LNCS 8259

José Ruiz-Shulcloper
Gabriella Sanniti di Baja (Eds.)

# Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications

18th Iberoamerican Congress, CIARP 2013
Havana, Cuba, November 2013
Proceedings, Part II

2 Part II

IAPR

Springer

# Lecture Notes in Computer Science 8259

José Ruiz-Shulcloper
Gabriella Sanniti di Baja (Eds.)

# Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications

## 18th Iberoamerican Congress, CIARP 2013
## Havana, Cuba, November 20-23, 2013
## Proceedings, Part II

Springer

Volume Editors

José Ruiz-Shulcloper
Advanced Technologies Application Center (CENATAV)
$7^a$ A#21406 esq. 214 y 216, Rpto. Siboney, Playa. C.P. 12200 La Habana, Cuba
E-mail: jshulcloper@cenatav.co.cu

Gabriella Sanniti di Baja
Institute of Cybernetics "E. Caianiello", National Research Council (CNR)
Via Campi Flegrei 34, 80078 Pozzuoli (Naples), Italy,
E-mail: g.sannitidibaja@cib.na.cnr.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-41826-6 e-ISBN 978-3-642-41827-3
DOI 10.1007/978-3-642-41827-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013951329

CR Subject Classification (1998): I.5, I.4, I.2.10, I.2.7, F.2.2, J.3

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

The 18th Iberoamerican Congress on Pattern Recognition CIARP 2013 (Congreso IberoAmericano de Reconocimiento de Patrones) is the yearly event of a series of pioneer conferences on pattern recognition in the scientific community active in this field in Iberoamerican countries.

As has been the case for previous editions of the conference, CIARP 2013 hosted worldwide participants with the aim to promote and disseminate ongoing research on mathematical methods and computing techniques for pattern recognition, in particular in biometrics, computer vision, image analysis, and speech recognition, as well as their application in a number of diverse areas such as industry, health, robotics, data mining, entertainment, space exploration, telecommunications, document analysis, and natural language processing and recognition. Moreover, CIARP 2013 was a useful forum in which the scientific community could exchange research experience, share new knowledge and increase cooperation among research groups in pattern recognition and related areas.

We like to underline that CIARP conferences have significantly contributed to the birth and growth of national associations for pattern recognition in Iberoamerican countries that are already members of the International Association for Pattern Recognition, IAPR, (Argentina, Brazil, Chile, Cuba, Mexico), or will soon be applying to become IAPR members (Colombia, Peru, Uruguay).

CIARP 2013 received 262 contributions from 37 countries (12 of which are Iberoamerican countries). After a rigorous blind reviewing process, where each submission was reviewed by three highly qualified reviewers, 137 papers by 355 authors from 31 countries were accepted. All the accepted papers have scientific quality above the overall mean rating.

As has been the case for the most recent editions of the conference, CIARP 2013 was a single-track conference in which 22 papers where selected for presentation in oral sessions, while the remaining 115 papers were selected for poster presentation with short poster teasers. Following the tradition of CIARP conferences, the selection of the presentation type does not signify at all a quality grading. CIARP 2013 presentations were grouped into nine sessions: Supervised and Unsupervised Classification; Feature or Instance Selection for Classification; Image Analysis and Retrieval; Signals Analysis and Processing; Biometrics; Applications of Pattern Recognition; Mathematical Theory of Pattern Recognition; Video Analysis; and Data Mining.

We would like to point out that the reputation of CIARP conferences is increasing, especially since the last 11 editions for which the proceedings have been published in the *Lecture Notes in Computer Science* series. Moreover, starting from CIARP 2008, authors of the best papers presented at the conference (orally or as posters) have been invited to submit extended versions of their papers to well-known journals so as to enhance the visibility of their conference submissions and to stimulate deeper insight into the treated topics. For CIARP 2013 two special issues of the *International Journal of Pattern Recognition and Artificial Intelligence IJPRAI* and in *Intelligent Data Analysis IDA* will be published. Moreover, a Special Section of Pattern Recognition Letters has been added to include the two papers of the researchers selected as the winners of the two prizes given at CIARP 2013, namely the IAPR-CIARP Best Paper Prize and the Aurora Pons-Porrata Medal, which is a new CIARP-Award.

The IAPR-CIARP Best Paper Prize has the aim of acknowledging and encouraging excellence, originality and innovativeness of new models, methods and techniques with an outstanding theoretical contribution and practical application to the field of pattern recognition and/or data mining. The Iberoamerican CIARP-Award Aurora Pons-Porrata Medal is given to a living woman in recognition of her outstanding technical contribution to the field of pattern recognition or data mining.

The selection of the winners is based on the wish of the authors to be considered as possible candidates for the prizes, the evaluation and recommendations of members of the Program Committee, for the IAPR-CIARP Best Paper Prize, and the proposal of the national associations on Pattern Recognition, for the Aurora Pons-Porrata Medal, and the evaluation of the respective Award Committees. The task of these committees, whose members are carefully chosen to avoid conflicts of interest, is to evaluate each paper nominated for the IAPR-CIARP Best Paper Prize by performing a second review process including the quality of the (poster or oral) presentation, and the recommendations for the Aurora Pons-Porrata Medal. We express our gratitude to the members of the two Award Committees: Josef Kittler (Surrey University, UK), Jian Pei (Simon Fraser University, Canada), Fabio Roli (University of Cagliari, Italy), Tieniu Tan (National Laboratory on Pattern Recognition of China), Isneri Talavera-Bustamante (Advanced Technologies Applications Center, CENATAV, Cuba), Rita Cucchiara (University of Modena-Reggio, Italy), and Rocio González-Díaz, (University of Seville, Spain).

Besides the 137 accepted submissions, the scientific program of CIARP 2013 also included the contributions of three outstanding invited speakers, namely, Jian Pei (Simon Fraser University of Canada), Fabio Roli (University of Cagliari, Italy) and Tieniu Tan (National Laboratory on Pattern Recognition of China). The papers of these two last keynotes appear in these proceedings. Furthermore, the three invited speakers and Gabriella Sanniti di Baja gave four tutorials on "Mining Uncertain and Probabilistic Data for Big Data Analytics", "Multiple Classifier Systems", "Fundamentals of Iris Recognition", and "Discrete Methods to Analyse and Represent 3D Digital Objects," respectively.

During the conference, the Annual CIARP Steering Committee Meeting was also held.

CIARP 2013 was organized by the Advanced Technologies Applications Center (CENATAV) and the Cuban Association for Pattern Recognition (ACRP) with the endorsement of the International Association for Pattern Recognition (IAPR), and the sponsorship of the Cuban Society for Mathematics and Computer Sciences (SCMC), the Argentine Society for Pattern Recognition (SARP-SADIO), the Special Interest Group of the Brazilian Computer Society (SIGPR-SBC), the Chilean Association for Pattern Recognition (AChiRP), the Mexican Association for Computer Vision, Neural Computing and Robotics (MACVNR), the Spanish Association for Pattern Recognition and Image Analysis (AERFAI), and the Portuguese Association for Pattern Recognition (APRP). We recognize and appreciate their valuable contributions to the success of CIARP 2013.

We gratefully acknowledge the help of all members of the Organizing Committee and of the Program Committee for their support and for the rigorous work in the reviewing process.

We also wish to thank the members of the Local Committee for their unflagging work in the organization of CIARP 2013 that led to an excellent conference and proceedings.

Special thanks are due to all authors who submitted to CIARP 2013, including those of papers that could not be accepted.

Finally, we invite the pattern recognition community to attend CIARP 2014 in Puerto Vallarta, Mexico.

November 2013

José Ruiz-Shulcloper
Gabriella Sanniti di Baja

# Organization

CIARP 2013 was organized by the Cuban Association for Pattern Recognition, endorsed by the International Association for Pattern Recognition (IAPR) and sponsored by the Advanced Technologies Applications Center (CENATAV), DATYS Technologies & Systems, Cuba.

## Co-chairs

| | |
|---|---|
| José Ruiz-Shulcloper | Advanced Technologies Applications Center, (CENATAV), Cuba |
| Gabriella Sanniti di Baja | National Research Council (CNR) of Italy, Naples, Italy |

## IAPR-CIARP 2013 Best Paper Prize Committee

| | |
|---|---|
| Josef Kittler | Surrey University, UK |
| Jian Pei | Simon Fraser University, Canada |
| Fabio Roli | University of Cagliari, Italy |
| Gabriella Sanniti di Baja | CNR, Napoli, Italy |
| Tieniu Tan | National Laboratory on Pattern Recognition, China |

## CIARP 2013 Aurora Pons-Porrata Award Committee

| | |
|---|---|
| Rita Cucchiara | University of Modena-Reggio, Italy |
| Rocío González-Díaz | University of Seville, Spain |
| Isneri Talavera-Bustamante | CENATAV, Cuba |

## Local Committee

Niusvel Acosta-Mendoza
José R. Calvo-De Lara
Marieli Capote-Rodríguez (DATYS)
Andrés Gago-Alonso
Edel García-Reyes
Eduardo Garea-Llano
Ricardo González-Gazapo
José Hernández-Palancar
Rainer Larín-Fonseca
Danis López-Naranjo
José Medina-Pagola
Heydi Méndez-Vázquez
Diana Porro-Muñoz
Maité Romero-Durán
Isneri Talavera-Bustamante

## CIARP Steering Committee

Eduardo Bayro-Corrochano, Mexico
Cesar Beltrán Castañón, Perú
Edel García-Reyes, Cuba
Marta Mejail, Argentina
Alvaro Pardo, Uruguay
Roberto Paredes Palacios, Spain
Olga Regina Pereira Bellon, Brazil
João Miguel Sanches, Portugal
Cesar San Martín, Chile

## Program Committee

| | |
|---|---|
| Sergey Ablameyko | Belarusian State University |
| José Aguilar | Universidad de Los Andes, Venezuela |
| René Alquézar | Universitat Politécnica de Catalunya, Spain |
| Akira Asano | Kansai University, Japan |
| Ali Ismail Awad | Faculty of Engineering, Al Azhar University, Egypt |
| Ildar Batyrshin | Kazan State Technological University, Russia |
| Eduardo Bayro-Corrochano | CINVESTAV, Unidad Guadalajara, IPN, México |
| Rafael Bello | Univ. Central “Marta Abreu” de Las Villas, Cuba |
| César Beltrán Castañón | Pontificia Universidad Católica del Perú |
| José Miguel Benedí | Universidad Politécnica de Valencia, Spain |
| Jón Atli Benediktsson | University of Iceland |
| Rafael Berlanga-Llavori | Universitat Jaime I Castelló, Spain |
| Gunilla Borgefors | Uppsala University, Sweden |
| Dibio Borges | University of Brasilia, Brazil |
| João Rogério Caldas Pinto | Universidad Técnica de Lisboa, Portugal |
| José Ramón Calvo de Lara | Advanced Technologies Applications Center, Cuba |
| Virginio Cantoni | Università di Pavia, Italy |
| Jesús Ariel Carrasco-Ochoa | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Mario Castelán | CINVESTAV, Unidad Saltillo, IPN, México |
| Eduardo Concepción | Universidad de Cienfuegos, Cuba |
| Mauricio Correa | Universidad de Chile |
| Marco Cristani | University of Verona, Italy |
| Isabelle Debled-Rennesson | LORIA, France |
| Alberto Del Bimbo | Universitá degli Studi di Firenze, Italy |
| Maria De Marsico | Sapienza University of Rome, Italy |
| Claudio De Stefano | Università di Cassino e del Lazio Meridionale, Italy |
| Robert P.W. Duin | Delft University of Technology, The Netherlands |

| | |
|---|---|
| Aytul Ercil | Sabanci University, Turkey |
| Boris Escalante | Universidad Nacional Autónoma de México |
| Alfonso Estudillo-Romero | Universidad Nacional Autónoma de México |
| Jacques Facon | Pontificia Universidade Católica do Paraná, Brazil |
| Carlos Ferrer | Univ. Central “Marta Abreu” de Las Villas, Cuba |
| Francesc J. Ferri | Universidad de Valencia, Spain |
| Ana Fred | Instituto Superior Técnico, Portugal |
| Maria Frucci | Istituto di Cibernetica E. Caianiello, CNR, Italy |
| Andrés Gago-Alonso | Advanced Technologies Applications Center, Cuba |
| Edel García-Reyes | Advanced Technologies Applications Center, Cuba |
| Eduardo Garea-Llano | Advanced Technologies Applications Center, Cuba |
| Alexander Gelbukh | CIC, Instituto Politécnico Nacional, México |
| Lev Goldfarb | University of New Brunswick, Fredericton, Canada |
| Pilar Gómez-Gil | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Jordi González | Universitat Autònoma de Barcelona, Spain |
| Manuel Graña-Romay | Universidad del País Vasco, Spain |
| Igor Gurevich | Dorodnicyn Computing Center, Russian Academy of Sciences |
| Michal Haindl | Inst. Information Theory and Automation, Czech Republic |
| Edwin Hancock | University of York, UK |
| Raudel Hernández | Advanced Technologies Applications Center, Cuba |
| José Hernández-Palancar | Advanced Technologies Applications Center, Cuba |
| Laurent Heutte | Université de Rouen, France |
| Jinwei Jiang | The Ohio State University, USA |
| Xiaoyi Jiang | Universität Münster, Germany |
| Martin Kampel | Vienna University of Technology, Austria |
| Sang-Woon Kim | Myongji University, Korea |
| Reinhard Klette | The University of Auckland, New Zealand |
| Vitaly Kober | CICESE, México |
| Walter Kosters | Universiteit Leiden, The Netherlands |
| Walter Kropatsch | Vienna University of Technology, Austria |
| Rainer Larín-Fonseca | Advanced Technologies Applications Center, Cuba |
| Denis Laurendeau | Université Laval, Canada |

| | |
|---|---|
| Manuel S. Lazo-Cortés | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Federico Lecumberry | Universidad de la República, Uruguay |
| Aurelio López-López | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Juan Valentín Lorenzo-Ginori | Univ. Central “Marta Abreu” de Las Villas, Cuba |
| Cris L. Luengo Hendriks | Swedish University of Agricultural Sciences, Sweden |
| José Francisco Martínez Trinidad | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Nelson Mascarenhas | Federal University of Sao Carlos, Brazil |
| José E. Medina Pagola | Advanced Technologies Applications Center, Cuba |
| Marta Mejail | University of Buenos Aires, Argentina |
| Heydi Méndez-Vázquez | Advanced Technologies Applications Center, Cuba |
| Manuel Montes-y-Gómez | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Marco Mora | Catholic University of Maule, Chile |
| Claudio Moraga | Technische Universität Dortmund, Germany |
| Eduardo Morales | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Alberto Muñoz | Carlos III University, Spain |
| Vittorio Murino | University of Verona, Italy |
| Pablo Muse | Universidad de la República, Uruguay |
| Michele Nappi | Università di Salerno, Italy |
| Heinrich Niemann | University of Erlangen-Nuremberg, Germany |
| Mark Nixon | University of Southampton, UK |
| Lawrence O’Gorman | Bell Laboratories, USA |
| Emanuele Olivetti | Neuroinformatics Lab, FBK, Trento, Italy |
| Kalman Palagyi | University of Szeged, Hungary |
| Alvaro Pardo | Inst. de Matemáticas y Estadística, Uruguay |
| Talita Perciano | University of São Paulo, Brazil |
| Airel Pérez-Suárez | Advanced Technologies Applications Center, Cuba |
| Pedro Pina | Instituto Superior Técnico, Portugal |
| Armando Pinho | University of Aveiro, Portugal |
| Sílvia Pinto | IME – USP, Brazil |
| Hemerson Pistori | Catholic University, Campo Grande, Brazil |
| Filiberto Pla | Universitat Jaime I Castelló, Spain |
| Carlos Pon | Universidad Catolica del Norte, Chile |
| Petia Radeva | Universitat Autónoma de Barcelona, Spain |
| Pedro Real | Universidad de Sevilla, Spain |

| | |
|---|---|
| Carlos A. Reyes-García | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Bernardete Ribeiro | University of Coimbra, Portugal |
| Daniel Riccio | Università di Napoli Federico II, Italy |
| Gerhard Ritter | University of Florida, USA |
| Roberto Rodríguez | Inst. de Cibernética, Mat. y Física, Cuba |
| Fabio Roli | University of Cagliari, Italy |
| Edgar Román-Rangel | University of Geneva, Switzerland |
| Alejandro Rosales-Pérez | Inst. Nac. Astronomía, Óptica Electrónica, México |
| Arun Ross | Michigan State University, USA |
| Luis Rueda | University of Windsor, Canada |
| Javier Ruiz-del-Solar | Universidad de Chile |
| Hichem Sahli | Vrije Universiteit Brussel, Belgium |
| João Sanches | Instituto Superior Técnico, Portugal |
| Dairazalia Sánchez-Cortes | Idiap Research Institute, Switzerland |
| Alberto Sanfeliu | Universitat Politecnica de Catalunya, Spain |
| César San Martín | Universidad de Concepción, Chile |
| Carlo Sansone | Universita di Napoli Federico II, Italy |
| Roberto Santana | University of the Basque Country, Spain |
| Angel Sappa | Universitat Autónoma de Barcelona, Spain |
| Basilio Sierra | University of the Basque Country, Spain |
| Ida-Maria Sintorn | Uppsala University, Sweden |
| Juan Humberto Sossa Azuela | CIC, Instituto Politécnico Nacional, México |
| Beatriz Sousa Santos | University of Aveiro, Portugal |
| Concetto Spampinato | University of Catania, Italy |
| Tania Stathaki | Imperial College London, UK |
| Robin Strand | Uppsala University, Sweden |
| Carmen Paz Suárez-Araujo | Universidad de las Palmas de Gran Canaria, Spain |
| Zhenan Sun | National Laboratory on Pattern Recognition, China |
| Alberto Taboada-Crispi | Univ. Central “Marta Abreu” de Las Villas, Cuba |
| Isneri Talavera | Advanced Technologies Applications Center, Cuba |
| Tieniu Tan | National Laboratory on Pattern Recognition, China |
| Mariano Tepper | Duke University, USA |
| Massimo Tistarelli | University of Sassari, Italy |
| Karl Tombre | Université de Lorraine, France |
| María Inés Torres | Universidad del País Vasco, Spain |
| Yulia Trusova | Dorodnicyn Computing Center, Russian Academy of Sciences |

| | |
|---|---|
| Ventzeslav Valev | Inst. Math. and Informatics, Bulgarian Academy of Sciences |
| Sandro Vega-Pons | Neuroinformatics Lab, FBK, Trento, Italy |
| Cornelio Yáñez-Márquez | CIC, Instituto Politécnico Nacional, México |
| Vera Yashina | Dorodnicyn Computing Center, Russian Academy of Sciences |
| Zhi-Hua Zhou | Nanjing University, China |

## Additional Reviewers

Michael Affenzeller
Danilo Benozzo
Marco Bertini
Battista Bigio
Maria Elena Buemi
Pablo Cancela
Qing Da
Luca Didaci
Fazel Famili
Francesco Fontanella
Luca Ghiani
Luis Gómez
Norberto Goussies
Gabriel Hernández Sierra
Michelle Horta
Svebor Karaman
Gisela Klette
Bruno Leitão
Alexandre Levada
Haiqing Li
Dongwei Liu
Noel Lopes
Itzamá López-Yáñez
Ana Luísa Martins
Pedro Martins
John Mason
Sérgio Matos
Igor Montagner
Antonio Neves
Bao Nguyen
Matias Nitsche
Tomás Oliveira e Silva
Darian Onchis
Caroline Petitjean
Ales Prochazka
Luca Pulina
John Rugis
Denis Salvadeo
Mario Sansone
Riccardo Satta
Alessandra Scotto di Freca
Lorenzo Seidenari
Augusto Silva
Yunlian Sun
César Teixeira
Ana Maria Tomé
Tiberio Uricchio
Susana Vieira
Lihu Xiao

## Sponsoring Institutions

Advanced Technologies Applications Center (CENATAV)
International Association for Pattern Recognition (IAPR)
Cuban Association for Pattern Recognition (ACRP)
Cuban Society for Mathematics and Computer Sciences (SCMC)
Argentine Society for Pattern Recognition (SARP-SADIO)
Chilean Association for Pattern Recognition (AChiRP)
Mexican Association for Computer Vision, Neural Computing and Robotics (MACVNR)
Special Interest Group of the Brazilian Computer Society (SIGPR-SBC)
Spanish Association for Pattern Recognition and Image Analysis (AERFAI)
Portuguese Association for Pattern Recognition (APRP)

# Table of Contents – Part II

## Keynote

## Applications of Pattern Recognition

## Biometrics

## Video Analysis

## Data Mining

# Table of Contents – Part I

## Keynote

## Mathematical Theory of PR

## Supervised and Unsupervised Classification

## Feature or Instance Selection for Classification

## Image Analysis and Retrieval

## Signals Analysis and Processing

# Recent Progress on Object Classification and Detection

Tieniu Tan, Yongzhen Huang, and Junge Zhang

Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences, (CASIA), Beijing, China
{tnt,yzhuang,jgzhang}@nlpr.ia.ac.cn

**Abstract.** Object classification and detection are two fundamental problems in computer vision and pattern recognition. In this paper, we discuss these two research topics, including their backgrounds, challenges, recent progress and our solutions which achieve excellent performance in PASCAL VOC competitions on object classification and detection. Moreover, potential directions are outlined for future research.

