of the 20 nearest neighbors of an image negative, the TPR and the FPR are in this case 0 % or near to 0 %. That is the way to generate the points on the ROC curve. To be sure that the curve reaches from the the point (0,0) (TPR=0, FPR=0) to the point (1,1) (TPR=1,FPR=1), we add these points to the curve. The point (0,0) can be interpreted as 21 of the the 20 nearest neighbors of an image have to be positive to classify the image as positive. This is not possible and so the TPR and FPR are 0 %. The point (1,1)

can be interpreted as 0 or more of the 20 nearest neighbors of an image have to be positive to classify the image as positive. This will always happens and so the TPR and FPR will be 100 %. In Figure 2 we see two examples of ROC-curves.

The AUC is computed by trapezoidal integration,

$$AUC = \sum_{i=1}^{21} ((TPR_i \cdot \Delta FPR_i) + 1/2(\Delta TPR_i \cdot \Delta FPR_i)) \tag{5}$$

where

$$\Delta TPR_i = TPR_{i-1} - TPR_i, \tag{6}$$

$$\Delta FPR_i = FPR_{i-1} - FPR_i, \tag{7}$$

and TPR_i or FPR_i are those TPR or FPR, where at least i positive nearest neighbors (out of the 20 nearest neighbors) are necessary to classify an image as positive.

The second method is the 20-Nearest Neighbor (denoted by 20-NN) classifier. We already know the 20 nearest neighbors from the AUC. An image is classified as positive, if more than the half (=10) of the nearest neighbors are positive, or as negative if more than the half of the nearest neighbors are negative. If there are 10 positive and 10 negative nearest neighbors, then the image is classified as its nearest neighbor (1-NN classifier). Classification accuracy is defined as the number of correctly classified samples divided by the total number of samples.

Before decomposing the images with the CWTs, we employ two preprocessing steps to improve the performance [8]. First we employ adaptive histogram equalization using the CLAHE (contrast-limited adaptive histogram equalization) algorithm with 8×8 tiles and a uniform distribution for constructing the contrast transfer function. Second, we blur the image by a Gaussian 3×3 mask with $\sigma = 0.5$.

The image database consists of a total of 273 bulbus duodeni and 296 pars descendes images and was taken at the St. Anna's Children Hospital using a standard duodenoscope without magnification. In order to condense information of the original endoscopic images, we cut out regions of interest of size 128×128 [8]. Table 1 lists the number of image samples per class. Tests were carried out with 6 levels of decomposition and RGB-images. We only consider the 2-class case.

The results for the AUC are given in Table 2. If we watch the results for the bulbus dataset, we can see that the results with DCT or without any further

Table 1. Number of image samples per Marsh type (ground truth based on histology)

Data set	Bulbus				Pars			
Marsh type	0	a	b	c	0	a	b	c
Number of images (4-class case)	153	45	54	21	132	42	53	69
Number of images (2-class case)	153	120			132	164		

Table 2. Area under the ROC curve for the two data sets (bulbus and pars) with features extracted from DT-CWT variants by computing the Classic or the Weibull distribution and none or a further manipulation of the feature vectors by DFT variants or the DCT

Feature		Classic				Weibull			
Manipulation		non	DFT abs	DFT real	DCT	non	DFT abs	DFT real	DCT
Bulbus	DT-CWT	0.97	0.82	0.96	0.98	0.97	0.78	0.95	0.97
	D^3T-CWT	0.98	0.81	0.96	0.98	0.98	0.78	0.97	0.98
	D^4T-CWT	0.98	0.83	0.96	0.98	0.98	0.80	0.96	**0.99**
Pars	DT-CWT	0.82	0.76	**0.88**	0.86	0.81	0.78	0.86	0.84
	D^3T-CWT	0.82	0.76	0.87	0.87	0.82	0.78	0.85	0.84
	D^4T-CWT	0.84	0.77	0.86	**0.88**	0.83	0.77	0.85	0.86

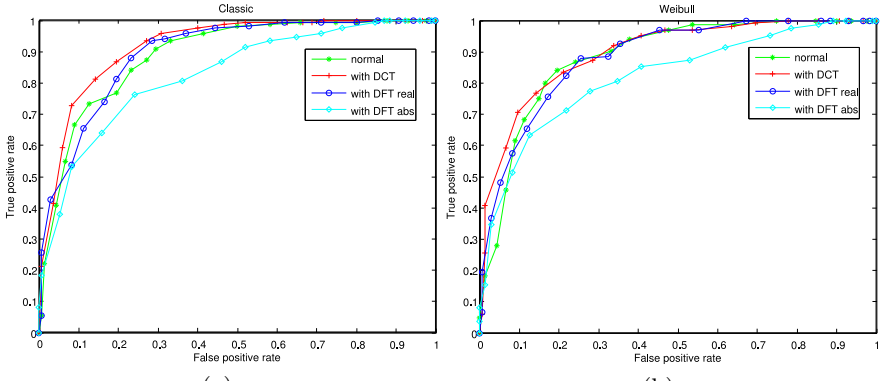


Fig. 2. ROC curves of the different feature vector manipulation methods (normal, DCT, DFT abs, DFT real) for the pars descendens dataset , with features extracted from the D^4T-CWT by the classic way (a) or the Weibull distribution (b)

manipulation of the feature vector are similar. The results with the absolute values of the DFT are much worse and the results with the real values of the DFT are a little bit worse than the results with DCT or without any further manipulation of the feature vector. In case of the pars descendens dataset, the results with DCT or the real part of the DFT are distinctly better than those without feature vector manipulation and the results with the absolute part of the DCT are much more worse compared to the other methods. The differences between the CWT-variants or the feature extraction methods (classic way or Weibull distribution) are very small. The best results for the both datasets in Table 2 are given in bold face numbers.