**Keywords:** Object classification, Object detection, PASCAL VOC.

## 1 Introduction

Object classification and detection are two core problems in computer vision and pattern recognition. They play fundamental and crucial roles in many applications, e.g., intelligent visual surveillance, image and video retrieval and web content analysis. Object classification and detection share some common components and face many common challenges (see Fig. 1), e.g., variability in illumination, rotation and scales, as well as deformation, clutter, occlusion, multi-stability and large intra-class variations.

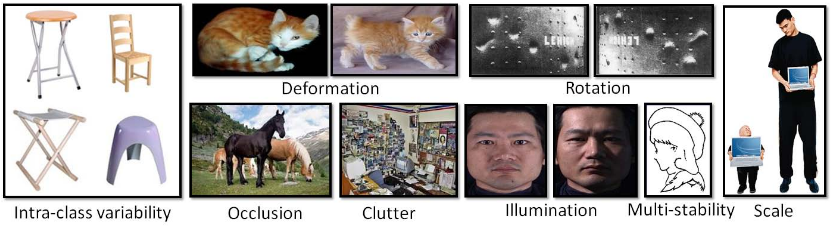


**Fig. 1.** Common challenges in object classification and detection

Despite the above challenges, great progress has been made in the past decades, and many algorithms have been proposed. In this paper we discuss the general framework of object classification and detection, and some classic methods in each

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 1–8, 2013.

component of the framework. Afterwards, we introduce our work on object classification and detection, especially our solutions which were ranked among the best in the PASCAL VOC competition [1,2]. Finally, we point out some potential directions for future work.

## 2 General Framework and Methods

### 2.1 General Framework

The general framework of object classification and detection is illustrated in Fig. 2, where the first and the last module are shared by object classification and detection. In this subsection, we discuss these two common modules, and then analyze the other modules in the following two subsections.

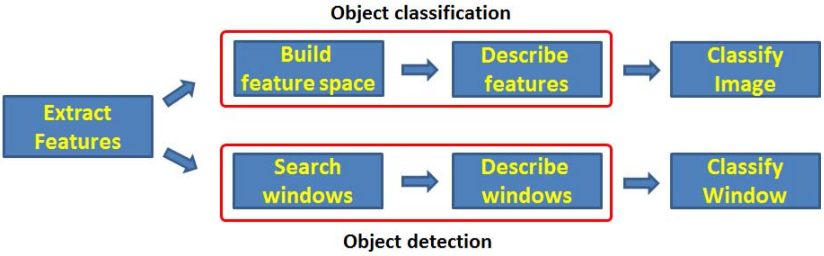


**Fig. 2.** A general framework of object classification and detection

The first module, i.e., feature extraction, usually includes two main steps: extracting image patches and representing image patches. Extracting image patches is implemented via sampling local areas of images, usually in a dense or a sparse manner. Representing image patches is implemented via statistical analysis over pixels of image patches. The representation vectors of image patches are called local features. Widely used features include: 1) appearance based ones, e.g., scale-invariant feature transform (SIFT) [19], histogram of oriented gradients (HOG) [8]; 2) color based ones, e.g., color descriptors [25]; and 3) texture based ones, e.g., local binary pattern [22] and Gabor filter [18].

The last module, i.e., classification, is a hot topic in machine learning. Many classic classifiers are used in object classification and detection, e.g., Boosting, SVM and KNN. Also, kernel tricks, e.g., inter section kernel [3], are often used to enhance the overall performance.

### 2.2 Object Classification

In this subsection, we discuss the other modules in object classification, i.e., building feature space and describing features with the feature space.

**Building Feature Space.** Feature space consists of a group of base vectors. In particular, in the well-known bag-of-features model, the feature space is composed of a set of dictionaries, which are also called visual codes or codebook. There are three strategies to build the feature space, explained as follows.

The first one randomly chooses patches from images as the base vectors. This method is adopted in some biologically inspired models [27,14,13]. It is fast but does not sufficiently reflect the characteristics of the feature space.

The second one is based on supervised learning, i.e., generating dictionaries via supervised learning over local features. This scheme builds the relation between features and image labels, and well reflects the structure of the feature space. However, it is time-consuming because it needs to iteratively resolve dictionaries. For more details, readers are referred to the literature [5,20].

The third one is based on unsupervised learning, i.e., obtaining the base vectors via unsupervised learning over local features. This strategy finds a good balance between accuracy and speed, and is widely used in recent methods.

**Describing Features.** Describing features is a very important component in object classification, and greatly influences image classification in both accuracy and speed. Existing coding strategies can be divided into the following categories:

*Voting-based methods* [7,12] describe the distribution of local features, reflecting the occurrence information of visual codes.

*Fisher coding-based methods* [23,24] calculate the distribution of local features with the Gaussian Mixture Models. Each Gaussian model describes a kind of local features.

*Reconstruction-based methods* [32,29] encode a feature by least-square-based optimization with constraints on the number of codewords for reconstruction.

*Local tangent-based methods* [33,38] firstly estimate the manifold of the feature space, based on which an exact description of local features is derived.

*Saliency-based methods* [15,31] depict a local feature by the saliency degree, e.g., the ratio of the distances from a local feature to the codewords around it.

For more details of feature coding, readers are referred to our recent paper [16], which provides a comprehensive study about feature coding.

### 2.3 Object Detection

A typical object detection system is composed of four major steps: window search, object representation, machine learning and optimization, and post-processing. For the sake of space, we only introduce window search and object representation in this subsection.

**Window Search:** Most existing approaches follow the sliding-window paradigm [8,10,28,35,26,9]. In this case, generic object detection is formulated as a binary classification task. In the detection stage, the detector model evaluates each sliding window across scales and locations in an image, and then thresholds it as an object or not. So the simplest method is applying exhaustive search. In contrast to exhaustive search, there are several approaches on heuristic window search for the purpose of narrowing search space and accelerating detection.

Exhaustive search can be found in typical detection methods [8,35,9]. The advantage of exhaustive search is that it can obtain a relative high recall rate, reducing missing rate. But the large search space makes it intractable for real time object detection. Motivated by this challenge, heuristic window search methods have been developed. Lampert *et al.,*[17] propose efficient subwindow search (ESS) with branch and bound strategy. This method is based on sparsely detected features. If it is built on densely extracted features, the efficiency of ESS cannot be guaranteed. We all know that the segmentation and salience information provide the prior for object location. Thus, there are also some studies on heuristic window search based on segmentation and salience analysis [28].

**Object Representation:** Representation models mainly include part based models [35,36,26,9,34] and rigid template models [8,28].

Part based representation was firstly proposed by Fischler and Eschlager [11]. In this model, an object is represented by several parts in a flexible and deformable configuration [11]. Each part is usually described or represented by a small rigid template. The spatial relationships between parts are considered by spring-like connections for structure description [6]. Therefore, part based model can be considered as a top-down structured model, which is robust to partial occlusion and appearance variations. Recently, excellent part based models [26,9] have shown their success on many difficult datasets [21]. Due to their robustness to deformation, occlusion, part based model is regarded as a promising method for object localization.

The other common representation model is the rigid template model [8,35]. Rigid template model describes an object in a holistic manner, so they cannot well capture the structure variations of objects. Therefore, they perform well on well conditioned databases but suffer from those challenging data with deformations and occlusions. Besides, both the part based model and the rigid template model are associated with low-level features. Thus, progress on low-level features helps the improvement of object representation greatly as well. One classic feature is histogram of oriented gradients (HOG) [8]. The others include scale-invariant feature transform (SIFT) [19] based bag-of-words feature, pairs of adjacent segments (PAS) [10], local binary patterns (LBP) [22],*etc.*

## 3 Our Work

In this section, we introduce our solutions to object classification and detection, which were ranked among the best in the PASCAL VOC competition.

### 3.1 Object Classification

Our system of object classification is based on the bag-of-features model [7], with four steps different from traditional ones as illustrated in Fig. 3 and explained as follows.

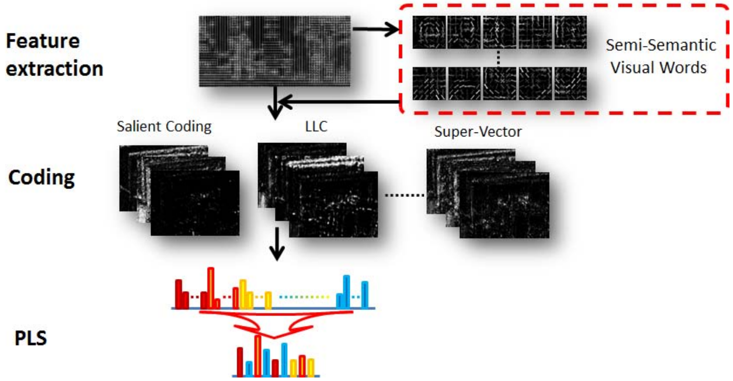


**Fig. 3.** Our system of object classification in VOC

1. We use different kinds of low level features: SIFT, HOG, LBP, color descriptor and Gabor filters. They play different roles in object description.
2. We apply parts learning from the deformable part-based model [9]. The learnt parts are augmented as a new set of dictionaries, which is demonstrated to be discriminative in differentiating objects from different classes.
3. We choose three methods for feature coding: super-vector coding [38], local constraint liner coding [29] and salient coding [15]. These methods complement each other in feature coding, which helps to improve the overall performance of object classification.
4. With different feature description and feature coding methods, the final dimensionality is very high. For dimensionality of more than a million for each image, which dimensionality reduction method should we use? We find that PLS [30] is good choice in this case.

### 3.2 Object Detection

Our detection system is based on the local structured model. In VOC2010, at the feature level, we propose Local Structured Descriptor and develop new descriptors from shape and texture information, respectively. Secondly, at the topology level, we present a local structured part representation with boosted feature selection and fusion scheme. Fig. 4(a) shows the framework we used in VOC2010.

The system includes two parts: building features and training local structured part detectors. The first part includes extracting local structured descriptors and feature selection of local structured LBP in a supervised manner. In the second stage, we firstly train the root model or holistic model using the learnt features from the first stage, then initialize local structured part appearance models from the root model. Linear SVM is applied to optimize the parameters. For more details, we refer readers to our previous papers [35,34].

The method mentioned above has two limitations: the model complexity is high and the model is still not "deformable" enough. Motivated by these challenges, we propose a data decomposition and spatial mixture modeling method

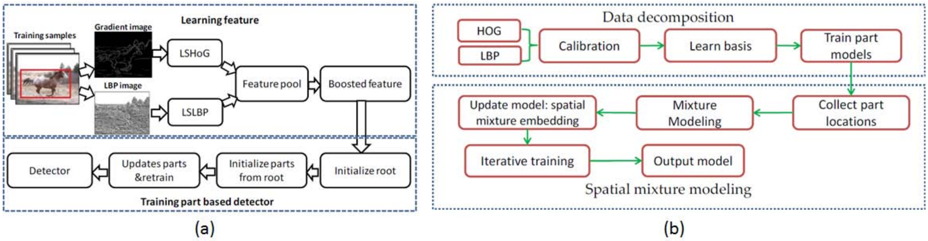


**Fig. 4.** System framework used in VOC2010 [35] and VOC2011 [36]

[36,37] in VOC2011 as shown in Fig. 4(b). Firstly, data decomposition is developed for the part based model, which not only largely reduces memory usage and computational cost but also outperforms other related systems. Secondly, a spatial mixture modeling method in which part location is described as a mixture distribution learnt from weakly labeled data, is proposed for more flexible structure description. Thirdly, we unify the spatial mixture model into the data decomposition framework. To the best of our knowledge, the presented system achieves the state-of-the-art performance compared with all other related methods from both the competition and the open literature. Due to limit of space, we refer readers to [36,37] for details.

## 4 Future Work and Conclusions

Object classification and detection, despite of decades research, remain two very active research topics in computer vision and pattern recognition. Every year, more than one hundred related papers appear in various top conferences and journals. Also, it should be recognized that there are still many challenges to be solved as we discussed in Introduction. Based on the analysis in this paper and our own experience, we think that the following directions deserve more attention:

- For object classification, the spatial relations of local features and the structure information of objects are still not well exploited. We believe that the progress in these two problems will greatly enhance current object classification models. Besides, representation learning [4] has shown good potential and provide new perspective on understanding object classification.
- For object detection, current methods mainly focus on learning structure parameters from data but ignore jointly learning structure topology and structure parameters. A potential direction is learning them together from data. Secondly, big data not only bring challenges but also opportunities for object detection. How to take advantage of big data with the latest machine learning techniques and biological observations is also promising.
- In addition, the integration of object classification and detection at the representation level is also an interesting direction for future work.

**Acknowledgments.** This work is jointly supported by National Basic Research Program of China (2012CB316300) and National Natural Science Foundation of China (61135002, 61203252).

## References

1. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/index.html
2. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/index.html
3. Barla, A., Odone, F., Verri, A.: Histogram intersection kernel for image classification. In: Proc. IEEE Inter. Conf. Image. Process. (2003)
4. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE T-PAMI 35(8), 1798–1828 (2013)
5. Bradley, D.M., Bagnell, J.A.: Differential sparse coding. In: Proc. Neu. Inf. Process. Sys. (2008)
6. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2005)
7. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Proc. Eur. Conf. Comput. Vis. (2004)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2005)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T-PAMI 32(9), 1627–1645 (2010)
10. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. IEEE T-PAMI 30(1), 36–51 (2008)
11. Fischler, M., Elschlager, R.: The representation and matching of pictorial structures. IEEE Trans. Comput. C-22(1), 67–92 (1973)
12. van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., Smeulders, A.W.M.: Kernel codebooks for scene categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
13. Huang, Y., Huang, K., Tao, D., Tan, T., Li, X.: Enhanced biological inspired model for object recognition. IEEE T-SMC-Part B 41(6), 1668–1680 (2011)
14. Huang, Y., Huang, K., Tao, D., Wang, L., Tan, T., Li, X.: Enhanced biological inspired model. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2008)
15. Huang, Y., Huang, K., Yu, Y., Tan, T.: Salient coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2011)
16. Huang, Y., Wu, Z., Wang, L., Tan, T.: Feature coding in image classification: A comprehensive study. IEEE T-PAMI (accepted, 2013)
17. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond sliding windows: Object localization by efficient subwindow search. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2008)
18. Lee, T.S.: Image representation using 2D gabor wavelets. IEEE T-PAMI 18, 959–971 (1996)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110 (2004)
20. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Supervised dictionary learning. In: NIPS (2008)
21. Mark, E., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV 88(2), 303–338 (2010)

22. Ojala, T., Petikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proc. IAPR Inter. Conf. Pattern Recognit. (1994)
23. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2007)
24. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
25. Sande, K., Gevers, T., Snoek, C.: Evaluation of color descriptors for object and scene recognition. IEEE T-PAMI 32(9), 1582–1596 (1998)
26. Schnitzspan, P., Roth, S., Schiele, B.: Automatic discovery of meaningful object parts with latent crfs. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2010)
27. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE T-TPAMI 29(3), 411–426 (2007)
28. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proc. IEEE Inter. Conf. Comput. Vis. (2009)
29. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2010)
30. Wold, H.: Partial least squares. In: Encyclopedia of Statistical Sciences (2004)
31. Wu, Z., Huang, Y., Wang, L., Tan, T.: Group encoding of local features in image classification. In: Proc. IAPR Inter. Conf. Pattern Recognit. (2012)
32. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2009)
33. Yu, K., Zhang, T.: Improved local coordinate coding using local tangents. In: Proc. Int. Conf. Mach. Learning (2010)
34. Yu, Y., Zhang, J., Huang, Y., Zheng, S., Ren, W., Wang, C., Huang, K., Tan, T.: Object detection by context and boosted hog-lbp. In: ECCV workshop on PASCAL VOC (2010)
35. Zhang, J., Huang, K., Yu, Y., Tan, T.: Boosted Local Structured HOG-LBP for Object Localization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2011)
36. Zhang, J., Huang, Y., Huang, K., Wu, Z., Tan, T.: Data decomposition and spatial mixture modeling for part based model. In: Proc. Asi. Conf. Compt. Vis. (2013)
37. Zhang, J., Yu, Y., Huang, Y., Wang, C., Ren, W., Wu, J., Huang, K., Tan, T.: Object detection based on data decomposition, spatial mixture modeling and context. In: International Conference on Computer Vision Workshop on Visual Object Classes Challenge (2011)
38. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 141–154. Springer, Heidelberg (2010)

# Directional Convexity Measure for Binary Tomography*

Tamás Sámuel Tasi, László G. Nyúl, and Péter Balázs

Department of Image Processing and Computer Graphics
University of Szeged
Árpád tér 2., 6720, Szeged, Hungary
{ttasi,nyul,pbalazs}@inf.u-szeged.hu

**Abstract.** There is an increasing demand for a new measure of convexity for discrete sets for various applications. For example, the well-known measures for h-, v-, and hv-convexity of discrete sets in binary tomography pose rigorous criteria to be satisfied. Currently, there is no commonly accepted, unified view on what type of discrete sets should be considered nearly hv-convex, or to what extent a given discrete set can be considered convex, in case it does not satisfy the strict conditions. We propose a novel directional convexity measure for discrete sets based on various properties of the configuration of 0s and 1s in the set. It can be supported by proper theory, is easy to compute, and according to our experiments, it behaves intuitively. We expect it to become a useful alternative to other convexity measures in situations where the classical definitions cannot be used.

**Keywords:** Binary Tomography, Discrete Geometry, Convexity Measure.

## 1 Introduction

Convexity is a crucial geometrical feature of a discrete set, e.g. in binary tomography[6], where the aim is to reconstruct binary images from their projections. Several reconstruction methods utilize preliminary information — such as horizontal (h), vertical (v) or both horizontal and vertical (hv) convexity — about the set to be reconstructed [2,4,5,7]. However, definitions for convexity are strict, in the sense that a change in a single pixel of the corresponding binary image could cause the set to lose its horizontal and/or vertical convexity,

* L.G. Nyúl was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. P. Balázs was supported by the OTKA PD100950 grant of the National Scientific Research Fund. The research of P. Balázs and T.S. Tasi was partially supported by the European Union and the State of Hungary co-financed by the European Social Fund under the grant agreement TÁMOP 4.2.4.A/2-11-1-2012-0001 ('National Excellence Program') and under the grant agreement TÁMOP-4.2.2.A-11/1/KONV-2012-0073 ('Telemedicine-focused research activities on the field of Mathematics, Informatics and Medical sciences').

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 9–16, 2013.

thus the previous methods cannot be applied anymore. Instead of providing a binary property to determine whether a set is convex or not, we prefer to assign each discrete set a degree of convexity. One could view this as *fuzzifying* the set of convex shapes by determining a membership value for each shape. Such a measure of convexity describes the image better than a binary value, and it should be more robust to noisy data. Therefore, it can be used to give a more detailed feature of the image, and even guide the reconstruction, in case such task is performed.

Various continuous and several discrete convexity measures have been introduced in image processing over the years, most of them belonging to a few, well-defined categories. Area based measures have been popular for quite some time [3,11,12], as well as boundary-based ones, like [13]. Other methods use simplification of the contour [8] to derive a shape hierarchy, or even use a probabilistic approach [9,10] to solve the problem. Our proposed method falls into the latter category, but takes a different approach. Instead of an approximation based on random inner points, or on certain pixels on the boundary, it treats all points equivalently.

The structure of the present paper is the following. In Section 2 we present the preliminaries for our problem and some basic features of images in binary tomography. Section 3 introduces our new convexity measure. In Section 4 we present few experimental results, and finally Section 5 is for the conclusion.

## 2 Preliminaries

### 2.1 Definitions and Notation

Let us consider the two-dimensional integer lattice $\mathbb{Z}^2$ on the plane. Any finite subset of this lattice is called a *discrete set*. A discrete set cannot only be represented by its elements, but also by a binary matrix or a binary image.

Adjacency in binary tomography is defined as follows: two positions $P = (p_1, p_2)$ and $Q = (q_1, q_2)$ in a discrete set are *4-adjacent* if $|p_1 - q_1| + |p_2 - q_2| = 1$. A discrete set $F$ is called *4-connected* if for two arbitrary positions $P, Q \in F$, there exists a sequence of distinct positions $P = P_0, P_1, \ldots, P_l = Q$ in the set $F$, so that $P_i$ is 4-adjacent to $P_{i-1}$, respectively, for each $i = 1, \ldots, l$. Sometimes we also call 4-connected discrete sets *polyominoes*. A discrete set $F$ is *h-convex* if no individual row contains any intervening 0s in the sequence of 1s, *v-convex* if no individual column contains any intervening 0s in the sequence of 1s, and *hv-convex* if both conditions are met.

Run-length encoding (RLE) is a simple, yet useful form of discrete data representation, mainly used for data compression purposes. Instead of storing data as is, each sequence in which the same data value appears in consecutive elements is stored as a single data value and a counter. This, of course, is effective only if there are relatively few, but preferably long runs of identical values in the data. This representation becomes highly beneficial when calculating convexity of rows or columns of discrete sets. A *token* or *run* is a maximal sequence in which the same data value occurs in consecutive data elements. The *length* of a given token is the number of occurences of the same data value in that particular token. A *1-token* is a token of 1s and a *0-token* is a token of 0s.

Consider, for example the following row of length 15: 001111100011111. We can represent this row in a more visible way, $0^2 1^5 0^3 1^5$, where the superscripts represent the length of each token (counters). The *span* of the data in a single row (column) is the distance between the outermost 1s in that row (column), while the *length* of the row (column) is the total number of bits present in that row (column). Obviously, the length is never smaller than the span. In our example the span of the row is 13, and its length is 15.

### 2.2 Basic Properties of a Convexity Measure

From a one-dimensional convexity measure we expect the following desirable, basic properties, several of which were also considered in [9,10]:

- It should give a value for all discrete sets in the interval [0, 1].
- For the least convex sets in the defined sense, it should give 0.
- It should give 1 if and only if the set is convex in the defined sense.
- It should be invariant under appropriate geometrical transformations.

In discrete geometry, thus in binary tomography the latter should be treated carefully. The measure is usually expected to be translation invariant, and it could be invariant with respect to uniform scaling. However, it is not expected to be rotation invariant, since elements of binary images are arranged in rows and columns, thus rotation is inherently problematic.

## 3 Introducing a New Convexity Measure

### 3.1 Preliminary Experiments

We carried out various early statistical experiments on discrete sets, mostly on the benchmark sets of [1] with added noise and distortions. We tried several combinations of the following features that, we believed, could truly describe the convexity of a sequence of binary digits: number of bit changes in the sequence, the span of the data, the length of the sequence, the number of 1-tokens in the sequence, the number of 0-tokens between the outermost 1-tokens in the sequence, etc. Although some results were promising, we concluded that each of these ad hoc measures unjustifiably favored some patterns as the most or least convex sets.

### 3.2 A Measure for Directional Convexity

To provide a measure that is objective and universally applicable, a theoretical approach is needed. One should consider the definition of convex shapes in the continuous domain. A planar shape $\mathbf{C}$ is said to be convex if for arbitrary points $A, B \in \mathbf{C}$, all points of the line segment $\overline{AB}$ belong to $\mathbf{C}$. One could determine the convexity of **C** by taking all possible pairs of points in **C** and measuring the proportion of points in the formed line segments that are not in **C** (the outer points). Of course, when **C** is convex, no line segment connecting points of **C** will contain outer points. Unfortunately, even if the shape is discretized, thus there is only a finite number of points, it is computationally too expensive to calculate all contributions of outer points in all possible line segments in **C**. To overcome this problem possible solutions so far include random sampling of inner points, randomly choosing points on the boundary only [10,13], or using probabilistic measures to estimate the convexity of the shape [9].

*Instead of limiting our calculations to only a random selection of points, we consider all pairs of points in the set and the line segments connecting them, but only in a few number of predefined directions.* This approach suits binary tomography better, since there are certain fundamental directions for describing and examining binary images, such as horizontal and vertical. Although throughout this paper we measure convexity in these two directions only, one could select any particular direction.

### 3.3 Calculating Directional Convexity

We compute *directional* convexity of a row (or column) in the following way. We split the row (column) into a sequence of 1-tokens and 0-tokens (see Section 2). From the construction above trivially follows that leading and trailing 0-tokens do not contribute to the measure, thus hereafter we shall omit them. The rest of the row (column) can be encoded as $1^{k_1}0^{l_1}1^{k_2}0^{l_2}\dots 1^{k_n}$, where $n$ is the number of 1-tokens and $k_1, l_1, k_2, l_2 \dots, k_n > 0$. Trivially, taking two 1s from the same 1-token, the line segment connecting them will not contain any 0s and will not contribute to the convexity measure. Now, let us take two arbitrary 1s from different 1-tokens, say the $i$th and $j$th, such that $i < j$. The contribution of 0s (outer points) in the line segment that connects them is

$$\sum_{t=i}^{j-1} l_t \,. \tag{1}$$

For two different 1-tokens ($i$th and $j$th), we can form $k_i k_j$ possible pairs of 1s, by picking one from each. The contribution of this particular 1-token pair is

$$k_i k_j \sum_{t=i}^{j-1} l_t \,. \tag{2}$$

Finally, to get the contributions for the entire row (column) one has to sum up (2) for all possible $\binom{n}{2}$ combinations of 1-token pairs:

$$\varphi = \sum_{1 \le i < j \le n} k_i k_j \sum_{t=i}^{j-1} l_t \,. \tag{3}$$

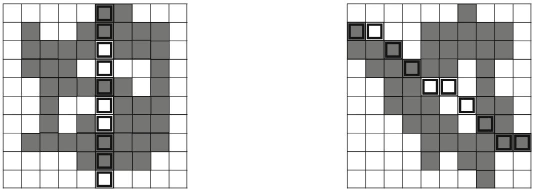


**Fig. 1.** A 4-connected discrete set with the highlighted column 1100100110 (left). For that particular column $\varphi = 2 \cdot 1 \cdot 2 + 2 \cdot 2 \cdot 4 + 1 \cdot 2 \cdot 2 = 24$. Another 4-connected discrete set (right), with a line segment in a different direction. The highlighted line segment to be used for the calculation in this case is 1011000111, with $\varphi = 32$.

The higher $\varphi$ is, the less convex the row (column) is. Therefore, $\varphi$ actually indicates the *directional non-convexity* of the row (column), rather than the directional convexity. Later, we shall describe a way to define a directional convexity measure based on $\varphi$.

Figure 1 shows an example discrete set on the left, represented by its binary image and the calculation of $\varphi$ of its highlighted column. On the right, it shows a different discrete set and a discrete line segment that connects two of its inner points, emphasizing on the fact that the same calculations can be made for arbitrary discrete directions.

### 3.4 Normalizing Directional Non-convexity

To obtain a normalized measure, it is required to know what is the maximum value a single row can produce for (3), i.e. which is the least convex row according to our measure. Initially we ran simulated annealing to find rows with such property, but eventually it became clear that such rows have the form of $1^{K/3}0^{K/3}1^{K/3}$, where $K$ is the length of the row and $K \equiv 0 \pmod 3$. (In case $K \equiv 1 \pmod 3$ is true, then $1^{\lfloor K/3 \rfloor}0^{\lceil K/3 \rceil}1^{\lfloor K/3 \rfloor}$ is the correct form, while if $K \equiv 2 \pmod 3$ is true, then $1^{\lceil K/3 \rceil}0^{\lfloor K/3 \rfloor}1^{\lceil K/3 \rceil}$ is the correct form.) The following two lemmas form the basis of this fact.

**Lemma 1.** *Let a row be given such that $1^{k_1}0^{l_1}1^{k_2}0^{l_2}\ldots 1^{k_n}$, with $k_1 = l_1 = \cdots = k_n = \frac{K}{2n-1}$. Then the non-convexity of the row is maximal if $n = 2$.*

*Proof.* There are $n-1$ pairs of 1-tokens having exactly one 0-token between them, $n-2$ pairs of 1-tokens having exactly two 0-tokens between them, and so on. Finally, there is one pair of 1-tokens having $n-1$ 0-tokens between them. For 1-tokens with exactly $i$ 0-tokens between them the non-convexity sum is

$$(n-i)\Big(\frac{K}{2n-1}\Big)^2 \frac{iK}{2n-1} . \tag{4}$$

Thus, the total non-convexity sum for the row is

$$\begin{aligned}\varphi_n &= \sum_{i=1}^{n-1}(n-i)\Big(\frac{K}{2n-1}\Big)^2\frac{iK}{2n-1} = \\ &= \frac{K^3}{(2n-1)^3}\sum_{i=1}^{n-1} i(n-i) = \frac{K^3}{(2n-1)^3}\left(n\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} i^2\right) = \\ &= \frac{K^3}{(2n-1)^3}\left(\frac{n^2(n-1)}{2} - \frac{(n-1)n(2n-1)}{6}\right) = \frac{K^3 n(n-1)(n+1)}{6(2n-1)^3}\ .\end{aligned} \tag{5}$$

Similarly,

$$\varphi_{n+1} = \frac{K^3(n+1)n(n+2)}{6(2n+1)^3}\ . \tag{6}$$

Then, for an arbitrary $n \geq 2$

$$\varphi_n - \varphi_{n+1} = \frac{K^3 n(n+1)}{6}\left(\frac{(n-1)(2n+1)^3 - (n+2)(2n-1)^3}{(2n-1)^3(2n+1)^3}\right) > 0 \tag{7}$$

since $(n-1)(2n+1)^3 - (n+2)(2n-1)^3 = 12n^2 - 16n + 1 > 0$. Thus, $\varphi_n$ is maximal if and only if $n = 2$. □

**Lemma 2.** *Let a row be given such that* $1^a0^b1^c$ *with* $a, b, c > 0$ *and* $K = a+b+c$. *Then the non-convexity of the row is maximal if* $a = b = c$.

*Proof.* From the definition it follows that $\varphi(1^a0^b1^c) = abc = ab(K-a-b)$. For the maximality of this expression the derivatives must be equal to 0, i.e.,

$$bK - 2ba - b^2 = 0 \quad \text{and} \quad aK - a^2 - 2ba = 0\ . \tag{8}$$

Knowing that $a, b > 0$ we get that

$$K - 2a - b = 0 \quad \text{and} \quad K - a - 2b = 0\ , \tag{9}$$

thus $2K - 3a - 3b = 0$ and therefore $\frac{2}{3}K = a + b$. Substituting $b = \frac{2}{3}K - a$ into (8) the lemma follows. □

Hence, the maximum value of non-convexity of a row (column) is $(K/3)^3$, where $K$ denotes the length of the row (column). Using this, the normalized non-convexity of a row (column) is

$$\hat{\varphi} = \frac{\varphi}{(K/3)^3}\quad . \tag{10}$$

### 3.5 Directional Convexity of a Two-Dimensional Discrete Set

The directional non-convexity of a two-dimensional discrete set can be defined as the mean of the normalized non-convexity value of all rows (columns). For

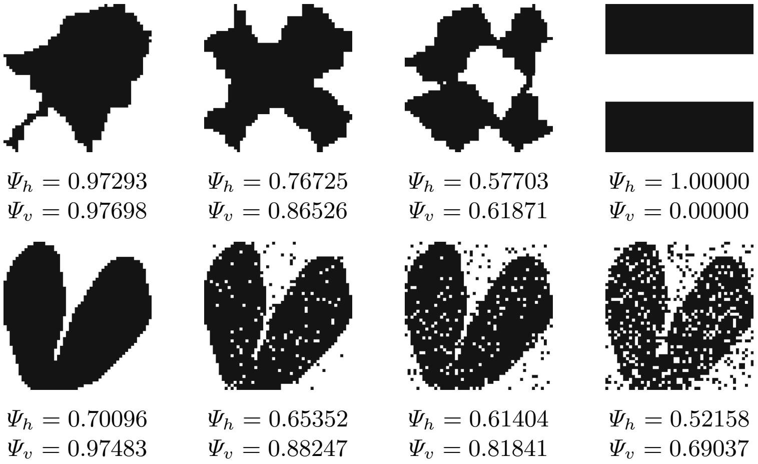


**Fig. 2.** Example binary images of size 50 × 50, with horizontal ($\Psi_h$) and vertical ($\Psi_v$) convexity shown. Bottom row: same image without, and with 5%, 10%, and 20% noise.

example, for an entire binary matrix consisting of $m$ rows, the directional non-convexity is

$$\Phi_h = \frac{\sum_{r=1}^{m} \hat{\varphi}_r}{m} \quad , \tag{11}$$

where $\hat{\varphi}_r$ is the normalized non-convexity of the $r$th row. From $\Phi_h$ one can simply derive a directional convexity measure for a discrete set by any monotonic continuous mapping from $[0, 1]$ to $[1, 0]$, e.g. $\Psi_h = 1 - \Phi_h$ can be considered such a measure. Analogously we can define the vertical convexity $\Psi_v$.

## 4 Experimental Results

The proposed method for measuring directional convexity of binary images along a defined direction has been tested thoroughly. Most of the images used were binary images derived from the 4-connected convex discrete sets of [1] by performing various operations resulting in a wide variety of non-convex images. Such operations included adding salt and pepper noise, applying morphological, topological, and set operations. Figure 2 shows a few display examples, along with the corresponding convexity measures.

We found that our method performs particularly well on noisy images, contrary to other methods that we tested that use the convex hull of the object or the span of each row or column. Another advantage is that the transition of the measure from convex to concave images is much smoother compared to other methods. With several other methods we experienced huge declines in function value caused by only a small distortion or noise in the image, which, we firmly believe, is unacceptable. Our model does not include any artificially favoured structure in the image to be considered the least convex.

## 5 Summary and Conclusion

In this paper we propose a new method to measure directional convexity of discrete sets. So far in all experiments we used horizontal and vertical directions, conventional in binary tomography, but the method works with any predefined direction as well. We are already working on the generalization to combine several directions to build a more global convexity measure for multidimensional discrete sets. We also have preliminary results on extending this work to measure convexity of not neccessarily binary discrete images, which may open ways to explore the connections between representations of binary, discrete, fuzzy, and continuous properties of shapes in images.

## References

1. Balázs, P.: A benchmark set for the reconstruction of hv-convex discrete sets from horizontal and vertical projections. Discrete Appl. Math. 157, 3447–3456 (2009)
2. Barcucci, E., Del Lungo, A., Nivat, M., Pinzani, R.: Medians of polyominoes: A property for the reconstruction. Int. J. Imag. Syst. Tech. 9, 69–77 (1998)
3. Boxter, L.: Computing deviations from convexity in polygons. Pattern Recogn. Lett. 14, 163–167 (1993)
4. Brunetti, S., Del Lungo, A., Del Ristoro, F., Kuba, A., Nivat, M.: Reconstruction of 4- and 8-connected convex discrete sets from row and column projections. Linear Algebra Appl 339, 37–57 (2001)
5. Chrobak, M., Dürr, C.: Reconstructing hv-convex polyominoes from orthogonal projections. Inform. Process. Lett. 69(6), 283–289 (1999)
6. Herman, G.T., Kuba, A. (eds.): Advances in Discrete Tomography and its Applications. Birkhäuser, Boston (2007)
7. Kuba, A., Nagy, A., Balogh, E.: Reconstruction of hv-convex binary matrices from their absorbed projections. Discrete Appl. Math. 139, 137–148 (2004)
8. Latecki, L.J., Lakamper, R.: Convexity rule for shape decomposition based on discrete contour evolution. Comput. Vis. Image Und. 73(3), 441–454 (1999)
9. Rahtu, E., Salo, M., Heikkila, J.: A new convexity measure based on a probabilistic interpretation of images. IEEE T. Pattern Anal. 28(9), 1501–1512 (2006)
10. Rosin, P.L., Zunic, J.: Probabilistic convexity measure. IET Image Process. 1(2), 182–188 (2007)
11. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision, 3rd edn. Thomson Learning, Toronto (2008)
12. Stern, H.: Polygonal entropy: a convexity measure. Pattern Recogn. Lett. 10, 229–235 (1989)
13. Zunic, J., Rosin, P.L.: A New Convexity Measure for Polygons. IEEE T. Pattern Anal. 26(7), 923–934 (2004)

# Biologically Inspired Anomaly Detection in Pap-Smear Images

Maykel Orozco-Monteagudo[1], Alberto Taboada-Crispi[1], and Hichem Sahli[2,3]

[1] Universidad Central de Las Villas, Cuba
{morozco,ataboada}@uclv.edu.cu
[2] Electronics & Informatics Dept. (ETRO), Vrije Universiteit Brussel (VUB),
Brussels, Belgium
hichem.sahli@etro.vub.ac.be
[3] Interuniverisity Microelectronics Center (IMEC), Leuven, Belgium

**Abstract.** Uterine Cervical Cancer is one of the most common forms of cancer in women worldwide. Papanicolau smear test is a well-known screening method of detecting abnormalities in the uterine cervix cells. In this paper we address the problem of anomaly detection in pap smear images. Our method avoids modeling the normal pap smear images which is a very complex task due to the large within class variance of the normal target appearance patterns. The problem is posed as a Visual Attention Mechanism. Indeed the human vision system actively seeks interesting regions in images to reduce the search export in tasks, such as anomaly detection. In this paper, we develop a new method for identifying salient regions in pap smear images and compare this to two previously reported approaches. We then consider how such machine-saliency methods can be used to improve human performance in a realistic anomaly detection task.

**Keywords:** microscopic images, anomaly detection, saliency, SVM classification.

## 1 Introduction

In general terms, Anomaly Detection (AD) is the process of discovering data that are different, in some sense, from the patterns defined by a observed data set. The main challenge in an AD system is how to define what an anomaly is, because, normally, we have the descriptions of the regularity of the problem. For that reason, AD is frequently defined in a negative way: the target is to determine what is not normal or what does not fill a specific rule.

In the special case of AD in medical image, a lot of works have been reported [1]. AD algorithms have been applied in the detection of tumours [2], malformations [3], abnormal cells [4], etc.

Cervical cancer, currently associated with the Human Papilloma Virus as one of the major risk factors, affects thousands of women each year. The Papanicolaou test (also known as Pap test) is used to detect pre-malignant and malignant

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 17–24, 2013.

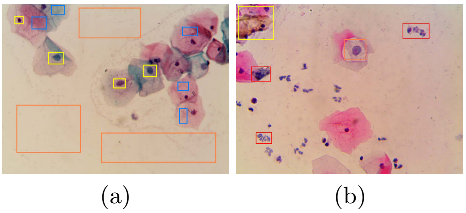


**Fig. 1.** Pap-smear cell images. (a) Dark blue parts (yellow squares) represents the nuclei. Pale blue and reddish parts (blue boxes) are the cytoplasms. Magenta parts (orange boxes) are the background. (b) Touching cells (surrounded in yellow). Isolated cells (surrounded in orange). Noise (surrounded in red).

changes in the cervix [5]. Cervical cancer can be mostly prevented by early detection of abnormal cells in smear tests.

As illustrated in Fig. 1, three classes of regions are considered: nucleus, cytoplasm, and regions that include background, noise, and other kinds of cells.

Developing automated algorithms for AD continues to pose interesting challenges. The goal of the present work is to develop an automated and computationally efficient algorithm for the detection of anomalies, associated principally with cancer, in Pap smear images.

Biological vision systems demonstrated a remarkable ability to recognize objects under adverse conditions, such as highly cluttered scenes. The use of saliency mechanisms is believed to play an important role in this robustness to clutter. They make salient locations "pop-out", driving attention to the appropriate regions of the visual field [6]. Biological vision systems rarely need to perform an exhaustive scan of a scene in order to detect an object of interest. These saliency detectors have been widely adopted in computer vision for applications such as object tracking, recognition and event analysis [7]. In these applications, saliency is used as a preprocessing step that saves computation and facilitates the design of subsequent stages.

Generally speaking, the detection of salient regions follows two paradigms: the object-based approaches [8,9], where the saliency is determined by the distribution of objects in the image, and the spatial-based approaches [10,11], where the saliency is determined by the selectivity of spatial locations. Both, object-based and spatial-based approaches, have drawbacks that one must take into account in developing biological vision systems. For the object-based approaches, it is necessary to develop robust mechanism for the detection of the objects in the image. On the other hand, spatial-based approaches usually fail when scenes are cluttered, the objects are overlapping, or objects distribution is heterogeneous or unbalanced. Many object-based saliency estimators have been reported in the literature. In most of them [12,13,7], the saliency of a region is determined from region's contrast, size, shape, and location.

In this work, anomaly detection based on Visual Attention Mechanism is proposed. The paper proposes a visual attention model to (i) first determine salient

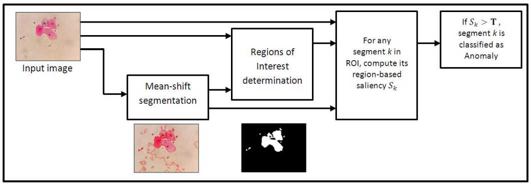

Input image
Mean-shift segmentation
Regions of Interest determination
For any segment $k$ in ROI, compute its region-based saliency $S_k$
If $S_k > \mathbf{T}$, segment $k$ is classified as Anomaly

**Fig. 2.** Framework for detection of anomalies in Pap smear images

areas defined as Region-of-Interest (ROI) where anomalies could lie, and (ii) a second a saliency measure characterizing abnormal cells for the final anomaly detection.