The results for the 20-NN classifier are given in Table 2. We can see that the results for the bulbus dataset are similar with DCT or without any further manipulation of the feature vector. The results with the real valued DFT are a little bit worse than the results mentioned before. In case of the pars descendens dataset, the results with DCT are worse than the results without any further

Table 3. Classification accuracy in % for the 20-NN classifier with the two data sets (bulbus and pars) and features extracted from DT-CWT variants by computing the Classic or the Weibull distribution and none or a further manipulation of the feature vectors by DFT variants or the DCT

Feature		Classic				Weibull			
Manipulation		non	DFT abs	DFT real	DCT	non	DFT abs	DFT real	DCT
Bulbus	DT-CWT	94.9	74.7	91.6	94.5	93.8	72.9	91.2	92.3
	D^3T-CWT	94.9	74.0	92.3	94.9	94.1	70.3	93.4	**95.2**
	D^4T-CWT	94.5	79.1	92.3	94.9	94.5	74.7	91.9	**95.2**
Pars	DT-CWT	82.4	70.3	**84.1**	77.4	82.1	73.3	70.3	76.0
	D^3T-CWT	82.1	68.2	82.1	78.0	81.4	70.3	80.7	74.3
	D^4T-CWT	82.4	69.3	82.1	78.4	81.8	71.6	80.1	79.4

feature vector manipulation and the results with the real part of the DFT are in case of the classic features similar and in case of the Weibull features worse than the results without feature vector manipulation. The results with the absolute part of the DFT are always worse than the other results. Once again, the results of the different CWT-variants are similar and the results of our feature exraction methods (classic and weibull) are also similar, apart from the case with the real valued DFT and the pars descendens dataset. The best results for each of the both datasets in Table 3 are given in bold face numbers.

4 Discussion

It is hard to interpret these results because the AUC and the overall classification results are often contradictory. For an example let us consider the results of the results of the DCT for the pars descendens dataset. The AUC is distinctly larger with DCT than without feature vector manipulation, whereas the classification accuracy for the 20-NN classifier is distinctly higher without feature vector manipulation than with DCT. If we watch the results, then it is impossible to say if we should favor a feature vector manipulation like the DCT or the real valued DFT or prefer no further feature vector manipulation. The advantages of a better balancing of different perspectives, zooms and distortions seems to be equal than the drawbacks like losing scale information by making the feature vector more scale invariant or by destroying information by the transformation. Maybe the results of the AUC are more significant than the overall classification results, because an overall classification result uses only the information whether there are more positive or negative nearest neighbors for an image, whereas the AUC uses the information how much nearest neighbors are positive or negative.

There are small improvements of the CWT's with additional scales in between dyadic scales (D^3T-CWT, D^4T-CWT) compared to the standard DT-CWT, but because of their higher computational complexity it is questionable if the small improvements justify their application.

The classic way and the Weibull distribution are equally suited to extract the information from the subbands of the CWT's.

References

1. Gschwandtner, M., Liedlgruber, M., Uhl, A., Vécsei, A.: Experimental Study on the Impact of Endoscope Distortion Correction on Computer-assisted Celiac Disease Diagnosis. In: Proceedings of the 10th International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu (2010)
2. Häfner, M., Uhl, A., Vécsei, A., Wimmer, G., Wrba, F.: Complex Wavelet Transform Variants and Scale Invariance in Magnification-Endoscopy Image Classification. In: Proceedings of the 10th International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu (2010)
3. Kwitt, R., Uhl, A.: Modeling the Marginal Distributions of Complex Wavelet Coefficient Magnitudes for the Classification of Zoom-Endoscopy Images. In: Proceedings of the IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2007), Rio de Janeiro, Brasil, pp. 1–8 (2007)
4. Lo, E.H.S., Pickering, M.R., Frater, M.R., Arnold, J.F.: Image segmentation from scale and rotation invariant texture features from the double dyadic dual-tree complex wavelet transform. Image and Vision Computing 29(1), 15–28 (2011)
5. Manthalkar, R., Biswas, P.K., Chatterji, B.N.: Rotation and scale invariant texture features using discrete wavelet packet transform. Pattern Recognition Letters 24(14), 2455–2462 (2003)
6. Kingsbury, N.G.: The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. In: Proceedings of the IEEE Digital Signal Processing Workshop, DSP 1998, Bryce Canyon, Utah, USA, pp. 9–12 (1998)
7. Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 20, 1145–1159 (1997)
8. Hegenbart, S., Kwitt, R., Liedlgruber, M., Uhl, A., Vécsei, A.: Impact of Duodenal Image Capturing Techniques and Duodenal Regions on the Performance of Automated Diagnosis of Celiac Disease. In: Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA 2009), Salzburg, Austria, pp. 718–723 (2009)

Author Index