The reminder of the paper is organized as follows. Section 2 describes the framework for the detection of anomalies in Pap smear images. Section 3 presents and discusses the obtained results. Finally, conclusions are presented in Section 4.

## 2 Materials and Methods

The general framework for anomalies detection in Pap smear images is shown in Fig. 2. This process has two main steps: the selection of Region of Interest (ROI), and the salient regions detection.

Both steps use a segmentation of the image obtained using the Mean-Shift segmentation algorithm [14]. Mean-shift is a popular method to segment images and videos. It has been widely used in cell images segmentation with good results [15,16,17].

### 2.1 Selection of the ROI

The goal of Region-of-interest (ROI) extraction is to separate the part of the image that contains the objects of interest (i.e. cells, tissues, parts of cells, cancerous cells, anomalies) from the parts that contain unusable information (i.e. background, small noise areas, dirt particles).

In our approach, ROI detection is based on Visual Attention Mechanism, using the Selective attention algorithm proposed in [18]. The saliency map is calculated using a frequency-tuned approach based on low level features of colour and luminance. For the objects, region saliency value is calculated by averaging the pixel-based saliency map over the pixels of a region. This method separates the background and very large cytoplasms from the rest of the segments in the image (see Fig. 3) . This method, by itself, is not able to detect cells with abnormal changes.

### 2.2 Salient Regions Detection

Our method for anomaly detection is based, principally, on the method used by pathologists to determine some kinds of anomalies. The main characteristic of an abnormal cell in Pap test are [19]:

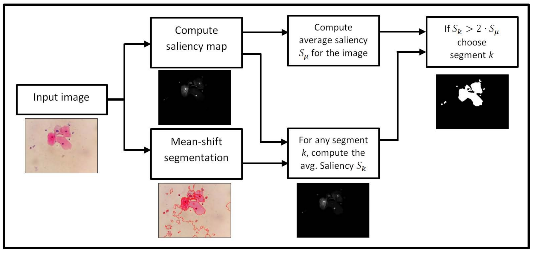

Input image
Compute saliency map
Mean-shift segmentation
Compute average saliency $S_\mu$ for the image
For any segment $k$, compute the avg. Saliency $S_k$
If $S_k > 2 \cdot S_\mu$ choose segment $k$

**Fig. 3.** Extraction of the ROI

- Nuclei are too big with an irregular texture without a clear pattern.
- Cytoplasms are too small.
- Cells are frequently crowded.

Formally, the saliency of a region $r_i$ at the segmentation $R = \{r_1, r_2, \cdots, r_n\}$ is given by:

$$\begin{aligned} S(r_i) = \zeta(r_i) \cdot \frac{1}{\phi(\Pr(r_i \in \text{Anomalies}))} \\ \cdot \frac{\text{AreaFactor}(r_i) \cdot \text{RelativeAreaFactor}(r_i)}{\text{Compactness}(r_i)} \end{aligned} \quad . \tag{1}$$

**ROI Belonging:** Function $\zeta(r_i)$ controls the belonging to the ROI.

$$\zeta(r_i) = \begin{cases} 0 \text{ if } r_i \notin \text{ROI} \\ 1 \text{ if } r_i \in \text{ROI} \end{cases} \quad . \tag{2}$$

**Area Factor:** The term AreaFactor$(r_i)$ is calculated as:

$$\text{AreaFactor}(r_i) = \frac{\text{Area}(r_i)}{\text{Area}(\widehat{r_i})} \tag{3}$$

where $\widehat{r_i}$ is the region in the ROI that contains $r_i$.

**Relative Area Factor:** The relative area factor of a region is the ratio of the area of the region and the area formed by the region and its neighbours. Formally, the relative area factor of a region is given by:

$$\text{RelativeAreaFactor}(r_i) = \frac{\text{Area}(r_i)}{\text{Area}(r_i) + \sum_{r_j \in \xi(r_i)} \text{Area}(r_j)} \tag{4}$$

where $\xi(r_i)$ is the set of neighbours of the region $r_i$.

**Compactness:** Compactness function takes its values up to 1 and it measures the compactness degree of a region where a perfect circle is the most compact region with compactness equal to 1.

$$\text{Compactness}(r_i) = \frac{\text{Perimeter}^2(r_i)}{4 \cdot \pi \cdot \text{Area}(r_i)} \quad . \tag{5}$$

**Potential of Being an Abnormal Cell:** Similar to [20], the parameter $\phi(\Pr(r_i \in \text{Anomalies}))$, represents the potential of being an abnormal cell, is here defined as follows:

$$\phi(\Pr(r_i \in \text{Anomalies})) = \frac{1}{1 + \Pr(r_i \in \text{Anomalies})} \tag{6}$$

where $\Pr(r_i \in \text{Anomalies})$ is the probability of $r_i$ to be an abnormal cell, given the feature vector $\mathbf{f}(r_i)$.

In our approach, $\Pr(r_i \in \text{Anomalies})$ is calculated by using the method of Platt [21], from the output of a two-class (normal or abnormal) SVM [22] trained by using as feature vector $\mathbf{f}(r_i)$, the mean of the $L$, $a$, and $b$ channels and the standard deviation of the $L$ channel of the region in the $Lab$ colour space. SVM training was carried out using a training set of images.

The parameters of the SVM classifier have been selected as follows. A linear kernel SVM and Gaussian kernel SVMs (with different values for $\sigma$) were trained by using a 10-fold cross-validation. A grid search method was used to select the best parameters of the SVM. The penalty parameter of the error $C$ was tested in $C = \{2^i : i = -1..14, \infty\}$, as well as the parameter of the Gaussian kernel $\sigma$ in $\sigma = \{2^i : i = -3..4\}$. The best performance was obtained for $C = 1024$ and a Gaussian kernel SVM with $\sigma = 1$. Finally,

$$\Pr(r_i \in \text{Anomalies}) = \frac{1}{1 + \exp(A \cdot f + B)} \ , \tag{7}$$

where $f$ is the output of the SVM, and the parameters $A$ and $B$ are fitted by using maximum likelihood estimation [21].

Finally, the detection of anomalies in Pap smear images will be done by the comparison with a threshold $T$. Regions with saliency greater that $T$ will be considered the foci of attention and classified as anomalies. Selection of $T$ was done empirically and it was fixed as $T = \max(0.01, p_{0.9})$, where $p_{0.9}$ is the 90% percentile of the saliencies of all the regions.

## 3 Results and Discussion

The proposed approach was applied to 40 images taken in the Gynaecological-Obstetrical Hospital of Santa Clara, Cuba. Pap smears were prepared by qualified technicians. The images were taken at 650x using a 319CU digital microscopy Camera. Regions that correspond with malignant or pre-malignant formations (abnormal growing of the nucleus) were manually segmented by experts as illustrated in Fig. 4(a) and (b).

The performance of the proposed anomaly detector, denoted as BIAD, has been evaluated via three classifier quality measures: precision, recall and F-measure [23]. True positive cases refer to the abnormal cells detected over the whole image. Abnormal cells that do not belong to the ROI are considered as false negatives.

In our evaluation, a detected anomaly region is considered as true anomaly if it intersects a ground truth anomaly region. Fig. 4 shows the anomalies detected using the proposed method in a typical pathological image.

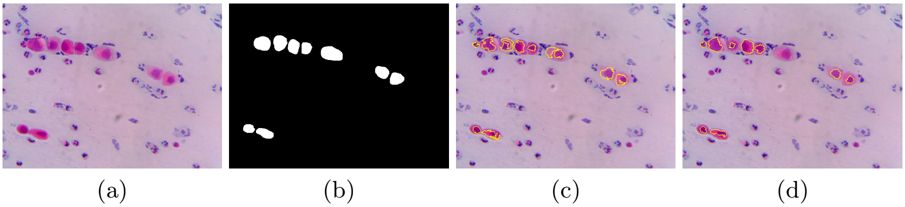


**Fig. 4.** Example of anomaly detection. (a) Original image. (b) Ground truth. (c) Detected anomalies using the proposed method over the original image. (d) Detected anomalies using the proposed method over the mean-shift smoothed image.

Table 1 shows the obtained results in terms of recall, precision, and F-measure, along with comparison to existing general purposes region-based focus of attention schemes, denoted as M1 [12], M2 [13], and M3 [7], respectively. The first three columns refer to the results using the original images, and the last three columns refer to the results using smoothed images. The Mean-Shift filter [14] has been used for image smoothing.

As it can be seen, the best results were obtained using the proposed method. Indeed, compared to the general purpose state-of-the-art methods, the introduced criteria used in the estimation of the saliency (area factor, relative area factor, and compactness) well-emulate some of the criteria used by pathologists. Moreover, the incorporation of a powerful machine learning technique, a SVM, contributes to the reduction of false negatives.

On the other hand, the results for the filtered images are worse than for the original images. Abnormal cells frequently lose important characteristics useful when you want to make an evaluation. This makes difficult to detect abnormal cell in these images. After this stage, these characteristics are smoothed. The key point is the delineation of the cytoplasm (Fig. 4). Our method tries to calculate the ratio between the area of the nucleus and the area of the cytoplasm. If cytoplasm is not well-segmented or, in the worst case, it is lost, then the evaluation of the saliency using our method does not provide good results. Indeed, if the segmentation method used to determine the regions is prone to oversegmentation, undersegmentation, or poor delineation of the regions the measure of the saliency will be widely biased.

**Table 1.** Results obtained with the proposed method and the comparison with another methods

| | **No Filtering** | | | **Mean-Shift Filtering** | | |
|---|---|---|---|---|---|---|
| | **Recall** | **Precision** | **F-measure** | **Recall** | **Precision** | **F-measure** |
| BIAD | 93.36% | 63.48% | 75.57% | 86.53% | 61.53% | 71.92% |
| M1 | 50.52% | 21.38% | 30.05% | 35.71% | 18.60% | 24.46% |
| M2 | 61.87% | 37.58% | 46.76% | 50.38% | 32.35% | 39.40% |
| M3 | 48.62% | 20.97% | 29.31% | 33.94% | 16.17% | 21.90% |

## 4 Conclusions

In this work, we introduced a visual attention mechanism for the detection of anomalies in images of the Papanicolaou test. First, a ROI extraction is performed in order to focus the analysis in suspected areas of the image, avoiding wasting computational resource in small regions, noise, dirt, etc. In the second step, we estimate the saliency of every region in the ROI by using a region-based saliency estimator. We use a supervised approach, via SVM, to estimate the probability of being an anomaly, in combination with cytological information of the abnormal cells. Finally, most salient regions (with a saliency greater than a predefined threshold) are considered as anomalies. As future work, we planned an extensive evaluation with pathologists.

**Acknowledgement.** This work was partially supported by the Canadian International Development Agency Project Tier II-394-TT02-00, the Flemish VLIR-UOS Programme for Institutional University Co-operation (IUC), and the Belgian Federal Science Polysy - European Space Agency PRODEX Programme.

## References

1. Taboada-Crispi, A., Sahli, H., Orozco-Monteagudo, M., Hernandez-Pacheco, D., Falcon-Ruiz, A.: Anomaly Detection in Medical Image Analysis. In: Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications. Medical Info Science Reference, pp. 426–446 (2009)
2. Strzelecki, M., Materka, A., Drozdz, J., Krzeminska-Pakula, M., Kasprzak, J.: Classification and segmentation of intracardiac masses in cardiac tumor echocardiograms. Computerized Medical Imaging and Graphics, 95–107 (2006)
3. Shinkareva, S., Ombao, H., Sutton, B., Mohanty, A., Miller, G.: Classification of functional brain images with a spatio-temporal dissimilarity map. NeuroImage, 63–71 (2006)
4. Goldberg, I., Allan, C., Burel, J.M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P., Swedlow, J.: The open microscopy environment (ome) data model and xml file: Open tools for informatics and quantitative analysis in biological imaging. Genome Biol. 5, 47 (2005)
5. Papanicolaou, G.: A new procedure for staining vaginal smears. Science 95, 438–439 (1942)
6. Itti, L., Koch, C., Niebur, E., et al.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 1254–1259 (1998)
7. Geerinck, T.: Visual attention framework: application to event analysis. PhD thesis, Vrije Universiteit Brussel (2009)
8. Baylis, G., Driver, J.: Visual attention and objects: evidence for hierarchical coding of location. Journal of Experimental Psychology: Human Perception and Performance, 451–470 (1993)
9. Vecera, S., Farah, M.: Does visual attention select objects or locations. J. Exper. Psychol. General, 146–160 (1994)
10. Wolfe, J.W.: Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review, 202–238 (1994)

11. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, 219–227 (1985)
12. Osberger, W., Maeder, A.: Automatic identification of perceptually important regions in an image. In: International Conference on Pattern Recognition, pp. 701–704 (1998)
13. Liu, H., Jiang, S., Huang, Q., Xu, C., Gao, W.: Region-based visual attention analysis with its application in image browsing on small displays. In: MM 2007, pp. 305–308 (2007)
14. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 603–619 (2002)
15. Li, N., Liu, J.X.: Automatic color image segmentation based on anisotropic diffusion and the application in cancer cell segmentation. In: 1st International Conference on Bioinformatics and Biomedical Engineering (2007)
16. Wenjia, B., Xiaobo, Z., Jinmin, Z., Liang, J., Wong, S.T.C.: Tracking of migrating glioma cells in feature space. In: 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (2007)
17. Bell, A.A., Herberich, G., Meyer-Ebrecht, D., Bocking, A., Aach, T.: Segmentation and detection of nuclei in silver stained cell specimens for early cancer diagnosis. In: IEEE International Conference on Image Processing (2007)
18. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 1597–1604 (2009)
19. Stanley, F., Patten, J.: Diagnostic cytopathology of the uterine cervix, 2nd edn. Monographs in clinical cytology, vol. 3. S. Kargel (1978)
20. Lucchi, A., Smith, K., Achanta, R., Lepetit, V., Fua, P.: A Fully Automated Approach to Segmentation of Irregularly Shaped Cellular Structures in EM Images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 463–471. Springer, Heidelberg (2010)
21. Platt, J.C.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In: Advances in Large Margin Classifiers, pp. 61–74 (1999)
22. Cristianini, N., Shawe-Taylor, J.: Introduction to Support Vector Machines and other kernel-basedlearning methods. Camridge University Press (2000)
23. Joshi, M.V.: On evaluating performance of classifiers for rare classes. In: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2002, pp. 641–655. IEEE Computer Society, Washington, DC (2002)

# Oriented Polar Snakes for Phase Contrast Cell Images Segmentation

Mitchel Alioscha-Perez[1], Ronnie Willaert[2], Helene Tournu[3,4], Patrick Van Dijck[3,4], and Hichem Sahli[1,5]

[1] Dept. Electronics & Informatics (ETRO), Vrije Universiteit Brussel (VUB), Belgium
[2] Research Group Structural Biology Brussels (SBB), Vrije Universiteit Brussel (VUB), Belgium
[3] Department of Molecular Microbiology, VIB, KU Leuven
[4] Laboratory of Molecular Cell Biology (MCB), KU Leuven, Leuven, Belgium
[5] Interuniversity Microelectronics Center (IMEC), Leuven, Belgium

**Abstract.** Noninvasive imaging of unstained living cells allows to study living specimens without altering them, and is a widely used technique in biotechnology for determining biological and biochemical roles of proteins. Fluorescence and contrast images are both used complementarily for better outcomes. However, segmentation of contrast images is particularly difficult due to the presence of lighting/shade-off artifacts, defocused scans, or overlapping. In this work, we make use of the optical properties intervening during the image formation process for cell segmentation. We propose the shear oriented polar snakes, an active contour model that implicitly involves phase information. Experimental results confirms the method suitability for cell images segmentation.

**Keywords:** active contours, image phase estimation, smart markers, image segmentation.

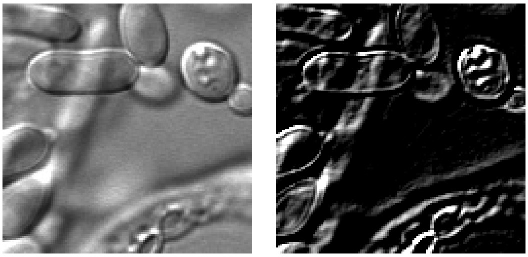
## 1 Introduction

Determining biological and biochemical roles of proteins is one of the critical tasks in biotechnology. Towards this goal, one of the most popular and successful technique is the noninvasive imaging of unstained living cells, which is mostly based on fluorescence and contrast imaging, and allows to study living specimens without altering them. Very often, fluorescence images with expression of (targeted) proteins and contrast images with individual cells information, are both used complementarily. Their combination provides better ways to measure the number of individual cells in populations, better sub-cellular localization of proteins, more precise individual cellular morphology measurements, cells population time-lapse evolution analysis, among other related tasks.

Considering images such as the one illustrated in Fig. 1(a), several approaches could be considered for segmentation, such as mathematical morphology [1], Level-sets [6] and active contours [7]. Authors in [4] estimate (restore) the phase information related to the image in question, and perform a subsequent level-set segmentation over the artifacts-free restored phase image, obtaining better results than segmenting the original phase contrast image. They showed how the resultant estimated image can be better segmented, even when using simple automatic thresholding techniques. Despite the

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 25–32, 2013.

mentioned improvements several drawbacks are still present. They use a more general/complex formulation for the phase estimation, but convexity on the related optimization problem can not be granted. Also, their strategy fails to cope with high-density overlapping (multilayer) cells cultures, where superposition of lighting/shade-off of individual cells introduces errors in the recovered phase (see Fig.1). As consequence, these errors have a high negative impact on the level-set segmentation since they use the estimated phase for the initialization of the level-set and for the segmentation as well.



(a) Phase Contrast Image. (b) Estimated Phase.

**Fig. 1.** Errors on the estimated phase are caused by the superposition of cells in multilayer cultures

Inspired by the work of Yin et al. [4], we estimate the phase using a simple and well known formulation that grants convergence towards the approximation, and we use a parametric active contour (snakes) with a novel energy functional.

In this paper we propose the shear oriented polar snakes [8], a parametric active contour model that: i) implicitly involves phase information on its energy functional using a provided shear-angle orientation, and ii) use the estimated phase image for the snakes initialization, but minimize the energy on the original image. The segmentation process is illustrated in Fig.(2).

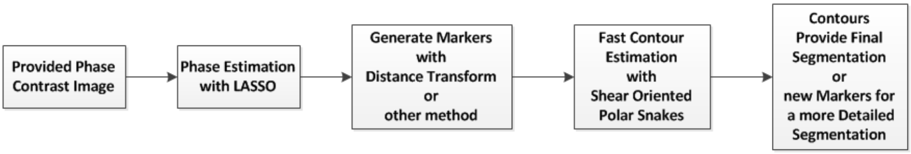

Provided Phase Contrast Image
Phase Estimation with LASSO
Generate Markers with Distance Transform or other method
Fast Contour Estimation with Shear Oriented Polar Snakes
Contours Provide Final Segmentation or new Markers for a more Detailed Segmentation

**Fig. 2.** The proposed segmentation process

The reminder of this paper is organized as follows: a review of phase-contrast optics and active contours models is provided in Section 2, along with the proposed energy model details. Experimental results are reported and discussed in Section 3, while conclusions are given in Section 4.

# 2 Oriented Polar Snakes

## 2.1 Phase Contrast Optics

Phase contrast imaging is used to image phase objects, which are almost transparent objects but with a refractive index different from their surrounding medium. The different optical path length, obtained from two mutually orthogonal light beams shifted by a vector $\tau(m, \theta)$, is transformed into intensity variations.

Let's consider a Point Spread Function (PSF) that represent the finite response of a focused optical system. Considering diffraction due to the microscope optics, a discrete PSF can be expressed as:

$$f(x, y) = (-x\cos\theta - y\sin\theta)\exp\left(-\frac{(x^2+y^2)}{\sigma^2}\right) \quad \forall x, y \in [-M, M] \tag{1}$$

which leads to the image model $I$ involving the PSF and the artifact-free phase image $\varphi$, as follows:

$$I(x, y) = \varphi(x, y) \otimes f(x, y) + \eta(x, y) \ . \tag{2}$$

being $\otimes$ a convolution operator, and $\eta(.)$ an additive noise function of unknown distribution.

The phase reconstruction problem consist in finding $\varphi$ from an observed $I$ (inverse problem). Despite of the simplicity of Eq. (2), its solution is seldom straightforward. Direct deconvolution is not possible to be performed due to the sensitiveness to noise.

One of the most popular ideas considers to find a $\varphi$ such that $\varphi \otimes f$ that is as close as possible to $I$, and use the squared residuals as penalization, named least squares. Since a regularization of $\varphi$ is preferable, the least absolute shrinkage and selection operator (LASSO) offers an attractive formulation [9]:

$$\min_{\varphi} ||I - \varphi \otimes f||^2 + \lambda||\varphi||_1 \tag{3}$$

The main benefit of this formulation is that it enforce sparsity on the solution due to the $\ell_1$-norm regularization, while keeping the optimization problem convex. In few words, it enforce the recovered phase to have as few pixels as possible.

## 2.2 Active Contour Models

The two main categories of active contours can be used: geometric and parametric snakes. In geometric snakes, the curve is described from a level set representation, while in parametric snakes it is described as a discrete collection of points. It is known that parametric snakes are much faster than geometric snakes, which support our model selection.

Let's consider the generic family of curves $\mathcal{C}_q$ depending on a $n$-dimensional parameter vector $q$ and defined in the image plane by:

$$\mathcal{C}_q : x(u) = x_c + r_q(u) \begin{pmatrix} \sin u \\ \cos u \end{pmatrix} \tag{4}$$

where $r_q(u)$ is the radius function depending on $u \in (0, 2\pi)$, and $x_c$ is a point inside the contour. Then, the total active contour energy can be defined as:

$$E_{ac} = \alpha \int_0^{2\pi} \left| \frac{\partial \mathcal{C}_q}{\partial u} \right|^2 du + \beta \int_0^{2\pi} \left| \frac{\partial^2 \mathcal{C}_q}{\partial^2 u} \right|^2 du + \int_0^{2\pi} E_{im}(\mathcal{C}_q) du \tag{5}$$

The first two terms, named potential energy or internal energy, accounts for elasticity (or extent) and curvature, respectively. The third term, the image energy, provides a cost evaluation of the contour according to the image. This terms is discretized as:

$$\int_0^{2\pi} E_{im}(\mathcal{C}_q) du = \sum_{k=0}^{N-1} E_{im}(\mathcal{C}_q, u_k) \ . \tag{6}$$

being $u_k = \frac{2\pi}{N} k$ (direction of $k$-th element). The external (image) energy can be defined in several ways, depending on the characteristics of the problem in question.

### 2.3 Energy Functional

In this work we propose the following external (image) energy functional:

$$E_{im}(\mathcal{C}_q, u_k) = \sin^2(u_k - \theta) \left| I_\mu(C_q(u_k)) \right| - \cos(u_k - \theta) I_\mu(C_q(u_k)) - \left| \nabla I(C_q(u_k)) \right| \tag{7}$$

with $\mu$ the intensity mean, $I_\mu(.) = I(.) - \mu$, $\nabla I(.)$ the intensity gradient, and $\theta$ the provided shear (angle) direction information.

The proposed energy enforce low energy values for $E_{im}$ whenever a lighting/shade-off effect is found in a parallel direction with respect to the provided shear angle, decreased even more if this match with a high gradient value.

### 2.4 Numerical Implementation

We opted for using a dynamic programming approach [2] to solve Eq.(5) based on the Bellman's Principle of Optimality [5], with a discrete-time finite horizon in a one-dimensional formulation (states related only to $r_q$). Since the initial number of markers can be considerable big, the use of a lower number of snakes elements (usually between $N = 45$ and $N = 90$) allowed us to find optimal solutions relatively fast. Depending on the parameters $\alpha$ and $\beta$, the execution time varies between 260 and 860 milliseconds per snake, using a Matlab implementation of the chosen numerical method. It is also worth noting that each individual snake can be optimized in parallel.

### 2.5 Segmentation

After the artifacts-free phase image has been estimated (see Fig. 1(b)), a distance transform is applied on it; then a local maxima points detection provides the set of initial markers, one point per marker. The set of automatically generated markers will be used for the initialization of the active contours (see Fig. 3a) as follows: each marker correspond to exactly one snake, that initially starts with radius $r_q(.) = 1$ and centroid $x_c$ on the corresponding marker. Once the snake models has been initialized, the next step is to optimize them by minimizing the active contour energy Eq.(5).

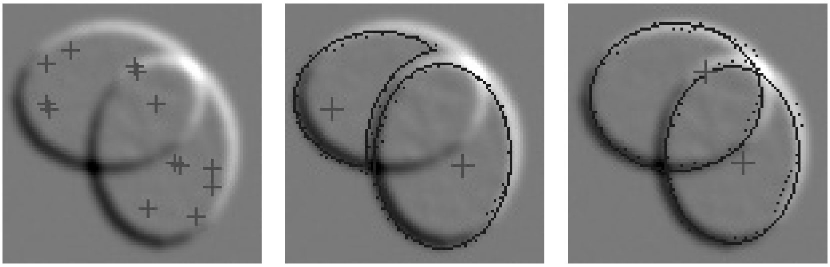
After all the snakes has been optimized, we expect that for initial markers located on the same region, their corresponding snakes converges towards the same boundaries. Then, likewise in [3], we perform a contour evaluation consisting in discarding those contours (and markers) overlapping in a high percent with another contour of higher priority, where the priority is given by evaluating the contour using Eq.(7).

## 3 Experimental Results

### 3.1 Qualitative Analysis

In order to highlight the benefits of the proposed energy functional of Eq.(7) with respect to the standard gradient-based $E_{im}(C_q, u_k) = -|\nabla I(C_q(u_k))|$, we used a simulated phase contrast image of two overlapping cells, and a real phase contrast image fragment where a cell is out of the focal plane.

The proposed model can find the appropriate contours (see Fig.3c), even without assuming any curvature in the contour ($\beta \approx 0$ in Eq. 5), while the gradient-based energy model fail to do so (see Fig.3b), requiring certain specific values of curvature parameter to find the appropriate contour.



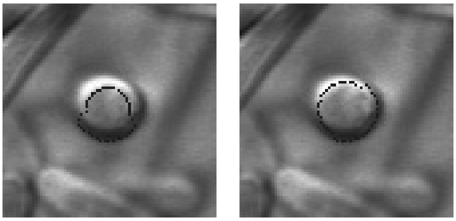
(a) Automatically generated initial markers.

(b) Gradient-based energy contours.

(c) Proposed energy contours.

**Fig. 3.** The proposed energy estimates the contours better than a standard gradient-based energy, even without assuming any curvature in the contour

We also show how the gradient-based energy model is very sensitive to the lighting/shade-off artifact, and the contour solution takes the extreme values of the gradient, resulting in a contour shifted in the shear-angle direction. The proposed model can better cope with this problem as can be seen in Fig.(4), where the most sensitive snake elements of our proposed solution are those orthogonal to the shear-angle direction.



(a) Gradient-based energy contour.

(b) Proposed energy contour.

**Fig. 4.** The proposed energy can better cope with lighting/shade-off artifacts

The proposed model can also deal properly with corners and broken lines, as shown in Fig.(5). In this simulated image with corners and shear angle $\theta = \frac{\pi}{4}$, the proposed energy prevented each contour's convergence to stop on the boundaries of the other rectangle, by enforcing high energy values (penalization) on dark boundaries located to the right of marker A; similarly on bright boundaries located to the left of marker B.

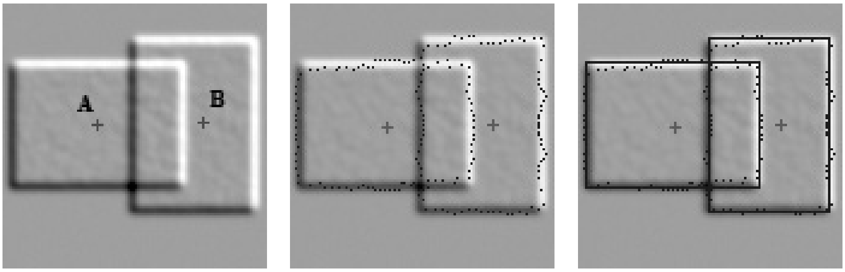
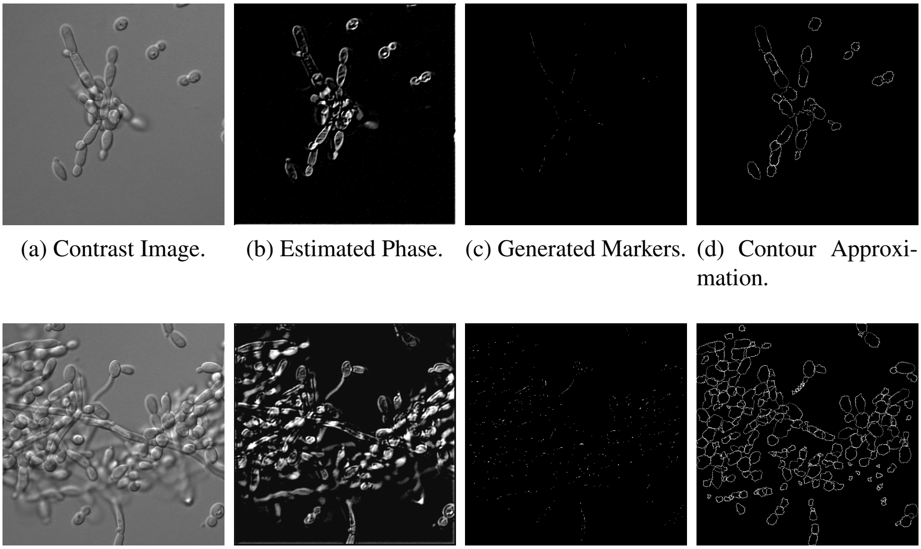


**Fig. 5.** The proposed energy can cope with other type of objects

### 3.2 Real Specimen Images

For the experiments on real specimen images, serum-induced cultures of the fungus Candida albicans were used in this study in order to obtain mixed populations of different cell morphologies of the fungus. Cells expressing two protein fusions to a monomeric GFP variant are displayed. Genomic copy tagging of HSP90 and RAS2 loci were performed in wild type strain SC5314. Cells were grown at 37°C for three hours and images were acquired by confocal microscopy (Olympus Fluoview$^{\text{TM}}$). Fluorescence images were obtained with Alexa Fluor 488 excitation laser.

The figures 6(a) and (e) depict the obtained images. As expected, the number of automatically generated markers (Fig. 6(c) and (g)) from the estimated phase (Fig. 6(b) and (f)) was higher than the number of cells in the image. The fact that most of the markers fall inside cell regions allowed the active contours to converge towards individual cell boundaries. As illustrated in Fig.(6(d) and (h)), the proposed contour evaluation process provides almost one contour/marker per individual cell in the image.



(a) Contrast Image. (b) Estimated Phase. (c) Generated Markers. (d) Contour Approximation.

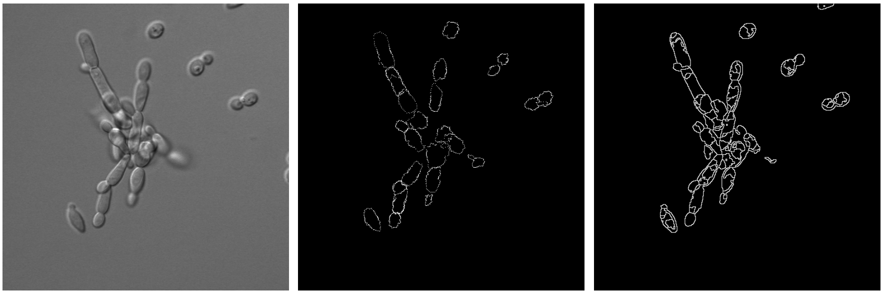
(e) Contrast Image. (f) Estimated Phase. (g) Generated Markers. (h) Contour Approximation.

**Fig. 6.** The estimated phase provides the basis to generate markers for snakes initialization

To assess the effectiveness of the proposed Oriented Polar Snakes, we compared it to the markers-controlled watershed segmentation. Both segmentation schemes have been initialized using exactly the same set of markers. As it can be seen from Fig.7, the proposed method produces well segmented individual cells compared to the over-segmented watershed. It has to be noted that, subsequent region merging of the obtained watershed segments is difficult due to the type of contrast image.

## 4 Conclusions

In this work we presented a segmentation method, based on an active contour model with a novel energy functional, that takes into account the nature of this type of images. The method has been developed specifically for contrast microscopy cells images, but can be easily extended and applied to any other type of phase contrast images. In future works we will perform a quantitative analysis of the method over different datasets of phase contrast images. We will also investigate different ways to enhance the initial marker selection in order to decrease the initial number of snakes.



(a) Original Contrast Image. (b) Proposed Segmentation. (c) Watershed Segmentation.

**Fig. 7.** The proposed segmentation provides more accurate individual region/cells than watershed segmentation, using the same markers for initialization

**Acknowledgments.** The Belgian Federal Science Policy Office (Belspo) and the European Space Agency (ESA) PRODEX program supported this work.

## References

1. Koyuncu, C.F., Arslan, S., Durmaz, I., Cetin-Atalay, R., Gunduz-Demir, C.: Smart Markers for Watershed-Based Cell Segmentation. PLOS ONE 7(11), e48664 (2012), doi:10.1371/journal.pone.0048664
2. Cohen, L.D., Kimel, R.: Global minimum for active contour models: A path approach. Int. Jour. Comp. Vis. 24, 57–78 (1997)
3. Wienert, S., Heim, D., Saeger, K., Stenzinger, A., Beil, M., Hunfnagl, P., Dietel, M., Denkert, C., Klauschen, F.: Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. Scientific Reports 2, 503–510 (2012)
4. Yin, Z., Kanade, T., Chen, M.: Understanding the phase contrast optics to restore artifact-free microscopy images for segmentation. Medical Image Analysis 16, 1047–1062 (2012)
5. Bellman, R., Dreyfus, S.: Applied Dynamic Programming. Princeton University Press, Princeton (1962)
6. Ambühl, M., Brepsant, C., Meister, J., Verkhovsky, A., Sbalzarini, I.: High-resolution cell outline segmentation and tracking from phase-contrast microscopy images. J. Microsc. 245(2), 161–170 (2012)
7. Seroussi, I., Veikherman, D., Ofer, N., Yehudai-Resheff, S., Keren, K.: Segmentation and tracking of live cells in phase-contrast images using directional gradient vector flow for snakes. J. Microsc. 247(2), 137–146 (2012)
8. Collewet, C.: Polar snakes: A fast and robust parametric active contour model. In: IEEE Int. Conf. on Image Processing (ICIP), pp. 3013–3016 (2009)
9. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. J. R. Static. Soc. 58(1), 267–288 (1996)

# Drug Activity Characterization Using One-Class Support Vector Machines with Counterexamples

Alicia Hurtado-Cortegana[1], Francesc J. Ferri[1,⋆], Wladimiro Diaz-Villanueva[1], and Carlos Morell[2]

[1] Departament d'Informàtica, Universitat de València, Spain
[2] Comp. Sci. Dept. Univ. Central "Marta Abreu" de Las Villas, Santa Clara, Cuba

**Abstract.** The problem of detecting chemical activity in drugs from its molecular description constitutes a challenging and hard learning task. The corresponding prediction problem can be tackled either as a binary classification problem (active versus inactive compounds) or as a one class problem. The first option leads usually to better prediction results when measured over small and fixed databases while the second could potentially lead to a much better characterization of the active class which could be more important in more realistic settings. In this paper, a comparison of these two options is presented when support vector models are used as predictors.

## 1 Introduction

Among supervised learning techniques developed and widely used in recent years, support vector machines (SVM) have received considerable attention due both to their success in solving practical problems and their mathematical soundness. One of the distinguishing trends of SVM is their capability of generalization in the context of hard learning problems. Consequently, the literature exhibits lots of classification, clustering or regression problems spanning diverse application domains that can be very conveniently solved using SVM [1,2,3].

Data domain description, also referred to as one-class classification (OCC) constitutes a different prediction task which consists of characterizing only one class of objects (and consequently rejecting the rest). Depending on how the problem is posed, the differences with regard to two-class classification can be very subtle. The most important difference is that OCC aims at modeling a particular class instead of separating objects from two classes which implies modeling their discriminating boundary. One of the main consequences is the way in which both approaches treat outliers and novelties [4].

OCC models can be learned either from examples only or both from examples and counterexamples. In any case, the problem consists of arriving at a decision function that covers all examples without including any other regions in the representation space and excluding also all counterexamples, if any.

[⋆] Work partially funded by FEDER and Spanish Government through projects TIN2009-14205-C04-03 and Consolider Ingenio 2010 CSD07-00018.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 33–40, 2013.

In the particular case of support vector based approaches, several formulations exist. In particular, One-Class Support Vector Machines (OC-SVM) [5] try to learn a hyperplane in the Reproducing Kernel Hilbert Space (RKHS) that keeps examples as far as possible from the origin. On the other hand, Support Vector Data Description (SVDD) [4] consists of obtaining a kernelized hypersphere that contains all examples. These two formulations have been shown to be equivalent under certain circumstances [6]. In a recent work, SVDD has been extended by introducing a separation margin between examples and counterexamples [7]. In this way, the model not only optimally represents the class of interest but also robustly separates both types of data at the same time.

The purpose of the present work is to study advanced OCC models on a particular difficult task in which binary SVM arrive at very good solutions. The goal consists of assessing possible benefits and disadvantages of using more complex models to solve these challenging problems.

## 2 Learning Problem Formulations

Only the SVDD formulation and extensions are to be considered in the present work. Assume that data belong to a $d$-dimensional vector space, $\mathbb{R}^d$. There is also a mapping $\phi$, from $\mathbb{R}^d$ to a RKHS, $\mathcal{H}$, which is implicitly given by a Mercer kernel function, $k : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^{\geq 0}$ in such a way that $k(\cdot, \cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$ [6] .

Let us suppose we have a non empty positive training set given by the examples corresponding to the class of interest, $\mathcal{X}^+ = \{x_1, \ldots, x_{\ell_1}\}$ and a negative training set which consists of zero or more counterexamples, $\mathcal{X}^- = \{x_{\ell_1+1}, \ldots, x_{\ell_1+\ell_2}\}$. The size of the overall training set, $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^-$, is given by $\ell = \ell_1 + \ell_2$. Each object from $\mathcal{X}$ has a corresponding label, $y_i$ such that $y_i = 1$ if $1 \leq i \leq \ell_1$ and $y_i = -1$ if $\ell_1 < i \leq \ell$.

When only positive examples are to be used, SVDD tries to enclose all objects into a minimal hypersphere in the RKHS [4]. The so-called soft formulation introduces additional slack variables controlled by a penalty term to allow objects outside the hypersphere.

The formulation of the problem using a $\nu$ parameter is

$$\min_{R,c,\xi} R^2 + \frac{1}{\nu \ell_1} \sum_{i=1}^{\ell_1} \xi_i, \tag{1}$$

$$\text{subject to:} \quad \left( \|\phi(x_i) - c\|^2 - R^2 \right) \leq \xi_i, \tag{2}$$

$$\xi_i \geq 0. \tag{3}$$

By introducing a Lagrange multiplier, $\alpha_i$, for each constraint it is possible to go from this primal formulation to its corresponding dual in which the optimization is over a vector, $\alpha = (\alpha_1, \ldots, \alpha_{\ell_1})$, which consists of all Lagrange multipliers in the primal problem.

$$\max_{\alpha} \sum_{i=1}^{\ell_1} \alpha_i k\left(x_i, x_i\right) - \sum_{i,j=1}^{\ell_1} \alpha_i \alpha_j k\left(x_i, x_j\right), \tag{4}$$

$$\begin{gathered}\text{subject to } 0 \leq \alpha_i \leq \frac{1}{\nu \ell_1}, \\ \sum_i \alpha_i = 1. \end{gathered} \tag{5}$$

This quadratic problem can be solved using the same methods as for binary SVMs. Once $\alpha$ has been obtained, the center of the hypersphere, $c$ can be obtained from the additional constraint $c = \sum_{i=1}^{\ell_1} \alpha_i \phi(x_i)$. Correspondingly, the radius, $R$, can be obtained exactly in the same way as the bias of the linear function is computed in the case of binary SVMs [5]. The final characterization of the positive class is then given by the following decision function

$$f\left(x\right) = sgn\left(R^2 - \left\|\phi\left(x\right) - c\right\|^2\right) \tag{6}$$

The basic approach can be extended by introducing negative objects (counterexamples) and the corresponding constraints that keep them outside the hypersphere [8]. This introduces a sign (the label $y_i$) in the constraints and a new summation term in Eq. 1. Both summation terms will be now weighted by $\frac{\gamma}{\nu\ell}$ and $\frac{1-\gamma}{\nu\ell}$, respectively. $\gamma$ is a new parameter that controls the relative importance of the constraint violations in both positive and negative cases which may be very important in specific practical problems exhibiting some kind of imbalance.

Apart from keeping positive data inside the hypersphere and negative data outside, it is possible to impose a (maximal) margin between the negative objects and the boundary of the hypersphere. This is the rationale of the Small Sphere and Large Margin (SSLM) approach [7]. The formulation of the corresponding primal problem in our particular context is:

$$\min_{R,c,\rho,\xi} R^2 - \eta\rho^2 + \frac{\gamma}{\nu\ell} \sum_{i=1}^{\ell_1} \xi_i + \frac{1-\gamma}{\nu\ell} \sum_{i=\ell_1+1}^{\ell} \xi_i \tag{7}$$

$$\begin{aligned} \text{subject to } & \left\|\phi\left(x_i\right) - c\right\|^2 \leq R^2 + \xi_i, && 1 \leq i \leq \ell_1 \\ & \left\|\phi\left(x_i\right) - c\right\|^2 \geq R^2 + \rho^2 - \xi_i, && \ell_1 < i \leq \ell \\ & \xi_i \geq 0, && 1 \leq i \leq \ell \end{aligned} \tag{8}$$

In this extended formulation, apart from the parameter $\nu$ that controls how strict the characterization must be, and the parameter $\gamma$ that controls the trade off between positive and negative outliers, a new parameter $\eta$ that moderates the maximization of the margin has been introduced. The margin is represented by a new variable, $\rho$.

These three OCC models constitute a family of predictors with increasing level of complexity. The more complex models need more parameters and the corresponding tuning gets harder. On the other hand, the more complex models are able to attain better characterizations with improved separation which will potentially lead to better generalization abilities.

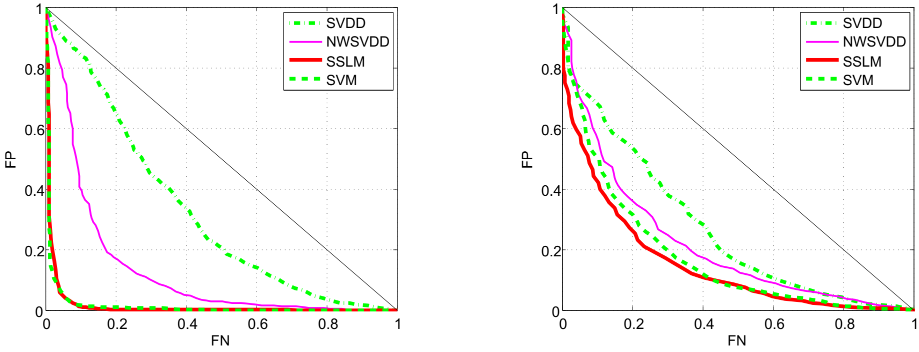
## 3 Drug Activity Prediction from Molecular Structure

The design of new medical drugs with desired chemical properties has a capital importance for the pharmaceutical industry. Several approaches are used in drug discovery, which can be grouped in three main categories: random screening of a large number of compound in a blind way, structural modifications of lead compounds and rational drug design [9]. Quantitative structure-activity (structure-property) relationships (QSAR/QSPR) constitute a methodology in the last category that is based on the fact that some properties of a set of molecules change with their molecular structure and therefore it is possible to find a relationship between this structure and the properties that the molecule exhibits. Once this relationship has been obtained it can be used to predict the properties of new, perhaps unknown, compounds.

Molecular descriptors used in QSAR can be empirical (derived from experimental data) or nonempirical. Among the nonempirical descriptors, the so-called topological indices have special relevance [10]. Topological indices are molecular descriptors derived from information on connectivity and composition of a molecule and can be easily derived from the hydrogen-suppressed molecular representation seen as a graph [11,12]. Some examples of topological indices are the popular Kier and Hall connectivity index [13] and Balaban index of average distance sum connectivity [14]. In this work, a set of 116 indices has been selected from three families considered that we will refer to as topological [15], the above mentioned Kier-Hall and the electro-topological or charge index[16]. Some experiments have been carried out using a reduced set formed by the 62 topological indices.

To properly assess the different predictors in this context, Receiver Operating Characteristic (ROC) curves and associate performance measures have been considered in this work [17]. Given a particular predictor whose output consists of a continuous value in a specified interval (as in this work), the ROC curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form (FP,TP). The performance of any classifier (with the corresponding threshold included) can be represented by a point in the ROC space. ROC curves move from the "all-inactive" point (0,0) which corresponds to the highest value of the threshold to the "all-active" point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A common method to compare classifiers is to calculate the area under



a) Antibacterial database b) Analgesic database

**Fig. 1.** ROC curves corresponding to the different predictors considered

the ROC curve (AUC). The value of the AUC will always be between 0.5 and 1.0, because random guessing produces the diagonal line between (0,0) and (1,1), which has an area of 0.5. The AUC has some important statistical properties [17] and is frequently used as a global measure of predictiveness.

## 4 Experimental Results

Several comparative experiments have been carried out using a wide range of settings for the algorithms considered. Two specific datasets containing chemical compounds have been considered. First, an small dataset of 434 compounds using 62 topological indices and exhibiting (218) or not (216) antibacterial activity have been considered [2]. Also, a more challenging and realistic dataset with 973 compounds where 111 of them exhibit analgesic properties have been used. In this second database, all 116 descriptors have been used to represent the compounds [15]. More details about data and availability are given in previous referenced works. Moreover, the experimental protocol including coding of all algorithms closely follows these previous studies.

As the main goal consists of an empirical comparison, a relatively wide range of settings has been tried for all the algorithms considered. To obtain appropriately averaged performance measures the $n$-fold cross validation procedure with $n = 10$ has been repeated four times. As a performance measure for each fold, the full ROC curve has been computed along with its AUC measure. Both ROC curves and AUC measures have been averaged over the different blocks in the cross validation procedure [17] and are shown in Figure 1 and Table 2, respectively. Only the results corresponding to the best settings for each algorithm have been presented. These settings for each particular algorithm and database are specified in Table 1. For all algorithms, a Gaussian kernel has been used whose parameter has been fixed as $\sigma = 0.125$ according to several previous studies using the same databases [2].

By observing the ROC curves obtained with the best settings for the Antibacterial database in Figure 1 it can be seen that there is a very significant difference between the two best algorithms (SVM and SSLM) and the rest. It is not surprising that the SVDD algorithm gives the worst results because it does not use negative examples. On the contrary, the poor result corresponding to the NWSVDD method was relatively unexpected. When considering the Analgesic database the performance of all algorithms gets significantly lower in all cases. This is due both to the fact that the problem is considerably more difficult and also because the database is severely unbalanced. For this database, SVDD and NWSVDD methods give virtually the same results along the ROC curve and SVM gives only slightly better results. The SSLM method gives the best results except for a small range in the curve. The AUC values shown in Table 2 numerically characterize the differences of performance among the different methods over the two databases. The AUC value corresponding to the LDA method has also been included as a baseline.

The particular AUC values obtained for each database in each of the 4 times 10 cross validation steps have been put together and a nonparametric Friedman test followed by a post-hoc Holm test [18] has been performed. Table 3 shows the obtained average rankings and adjusted $p$-values when comparing each method to SSLM. According to this, it can be said that the SSLM gives the best AUC results at a significance level of $\alpha = 0.05$.

These results illustrate the fact that OC predictors with enough information (counterexamples) and flexibility (in particular using a margin to separate examples from counterexamples) are able to improve on good binary classifiers (SVM). Nevertheless, the amount of improvement attained is relatively moderate. Apart from this improvement on the overall performance, the one-class predictors are interesting also because of its ability to adapt to different situations. In particular, in specific applications as the ones considered in this paper, it is possible to adapt the predictors to specific operating ranges of the ROC curve that correspond to specific situations. In other words, instead of looking for a unique model that gives rise to a good ROC curve, we can learn a specific model that is good only in a small range in the curve. This capability of the

**Table 1.** Best parameters for each one of the algorithms on each database

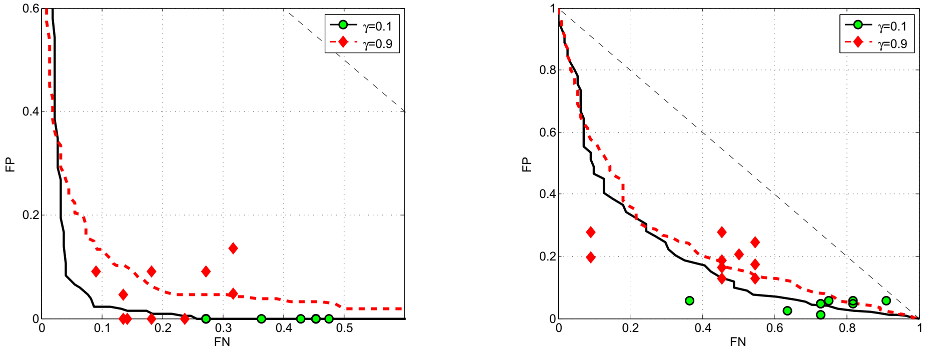
| | SVM | SVDD | NWSVDD | SSLM |
|---|---|---|---|---|
| Antibacterial | $\nu = 0.0319$ | $\nu = 0.25$ | $\nu = 0.0125$, $\gamma = 0.25$ | $\nu = 0.0001$, $\gamma = 0.1$, $\eta = 50$ |
| Analgesic | $\nu = 0.1194$ | $\nu = 0.9$ | $\nu = 0.0040$, $\gamma = 0.35$ | $\nu = 0.001$, $\gamma = 0.9$, $\eta = 40$ |

**Table 2.** AUC measure for each algorithm on each database

| | LDA | SVM | SVDD | NWSVDD | SSLM |
|---|---|---|---|---|---|
| Antibacterial | 0.966 | 0.976 | 0.686 | 0.871 | **0.985** |
| Analgesic | 0.834 | 0.829 | 0.732 | 0.788 | **0.852** |

**Table 3.** Average rankings and adjusted $p$-values

| Algorithm | SSLM | SVM | LDA | NWSVDD | SVDD |
|---|---|---|---|---|---|
| Ranking | 1.669 | 2.256 | 2.513 | 3.825 | 4.737 |
| Adjusted $p$-value | | 0.03755 | 0.00221 | $< 10^{-16}$ | $< 10^{-32}$ |



a) Antibacterial database

b) Analgesic database

**Fig. 2.** Best one-class models (SSLM) obtained for each database by forcing the algorithm to minimize either FP or FN rates by forcing the parameter $\gamma$

models has not been fully exploited in this work but Figure 2 shows two specific models specialized at the different endings of the ROC curve. In these figures particular predictors obtained at each one of ten runs are shown along with the corresponding averaged curves. In the case of Antibacterial database, it is possible to obtain predictors able to minimize one of the two types of errors but at different rates. In the Analgesic database a similar behavior can be observed. In both cases, the variability in the false negative rate is higher than the one in false positive rate.

## 5 Concluding Remarks

In this work, several different one-class predictors have been applied to a particular challenging problem related to drug activity characterization. In particular, recently proposed one-class predictors using counterexamples and a separation margin have been shown to give very interesting solution for this kind of problems. The behavior of the different models has been characterized by their corresponding ROC curves and AUC measures. Apart from the overall performance results it has been shown that the models can be adapted to different specifications in terms of maximum rates of each type of error. Further work is currently directed towards the specific problem of obtaining one or several one-class predictors optimized at different specific error rates.

## References

1. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining 2(2), 121–167 (1998)
2. Ferri, F.J., Diaz-Villanueva, W., Castro, M.: Experiments on automatic drug activity characterization using support vector classification. In: IASTED Intl. Conf. on Computational Intelligence (CI 2006), San Francisco, US, pp. 332–337 (2006)
3. Yepes, V., Pellicer, E., Ferri, F.: Profit forecasting using support vector regression for consulting engineering firms. In: 9th International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy (2009)
4. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
5. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471 (2001)
6. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
7. Wu, M., Ye, J.: A small sphere and large margin approach for novelty detection using training data with outliers. IEEE Trans. Pattern Anal. Mach. Intell. 31(11), 2088–2092 (2009)
8. Cao, L.J., Lee, H.P., Chong, W.K.: Modified support vector novelty detector using training data with outliers. Pattern Recogn. Lett. 24(14), 2479–2487 (2003)
9. Gozalbes, R., Doucet, J., Derouin, F.: Application of topological descriptors in qsar and drug design: History and new trends. Current Drug Targets – Infectious Disorders 2, 93–102 (2002)
10. Katritzky, A.R., Gordeeva, E.V.: Traditional topological indices vs electronic, geometrical, and combined molecular descriptors in qsar/sqpr research. J. Chem. Inf. Comput. Sci. 33, 835–857 (1993)
11. Basak, S., Bertelsen, S., Grunwald, G.: Application of graph theoretical parameters in quantifying molecular similarity and structure-activty studies. J. Chem. Inf. Comput. Sci. 34, 270–276 (1994)
12. Seybold, P., May, M., Bagal, U.: Molecular structure-propertiy relationships. J. Chem. Educ. 64, 575–581 (1987)
13. Kier, L.B., Hall, L.H.: Molecular Connectivity in Structure-Activity Analysis. John Willey and Sons, New York (1986)
14. Balaban, A.T.: Highly discriminating distance-based topological index. Chem. Phys. Lett. 89, 399–404 (1982)
15. Gálvez, J., García-Domenech, R., de Julián-Ortiz, J., Soler, R.: Topological approach to drug design. J. Chem. Inf. Comput. Sci. 35, 272–284 (1995)
16. Galvez, J., Garcia, R., Salabert, M., Soler, R.: Charge indexes. new topological descriptor. J. Chem. Inf. and Comp. Sciences 34, 502–525 (1994)
17. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)
18. García, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9, 2677–2694 (2008)

# Segmentation Based Urdu Nastalique OCR

Sobia Tariq Javed[1] and Sarmad Hussain[2]

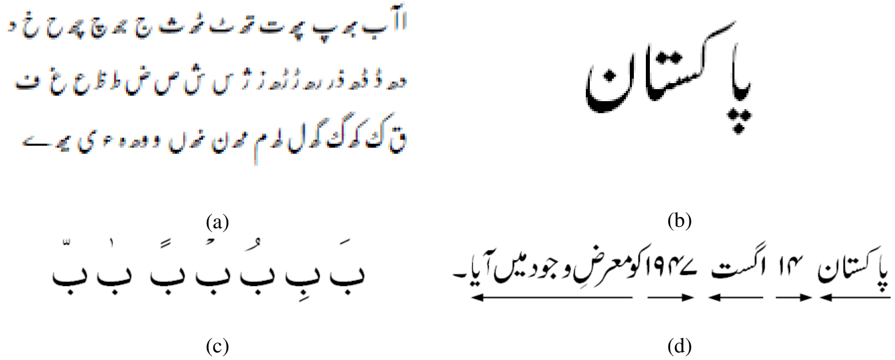
[1] National University of Computer and Emerging Sciences, Lahore, Pakistan
sobia.tariq@nu.edu.pk
[2] Al-Khawarizmi Institute of Computer Science,
University of Engineering and Technology, Lahore, Pakistan
sarmad.hussain@kics.edu.pk

**Abstract.** Urdu Language is written in Nastalique writing style, which is highly cursive, context sensitive and is difficult to process as only the last character in its ligature resides on the baseline. This paper focuses on the development of OCR using Hidden Markov Model and rule based post-processor. The recognizer gets the main body (without diacritics) as input and recognizes the corresponding ligature. Accuracy of the system is 92.73% for printed and then scanned document images at 36 font size.

**Keywords:** Nastalique, Urdu OCR, Urdu Segmentation.

## 1 Introduction

Urdu is written using Arabic script in Nastalique writing style. Urdu has an extended Arabic character set as given in the figure below [13, 14, and 15]. Urdu characters are constituted by a main body with zero or more diacritics for specifying the consonants and (optionally) vowels. Nastalique writing style is very cursive with context sensitive shaping [9, 10]. The characters join together to form a *Ligature* . One or more ligatures form a word. For example word Pakistan shown in the Figure 1 (b) has three ligatures. Moreover, Urdu is bidirectional [3].



(a) (b)

(c) (d)

**Fig. 1.** Different characteristics of Urdu Writing System (a) Urdu character Set [3] (b) Word *Pakistan* written in Natalique (c) Diacritical marks on letter bay (d) Bidirectional Urdu script

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 41–49, 2013.

The character set of Urdu shown in Figure 1 (a) can be sub-categorized into classes which contain same base forms (if the dots and marks are ignored). This categorization is given in Table 1, and is the basis of segmentation and recognition. Some letters belong to multiple classes because they contain isolated and final forms different from initial and medial forms. For example, letter *Fay* has same initial and medial forms as Qaf (e.g. فب بفب compared with قب بقب) but different isolated and final forms (e.g. بف ف compared with بق ق). These letters are listed in the last row, in addition to being listed with other classes. In the rest of the paper we refer to classes and not the individual characters.

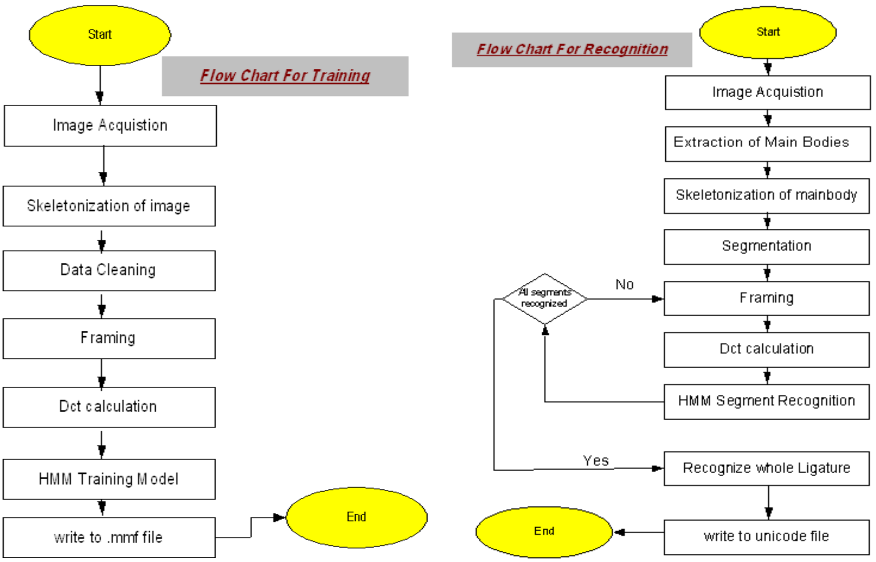
**Table 1.** Classification of Urdu letters based on their shapes

| Member Letter(s) | Class | | Member Letter(s) | Class | | Member Letter(s) | Class |
|---|---|---|---|---|---|---|---|
| م | م | | ص | ص | | ا آ | ا |
| و | و | | ط ظ | ط | | ب پ ت ٹ ث ن | ب |
| ہ | ہ | | ع غ | ع | | ج چ ح خ | ج |
| ھ | ھ | | ف ق | ف | | د ڈ ذ | د |
| ی | ی | | ک گ | ک | | ر ڑ ز ژ | ر |
| ے | ے | | ل | ل | | س ش | س |
| | | | | | | | |
| ق | ق | | ف | ف | | ن ں | ن |

## 2 Methodology

The current system uses HMMs for pattern matching, as it can accurately handle large data sets and can be trained to handle noise and distortion to some extent [5, 6, 12, 16, 17, 18]. The recognition process is shown in the Figure 2.

The system takes a monochrome scanned image with 150 dpi containing Urdu text as input. As the document images are generated by the authors in the lab to test the process, there is no pre-processing and it is assumed that the document images are not skewed and with minimum noise and distortion. Main bodies are extracted by first separating lines of text within the page, then identifying the baseline and finally separating the main bodies from diacritics using the baseline [2, 3]. After extracting the main body, they are skeletonized using Jang Chin algorithm [4], as shown in Figure 3.



(a) (b)

**Fig. 2.** Flow Charts for (a) Training and (b) Recognition

اللہ ہم کو بخش دے۔تو رحمن ہے۔تو رحیم ہے

(a) (b)

**Fig. 3.** (a) Original Image and (b) Sketetonized Image

The skeletonized image is then segmented after determining the ending point of the ligature. In Nastalique it is very difficult to determine the exact starting point of the ligature so instead of that we start with the ending point of the ligature [1] which is more deterministic, as shown in Figure 4.

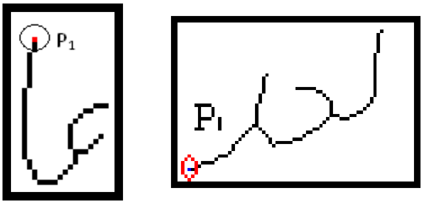


**Fig. 4.** The Ending Points P1 of the Ligatures are Circled

Therefore, as Urdu is written from right to left, the ligatures are traversed from left to right. During the traversal the ligatures are segmented at branching points as shown in Figure 5 below. This process results in multiple segments from a ligature, obtained in the reverse writing order as shown in Table 2.

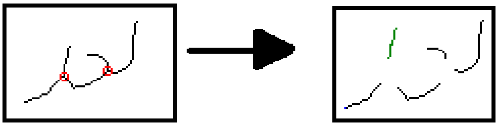
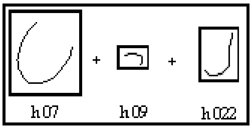
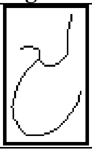
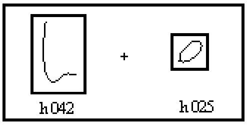
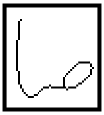
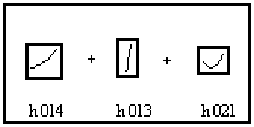
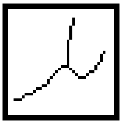
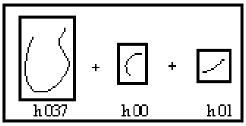
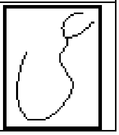


**Fig. 5.** Segmentation of a ligature using branching points [1]

**Table 2.** Segmentation of the Ligatures

| Sr. | Segments | Ligature | Sr. | Segments | Ligature |
|---|---|---|---|---|---|
| 1 |  |  | 3 |  |  |
| 2 |  |  | 4 |  |  |

These skeletonized ligature segments are framed and used to train the HMMs, as shown in Figure 6. The system is tested with non-overlapping frame sizes of 5x5, 8x8, 9x9, 12x12 and 16x16 pixels. 8x8 is found to give the best results for the 36 font size.

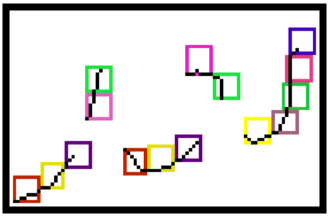
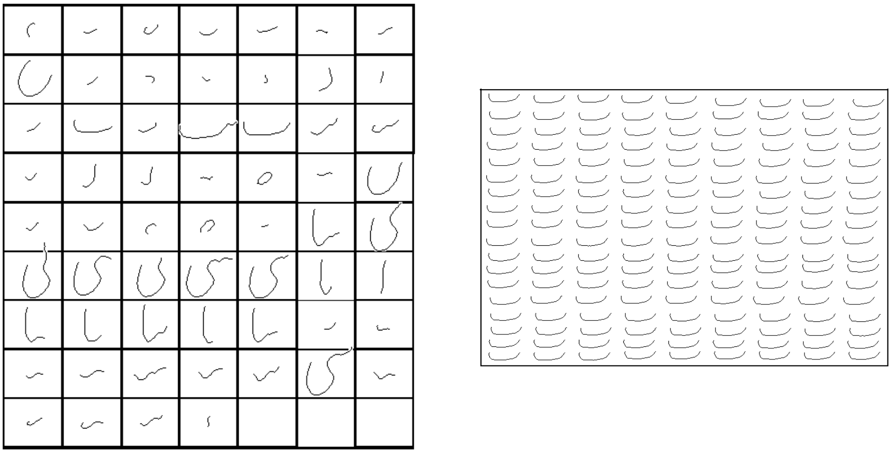


**Fig. 6.** The Framing of Segmented Word بعد*(baad)*

When this pre-processing is complete, before starting training process, the HMM parameters are initialized with training data in order to allow convergence of the training algorithm. Each segment is considered as a separate HMM. Sixty HMMs were extracted from all shapes (isolated, initial, final and medial) of a sub-set of six classes of Urdu characters, including *Alif, Bay, Dal, Swad, Ain* and *Yeh* classes which are given in the Figure 7 (a). In order to cater variations in the image 100 samples of each shape are collected for training the HMMs. Samples of isolated *Bay* are given below in Figure 7 (b).

After training the model, the recognition process is performed. In recognition process the skeletonized ligature is first segmented and then each segment is fed to the HMM for recognition.



(a) (b)

**Fig. 7.** (a) Segments modeled by HMMs and (b) Variation in training samples used for the HMMs

As the segments are of varied sizes, and the framing window size is fixed, for more precisely modeling each segment different number of states are defined for the different segments, with some examples illustrated in Table 3.

**Table 3.** HMM State Analysis

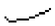
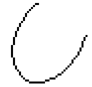
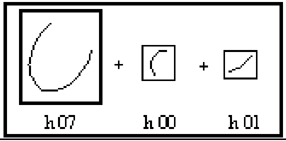
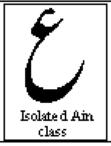
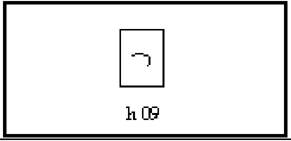
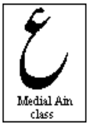
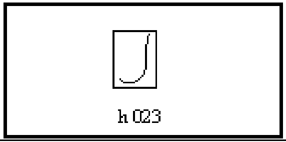
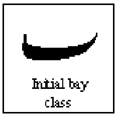
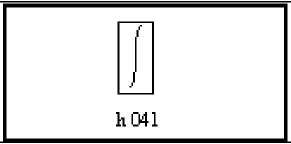
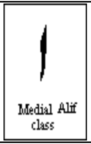
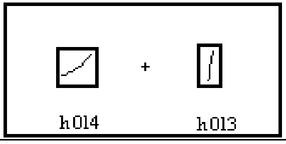
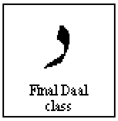
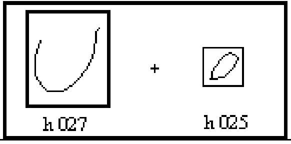
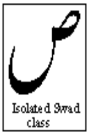
| Sr. | HMM Name | HMM | No. of Frames | No. of states | No. of samples |
|---|---|---|---|---|---|
| 1 | h00 | | 3 | 5 | 100 |
| 2 | h01 | | 2 | 4 | 100 |
| 3 | h02 | | 3 | 5 | 109 |
| 4 | h03 | | 3 | 5 | 136 |
| 5 | h04 |  | 3 | 5 | 100 |
| 6 | h05 | | 3 | 4 | 110 |
| 7 | h06 | | 3 | 4 | 122 |

**Table 3.** (*Continued*)

| | | | | | |
|---|---|---|---|---|---|
| 8 | h07 |  | 10 | 11 | 108 |
| 9 | h08 | | 2 | 4 | 137 |
| 10 | h09 | | 2 | 4 | 114 |

After training, the model is developed, the recognition process is performed. In the recognition process the skeletonized ligature is first segmented and then each segment is sent to the HMM for recognition. Once the constituent segments are recognized, rules are applied to order them to form the corresponding ligature, as shown in Table 4.

**Table 4.** Rules for Forming Ligatures from Constituent Segments

| Sr. | Segments | Letter | Sr. | Segments | Letter |
|---|---|---|---|---|---|
| 1 |  |  | 4 |  |  |
| 2 |  |  | 5 |  |  |
| 3 |  |  | 6 |  |  |

## 3 Results

A total of 1692 ligatures, which are formed from the six base forms mentioned above, are extracted from the 18600 high frequency words in a corpus-based dictionary [7]. These classes are used in these ligatures in a variety of contexts. The Urdu words were written in font Noori Nastalique and font size 36. The pages are printed and then scanned at 150 DPI. Out of these 1692 ligatures, 1569 ligatures were identified correctly giving an accuracy of 92.73%.

## 4 Discussion

The focus of the paper has been to explore segmentation based system capable of recognizing Urdu Nastalique font. The results are promising; however, some letters are not recognized correctly due to the following problems. The variation in the images affects the output of the recognizer. The variation may be introduced due to scanning, binarization and thinning processes, giving wrong recognition results, as shown in Figure 8.

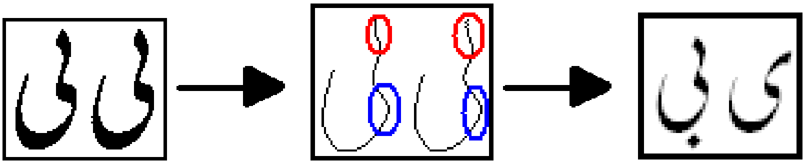
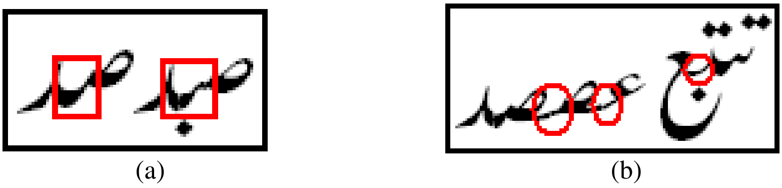


**Fig. 8.** Variation in the images causing ligature misrecognition

This problem may be resolved by increasing the number of training samples and by giving original segment as HMM input instead of giving skeletonized segment.

The similarity in the shapes of different characters can also lead to the recognizer confusion. For example, the shape of the letter *Bay* and last stroke in *Swad* (in Figure 9) are similar to each other when written in Noori Nastalique. Diacritics can disambiguate such cases.



(a)

(b)

**Fig. 9.** Similarity in Shapes *Swad* and *Bay* in Different Contexts

Inconsistency in font can also cause some variation causing recognition errors. The Noori Nastalique font used shows such behavior in some cases, e.g. main bodies of ligatures change with change in diacritics as shown in Figure 10. This variation is because the font uses hand written ligatures.

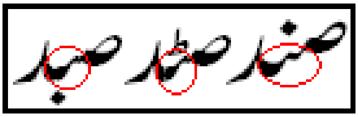


**Fig. 10.** Dissimilarity in Shapes of Same Ligature with Different Diacritic Placement

## 5 Conclusion

The Nastalique style used to write Urdu language is complex due to its diagonal, context sensitive and cursive nature. In this paper we have presented a technique to develop a segmentation based OCR for Nastalique. A ligature is first segmented and each segment is recognized using an HMM based recognizer. Then a set of rules are used to identify the ligature corresponding to the sequence of recognized segments. The accuracy of system is 92.73% for six base forms using fabricated documented images at 36 font size. The technique still needs to be tested on real data and extended to cover the entire set of Urdu letters at a variety of font sizes.

## References

1. Javed, S.T., Hussain, S.: Investigation into a Segmentation Based OCR for the Nastalique Writing System. Master's thesis report at National University of Computer and Emerging Sciences, Lahore (2007), `http://www.cle.org.pk/resources/theses.htm`
2. Javed, S.T., Hussain, S., Maqbool, A., Asloob, S., Jamil, S., Moin, H.: Segmentation Free Nastalique Urdu OCR. Journal of World Academy of Science, Engineering and Technology (70) (2010), `http://www.waset.org/journals/waset/v70.php`
3. Javed, S.T., Hussain, S.: Improving Nastalique Specific Pre-Recognition Process for Urdu OCR. In: The Proceedings of 13th IEEE International Multitopic Conference 2009 (INMIC 2009), Islamabad, Pakistan (2009)
4. Jang, B.-K., Chin, R.T.: Analysis of thinning algorithms using mathematical morphology. IEEE Transactions on Pattern Analysis and Machine Intelligence (1990)
5. Rabiner, L., Juang, B.-H.: Theory and Implementation of Hidden Markov Models. In: Fundamental of Speech Recognition, ch. 6 (1993)
6. Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (1995)
7. Ijaz, M., Hussain, S.: Corpus Based Urdu Lexicon Development. In: The Proceedings of Conference on Language Technology (CLT 2007), University of Peshawar, Pakistan (2007)
8. Pal, U., Sarkar, A.: Recognition of Printed Urdu Text. In: The Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR (2003)
9. Hussain, S.: www.LICT4D.asia/Fonts/Nafees_Nastalique. In: The Proceedings of 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society. Asian Media Information Center, Singapore (2003)
10. Wali, A., Hussain, S.: Context Sensitive Shape-Substitution in Nastaliq Writing system: Analysis and Formulation. In: The Proceedings of International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering, CISSE (2006)
11. Lu, Z., Bazzi, I., Kornai, A., Makhoul, J.: A Robust, Language-Independent OCR System. In: The 27th AIPR Workshop: Advances in Computer Assisted Recognition, SPIE (1999)
12. Bojovic, M., Savic, M.D.: Training of Hidden Markov Models for Cursive Handwritten Word Recognition. In: The Proceedings of the15th International Conference on Pattern Recognition (ICPR), vol. 1 (2000)
13. Hussain, S., Afzal, M.: Urdu Computing Standards: UZT 1.01. In: The Proceedings of the IEEE International Multi-Topic Conference, Lahore, Pakistan (2001)
14. Hussain, S.: Letter to Sound Rules for Urdu Text to Speech System. In: The Proceedings of Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, Switzerland (2004)
15. Hussain, S., Durrani, N.: Urdu. In: A Study on Collation of Languages from Developing Asia, Center for Research in Urdu Language Processing, NUCES, Pakistan (2007)
16. El-Hajj, R., Likforman-Sulem, L., Mokbel, C.: Arabic Handwriting Recognition Using Baseline Dependant Features and Hidden Markov Modeling. In: The 8th International Conference on Document Analysis and Recognition (ICDAR), South Korea (2005)

17. Elms, A.J.: A Connected Character Recognizer Using Level Building of HMMs. In: The Proceedings of 12th International Conference on Pattern Recognition (1994)
18. Safabakhsh, R., Abidi, P.: Nastaaligh Handwritten Word Recognition Using a Continuous-Density Variable-Duration HMM. The Arabian Journal for Science and Engineering (2005)
19. Shah, Z., Saleem, F.: Ligature Based Optical Character Recognition of Urdu, Nastaliq Font. In: The Proceedings of International Multi Topic Conference, Karachi, Pakistan (2002)
20. Husain, S.A., Amin, S.H.: A Multi-tier Holistic approach for Urdu Nastaliq Recognition. In: The Proceedings of International Multi Topic Conference, Karachi, Pakistan (2002)
21. Ahmad, Z., Orakzai, J.K., Shamsher, I., Adnan, A.: Urdu Nastalique Optical Character Recognition. In: The Proceedings of World Academy of Science, Engineering and Technology (2007)
22. Shamsher, I., Ahmad, Z., Orakzai, J.K., Adnan, A.: OCR for Printed Urdu Script Using Feed Forward Neural Network. In: The Proceedings of World Academy of Science, Engineering and Technology (2007)
23. Malik, S., Khan, S.A.: Urdu online handwriting recognition. In: Proceedings of the IEEE Symposium on Emerging Technologies (2005)

# Misalignment Identification in Induction Motors Using Orbital Pattern Analysis

José Juan Carbajal-Hernández[1,*], Luis Pastor Sánchez-Fernández[1],
Victor Manuel Landassuri-Moreno[2], and José de Jesús Medel-Juárez[1]

[1] Center of Computer Research – National Polytechnic Institute. Av. Juan de Dios Bátiz s/n,
Nueva. Industrial Vallejo, Gustavo A. Madero, México D.F., C.P. 07738, México
[2] Mexico Valley University Center (CUUAEM-VM) – Autonomous University of the State
of Mexico. Boulevard Universitario, Predio San Javier, Atizapán de Zaragoza,
Estado de México, C.P. 54500, México
{jcarbajalh,lsanchez,jjmedelj}@cic.ipn.mx,
vmlandassurim@uaemex.mx

**Abstract.** Induction motors are the most common engine used worldwide. When they are summited to extensive working journals, e.g. in industry, faults may appear, generating a performance reduction on them. Several works have been focused on detecting early mechanical and electrical faults before damage appears in the motor. However, the main drawback of them is the complexity on the motor's signal mathematical processing. In this paper, a new methodology is proposed for detecting misalignment faults in induction motors. Through signal vibration and orbital analysis, misalignment faults are studied, generating characteristically patterns that are used for fault identification. Artificial Neural Networks are evolved with an evolutionary algorithm for misalignment pattern recognition, using two databases (training and recovering respectively). The results obtained, indicate a good performance of Artificial Neural Networks with low confusion rates, using experimental patterns obtained from real situations where motors present a certain level of misalignment.

**Keywords:** Orbital analysis, patterns recognition, neural networks evolution, motor fault, misalignment.

## 1 Introduction

Motor fault analysis is a common industrial practice where machinery is summited to extensive working journals. Induction motors are based on different electrical and mechanical components that can suffer some kind of wearing down with prolonged use [1]. This work has been oriented to electrical induction motors, since they are most used in industry worldwide. Historically, some works have been focused to detect some faults (most of them commonly identified) avoiding future problems and damages if correctly maintenance is early provided [1-3]. Induction motors can be mainly classified by size, power, number of phases, etc. However, misalignment in

[*] Corresponding author.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 50–58, 2013.

rotor bars is a common fault in all kinds of motors that generates different levels of external vibrations [4].

Several techniques can be used in fault motor analysis such as support vector machine [5], Fourier spectrum [6], wavelet filtering [7], among others as [8-11]. However, the complexity of the mathematical processing in the motor signal is a drawback on them, where the implementation in real systems may be a difficult task. Thus, a new technic using orbital analysis is proposed, offering a practical and easy way to recognize misalignment faults in induction motors. Orbital analysis in motor faults has been used for modeling normal operation [12, 13], and they represent a basis of this research. Different motor misalignments levels present a particular characteristic vibration orbit, which can be used to determine when a motor presents a fault, before a serious damage appears in the machine. The main contribution of this work is the development of a new computational model for induction motors fault recognition, using artificial intelligence technics as Artificial Neural Networks (ANNs). They are evolved with an evolutionary algorithm called FS-EPNet to optimize the networks architectures for orbital analysis.

The rest of the paper is organized as follows: firstly, section 2 presents how electrical signals are measured and preprocessed using sensors, and how an orbit is created. Section 3 explains the representation of different misalignment faults in orbital patterns and the main characteristics over normal and bad orbits. In section 4, ANNs and the FS-EPNet algorithm are shown for the recognition of orbital patterns. Thus, section 5 shows experimental results using two databases from real measured patterns: learning and recalling phases respectively. Finally, section 6 presents the conclusions reached about the advantages and disadvantages of the proposed model.

## 2 Signal Acquisition and Preprocessing

This section is aimed to present the procedure of obtaining a characteristic orbit, i.e. sampling, positioning and signal preprocessing to extract and separate it from a measured signal.

*Sampling*

Vibration is a common symptom derived from mechanical faults in induction motors. Such vibrations can be measured using a piezoelectric accelerometer sensor, which generates an electrical signal that is proportional to the acceleration vibration of a seismic mass [14]. As each motor have a different rotation speed, standards as [15] and [16] have established sampling frequency rates for motor measuring. According to them, this work used a sampling frequency of 50 kHz, being large enough to obtain a good quality signal, over tested induction motors.

*Positioning*

Orbital patterns are built using two signals that are plotted together. In order to obtain those signals, two piezoelectric accelerometers are placed orthogonally and radial to the motor chassis bearing (Fig. 1).

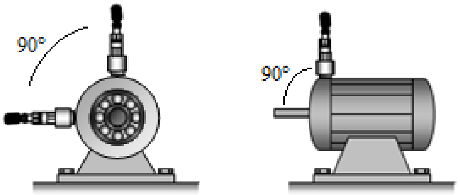


**Fig. 1.** Accelerometer placement at 90° over the engine

*Signal Preprocessing*
Accelerometer signals are measured in acceleration units and must be converted to displacement units, using a double integration process as follows [14]:

$$v(t) = \int_0^t a(t)dt + v_0 \tag{1}$$

$$d(t) = \int_0^t v(t)dt + d_0 \tag{2}$$

where $a(t)$ is acceleration, $v(t)$ is velocity, $d(t)$ is displacement, $v_0$ and $d_0$ are the initial velocity and displacement conditions respectively.

Vibration signals in displacement units are compounded by several harmonics; each of them can be related with the normal operation of the engine or with a motor fault. Undesirable harmonics can distort the shape of the orbit, changing notably the main characteristics of a fault shape. In this sense, those harmonics must be avoided in order to have a good quality orbit. A Butterworth passband filter was implemented for removing those spurious harmonics according to the following magnitude response [17]:

$$|H(\omega)|^2 = \frac{1}{1 + \left(\frac{c - \cos\omega}{\Omega_0 \sin\omega}\right)^{2N}} \tag{3}$$

where $\omega = 2\pi f/f_s$, $f_s$ is the sampling frequency, $\Omega_0 = \tan(\omega_0/2)$ and $c$ can be expressed as follows:

$$c = \frac{\sin(\omega_{pa} + \omega_{pb})}{\sin\omega_{pa} + \sin\omega_{pb}} \tag{4}$$

where $\omega_{pa} = 2\pi f_{pa}/f_s$, $\omega_{pb} = 2\pi f_{pb}/f_s$ and $[f_{pa}, f_{pb}]$ is the passband.

An unfiltered orbit has an irregular form, making no possible fault discovery; however, a remarkable shape may be clearly seen in a filtered orbit (Fig. 2).

Each filtered signal generates continuous orbits with the same shape (Fig. 2); nevertheless, just one orbit is required in this work for analyzing its characteristics

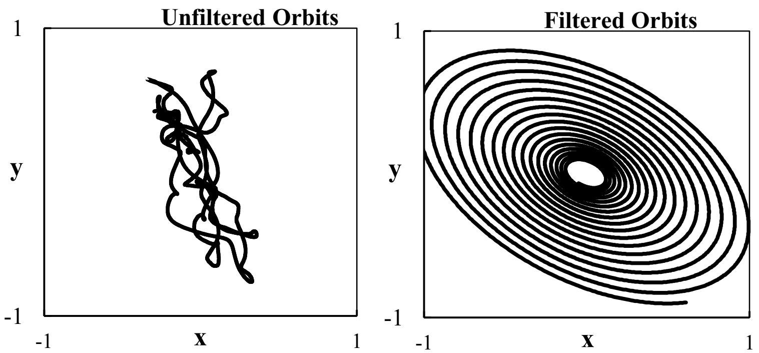


**Fig. 2.** Example of the orbital filtering using the Butterworth passband filter, which removes spurious harmonics allowing clear orbits

and detecting whether a misalignment is present. In this case, we search two points into a filtered signal (starting and ending), with a low distance defined by a tolerance. This tolerance was obtained computing the average of the distance of all points into the signal. There is not a rule for establishing the tolerance; however, this value was enough for obtaining good shape orbits. The distance criteria were obtained using the Euclidian equation as follows [17]:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

Where *d* is the distance between points, and (*x*, *y*) are the orbit points coordinates respectively. Finally, extracted orbits should be normalized due to differences in their size. Therefore, all orbits are resized in a [-1, 1] range according with the following equation:

$$s_{1,2}(n) = \frac{d_{1,2}(n)}{max\{|d_1(n)|, |d_2(n)|\}}, \quad \forall\, n = 0, 1, 2, \ldots, N - 1 \tag{6}$$

## 3 Orbital Analysis

There is a correspondence between orbit shapes and motor faults, i.e. when an induction motor is in good condition (no faults are present), the corresponding orbit is a circumference; on the other hand, when a misalignment fault is present in the motor, the orbit shape suffers a deformation in one part of the circumference. A misalignment may appear in different intensities: a slight misalignment almost deforms the circumference; by contrast, a strong misalignment deforms considerably the orbit shape, generating a kind of “8” in the circumference. Fig. 3 shows the motor’s orbit shapes of different misalignment intensities.

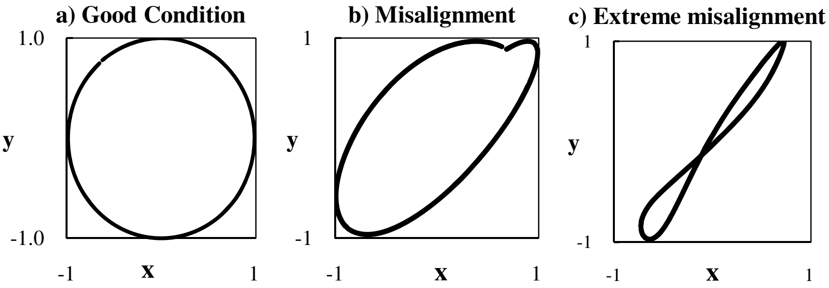


**Fig. 3.** Examples of orbit shapes: a) good motor conditions have perfect circumferences, b) a slight misalignment fault is present in the motor and c) an extreme misalignment have a shape of an "8" number.

# 4 Pattern Recognition

There are several technics for pattern recognition that can be used for orbits shape identification [1–11]. In this work, Artificial Neural Networks (ANN) are used as classifiers, because they have proved being a very effective learning model with high rates of effectiveness. However, the construction of an ANN is not an easy task; for this reason, evolutionary algorithms are used for building ANN architectures, establishing criteria for a better selection of the ANN's parameters. In this sense, the final ANN topology chosen by the evolutionary algorithm guarantees the best performance of the ANN. Also, connectivity reduction tests help to avoid computational burden with a high efficiency of the ANN. This process is known as Evolutionary Artificial Neural Networks (EANNs) or Neuroevolution.

*Artificial Neural Network*

Evolution of Artificial Neural Networks have been remarkably useful at optimizing networks' parameters during evolution [18-21], also local minima may be avoided than using traditional gradient-based search algorithms [18].

The Feature Selection EPNet algorithm (FS-EPNet) [19, 21] allows the ANNs' parameter evolution, including the input adaptation of the networks (Feature Selection Evolution). The FS-EPNet is a steady-state algorithm based in Lamarkian inheritance, where information learned by parents is passed to children; also, no crossover operator is used to avoid the permutation problems [18]. In this way, nine different mutations are used to carry out the evolution of individuals (ANNs): (1) hybrid training, composed of training with the Modified Back Propagation (MBP) algorithm and Simulated Annealing (SA); (2) node deletion; (3) connection deletion; (4) input deletion; (5) delay deletion; (6) connection addition; (7) node addition; (8) input addition; and (9) delay addition. Only one such mutation is performed on the selected individual in each generation. The hierarchical order of the mutations permit to maintain networks sizes to the minimum; however, if the problem cannot be solved more accurately, it will start to add nodes and connections, increasing the average networks sizes over the population. A detailed description of FS-EPNet algorithm may be seen in [20, 21].

*Pattern Building*

According with the orbital signal analysis, orbit shapes where used for creating motor fault patterns. However, resulting signal orbits are not practical to be used in a neural network due to they have different lengths. In order to have uniform patterns, all orbits signals were resampled for having 100 points of length, where each one is a bidimensional pattern ($x$, $y$). A database of 386 patterns was created to be used in the learning phase of the ANN (from here, an inside test set is obtained to evolve the networks with the FS-EPNet algorithm). Orbit shapes of this database were measured from different induction motors, which had different misalignment levels: 275 regular misalignment patterns, 106 extreme misalignments patterns and 5 patterns of good condition motors.

## 5 Experimental Results

An experimental database was used for validating the performance of the proposed system as part of a recovering process. This database was built using different kind of motors and with different levels of misalignments. It is important to remark that this database was compounded by different motor measurements than those used in the database of the learning process. In this case, the size of the database was of 118 motor fault patterns as follows: 73 regular misalignment patterns, 35 extreme misalignments patterns and 10 patterns of good condition motors. From those patterns, the final test set was obtained, applied after the evolutionary process has finished. Preliminary experiments allow setting up some common parameters in this study: population size 30, generations of evolution 100 (stopping criteria), initial connection density 100% and 30%. Initial learning rate 0.15, minimum learning rate 0.01, epochs for learning rate adaptation 5, number of mutated hidden nodes 1, number of mutated connections 1-3. The inputs are fixed at 200, where the first half is for $x$-axis and the other half for $y$-axis. The hidden nodes are initialized between 2 and 10 nodes randomly. Partial training settle at 100 epochs, whereas 1000 epochs of further training at the end of the algorithm. 30 independent runs were performed to ensure statistical validity of the results. It was used two test sets to evaluate the performance of the algorithm, one inside of the evolutionary algorithm (100 partners from the available data to train) and one final test set (experimental database) to evaluate the final performance of the algorithm.

Figure 4 presents the Average Classification error (Fig. 4a), the average error in terms of the Normalize Root Mean Squared Error (NRMSE, Fig. 4b), the Average connections (Fig. 4c) and the Average hidden nodes (Fig. 4d) for the orbit motor fault recognition process over 100 generations of evolution with 100% and 30% of connectivity at network initialization. It can be seen in Fig. 4a, that initializing the networks with 100% of connectivity allows a perfect classification error in the test set inside the FS-EPNet algorithm, using the *winner takes all* method from the first

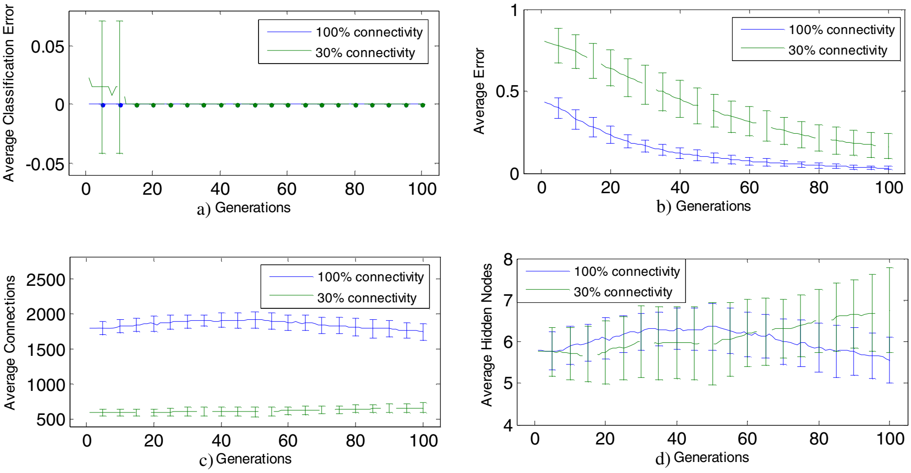


**Fig. 4.** Evolved parameters with the FS-EPNet over 100 generations of evolution for initial connectivity of 100% and 30% for the orbit motor fault recognition process: a) average classification error (*winner takes all*); b) average error in terms of the NRMSE; c) average connections and d) average hidden nodes.

generation. These results indicate that at the random initialization and initial partial training, the networks in the population can solve the problem without any effort; however, that is not maintained for the final test set (see final line in Table 1, for the final test set). On the other hand, a considerable reduction in the connectivity (30%) produces average errors over 0.01 during the first 10 generations (Fig. 4a). Thereafter, the networks can solve the problem with the same accuracy, and with fewer numbers of connections (Fig. 4c). A similar behaviour is presented using the NRMSE over the test set inside the evolutionary algorithm, where both errors started to converge as the generation advance. Finally, as it can be seen in Fig. 4d, networks that are initialized with a reduce number of connections, started to increase the number of hidden nodes, as more resources continue to solve the problems accurately (also connections are slightly increased, Fig. 4c). It is clear that in both cases, 100 generations of evolutions is enough to achieve perfect classification errors on the test set inside the algorithm.

**Table 1.** Orbit motor fault recognition results with 100% and 30% of connectivity with the FS-EPNet

| Parameter | 100% Connectivity | | | | 30% Connectivity | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Std Dev | Min | Max | Mean | Std Dev | Min | Max |
| Number of Inputs | 200 | 0 | 200 | 200 | 200 | 0 | 200 | 200 |
| Number of Hidden Nodes | 5.966 | 0.999 | 4 | 8 | 6.633 | 1.325 | 4 | 9 |
| Number of Connections | 1830.4 | 208.01 | 1419 | 2256 | 656.26 | 93.657 | 463 | 822 |
| Error Test Set EPNet | 0.034 | 0.01 | 0.018 | 0.061 | 0.168 | 0.084 | 0.023 | 0.333 |
| Classification Error inside | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Error Final Test Set** | **4.054** | **0.207** | **3.753** | **4.405** | **4.281** | **0.427** | **3.297** | **5.054** |
| **Classification Error Final Test Set** | **6.371** | **0.987** | **5.128** | **8.158** | **6.736** | **1.028** | **4.895** | **8.624** |

Table 1 presents the results of evolving ANNs with both values of connectivity. There is appreciated that classification errors tested during the evolution of the networks is perfect, as commented before (over the classification error inside the evolutionary algorithm); nevertheless, that is not maintained for the final test set, last line of Table 1.

## 6 Discussion and Conclusions

In this work, the use of orbital analysis and evolved Artificial Neural Networks (ANNs) for fault recognition in induction motor were proposed. Although several methodologies for detecting mechanical faults in induction motor have been developed, the proposed model represents a feasible and alternative way for motor misalignment fault detection. One disadvantage of this model is the number of preprocessing steps implemented before the ANN classification step. However, a misalignment was clearly shown to distort considerably an orbit, having a characteristic shape, which can be perfectly identified by classifiers as ANNs (designed with the FS-EPNet algorithm). On the other hand, the evolution of Artificial Neural Network provides a good topology optimization, avoiding computational burden for recovering phase, and giving an accurate assessment in the classification of misalignment orbits. It may be worth to say that this paper provides a preliminary study of misalignment identification in induction motors using orbital pattern analysis, and future works is needed, e.g. including additional patterns from different mechanical faults in order to expand the capacities of the system, or use lower connectivity values to initialize ANNs (before evolution starts) to generate smaller architectures. Anyhow, this model can be used as an important tool for preliminaries motor analysis, when the good functioning of the machine is essential in critical time production industry.

## References

1. Chow, M.: Methodologies of using neural network and fuzzy logic technologies for motor incipient fault detection. World Scientific, Singapore (1997)
2. Wang, J., Gao, R., Yan, R.: Broken-rotor-bar diagnosis for induction motors. Journal of Physics 305, 1–10 (2011)
3. Filippetti, F., Franceschini, G., Tassoni, C.: Neural networks aided on-line diagnostics of induction motor rotor faults. IEEE Transactions on Industry Applications 31(4), 892–899 (1995)
4. Lee, Y., Lee, C.: Modelling and analysis of misaligned rotor–ball bearing systems. Journal of Sound and Vibration 224, 17–32 (1999)
5. Matić, D., Kulić, F.: SVM broken bar detection based on analysis of phase current. In: 5th International Power Electronics and Motion Control Conference and Exposition, vol. (6397456), pp. LS4c.21–LS4c.24 (2012)
6. Climente, V., Antonino, J., Riera, M., Puche, R., Escobar, L.: Application of the Wigner–Ville distribution for the detection of rotor asymmetries and eccentricity through high-order harmonics. Electric Power Systems Research 91, 28–36 (2012)

7. Kechida, R., Menacer, A., Talhaoui, H.: Approach signal for rotor fault detection in induction motors. Journal of Failure Analysis and Prevention 13(3), 346–352 (2013)
8. Füssel, D., Ballé, P.: Combining neuro-fuzzy and machine learning for fault diagnosis of a D.C. motor. In: Proceedings of the American Control Conference, pp. 37–41 (1997)
9. Gaylard, A., Meyer, A., Landy, C.: Acoustic evaluation of faults in electrical machines. In: Electrical Machines and Drives, Conference Publication; 412, pp. 147–150 (1995)
10. Liu, D., Zhao, Y., Yang, B., Sun, J.: A new motor fault detection method using multiple window S-method time-frequency analysis. In: International Conference on Systems and Informatics, pp. 2563–2566 (2012)
11. Kim, K., Parlos, A.: Induction motor fault diagnosis based on neuropredictors and wavelet signal processing. IEEE/ASME Transactions on Mechatronics 7(2), 201–219 (2002)
12. Ha, K., Hong, J., Kim, G., Chang, K., Lee, J.: Orbital analysis of rotor due to electromagnetic force for switched reluctance motor. IEEE Transactions on Magnetics 36(4), 1407–1411 (2000)
13. Dongfeng, S., Lianfsheng, O., Ming, B.: Instantaneous purified orbit: A new tool for analysis of nonstationary vibration of rotor system. International Journal of Rotating Machinery 7(2), 105–115 (2001)
14. Han, S.: Retrieving the time history of displacement from measured acceleration signal. Journal of Mechanical Science and Technology 17(2), 197–206 (2003)
15. ISO 10816. Mechanical vibration: evaluation of machine vibration by measurements on non-rotating parts (1995)
16. VDI 2056. Standards of evaluation for mechanical vibrations of machines, Germany (1964)
17. Proakis, J., Manolakis, D.: Tratamiento digital de señales. Pearson Education, vol. 1(4a). Ed. España (2007)
18. Yao, X., Liu, Y.: A new evolutionary system for evolving artificial neural networks. IEEE Transactions on Neural Networks 8(3), 694–713 (1997)
19. Yao, X.: Evolving artificial neural networks. Proceedings of the IEEE 87(9), 1423–1447 (1999)
20. Bullinaria, J.: Evolving neural networks: Is it really worth the effort? In: Proceedings of the European Symposium on Artificial Neural Networks, pp. 267–272. d-side, Evere (2005)

# Bus Detection for Intelligent Transport Systems Using Computer Vision

Mijail Gerschuni and Alvaro Pardo

Department of Electrical Engineering, School of Engineering and Technologies,
Universidad Catolica del Uruguay
mgerschu@gmail.com, apardo@ucu.edu.uy

**Abstract.** In this work we explore the use of computer vision for bus detection in the context of intelligent transport systems. We propose a simple and efficient method to detect moving objects using a probabolistic modelling of the scene. For classification of the detected moving regions we study the use of eigenfaces.

## 1 Introduction

In recent years there has been an increasing interest on improving public transport services. This is due to both direct and indirect benefits that can be obtained by cities from a logistical, environmental and social point of view. As a direct benefit, a well organized and efficient public transport system allows to reduce travel times, reduce traffic congestion while offering a comfortable alternative to family cars, and therefore also reduces pollution levels. On the other hand, one of the indirect benefits is the impact on the economy of the city. Big cities are increasingly important in the economy of the countries; providing appropriate logistical infraestructure helps to attract more business and investment. Governments recognize this reality and are increasingly investing in infrastructure, highways, subways, etc.

In order to improve public transport systems the main infrastructure investment is in exclusive bus corridors. However, the high costs associated with them make it not feasible to do it in all arteries of the city. Therefore, an alternative solution is the demarcation of preferential lanes for public transport in already existing streets and avenues. To be truly effective this solution requires the classification of vehicles traveling on them to detect buses and give them right of way at traffic lights.

In recent years the use of Computer Vision has expanded its application in Intelligent Transportation Systems, in particular for vehicle classification [2,5,3]

In the case of exclusive corridors, the detection of buses can be easily implemented because there is a physical separation between them and private vehicles. In the case of preferential lanes this is not the case. A computer vision based system has two major advantages, its cost and its scalability. It is of low cost, compared to other solutions, because it enables the use of preferential lanes and reduces the construction cost associated with exclusive lanes. Computer vision

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 59–66, 2013.

it is also an interesting technology that can be used for other purposes such as vehicle counting, detecting special vehicles, speeding control and surveillance.

In this work we present a system that applying computer vision automatically detects buses preferential lanes to enable the synchronization of traffic lights on a main road in order to minimize travel times.

## 2 Existing Solutions

To solve the problem of vehicle detection a sensor is required. Sensors can be classified into two types: intrusive and non-intrusive.

Intrusive sensors include inductive loops, piezoelectric cables and magnetometers among others. These are installed directly on the floor, either on it or under by pipeline as is shown in Figure 1. The operation thereof is generally simple and well known because they are mature technologies. The major disadvantage is that they require traffic disruptions for installation and maintenance. They also have many flaws associated with pavement condition and life depends a lot on the installation procedures.

On the other hand, non-intrusive sensors may be installed and maintained with minimal traffic distortion (Figure 1). These sensors include computer vision based solutions, microwave radar, laser, infrared detection etc. They allow the supervision of several lanes and are able to provide further information such as the vehicle type detected.

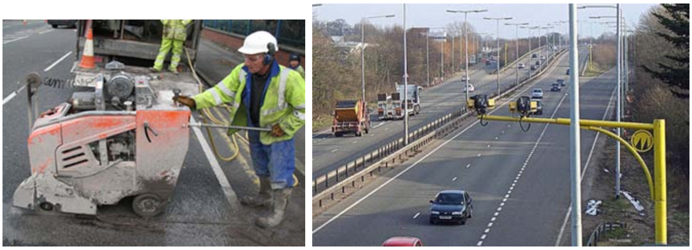


**Fig. 1.** Left: Installation of intrusive loops. Right: Computer Vision sensors.

In the case of exclusive corridors, intrusive sensors can be used for bus detection because there is a physical separation between them and private vehicles. For preferential lanes (most widespread solutions due to its lower costs) these sensors can not ensure the correct detection due to private vehicles. Some recent works that address the problem of sorting vehicles according to their magnetic signature captured with inductive loops was presented in [1].

The advantage of computer vision solution is the high value that can be generated due to its greater scalability. Most large scale traffic control systems use them for a variety of applications. First, they let you have a visual view of the traffic state, particularly useful when analyzing the causes of accidents that may result in roads. They also allow more elaborated statistical data analysis such as the level of congestion and the use of avenues. This can be obtained by calculating the occupation times and queues lengths. Computer vision solutions also allow to calculate the speed of vehicles that can be used for speeding infractions. It is also possible to do plate recognition, which can be used in shadow tolling, among other things.

In the next section we review the literature on computer vision based solutions for intelligent transport systems.

## 3 Computer Vision for Intelligent Traffic Systems

Although the work presented in [1] it is not based in computer vision it proposes a pattern recognition approach to classify vehicles by their magnetic signature obtained using inductive loops. The magnetic signature is a characteristic of each vehicle which depends on their geometry and distribution of metal parts. One way to obtain the magnetic signature is via the oscillation frequency versus time of an oscillator that uses an inductive loop during the time the vehicle passes over it. The database used is rather small and contains 34 cars, 9 vans and 18 buses. With this database the best classification is obtained with a naive Bayes classifier in the dissimilarity space with 99,3% of the vehicles correctly classified. Due to the small size of the database this can be considered as an upper bound of the classification performance. Nevertheless, the use of inductive loops it is a very robust method for this problem, and although it is an intrusive method, we include these results here to use it as comparison of our proposal.

In [2] a real-time vehicle detection based on background learning of the scene is proposed. The method has two processes running in parallel, one at high level and another one at low level. The low level process is the estimation of the background and runs in real time. The high level one runs at a lower frequency and is responsible of classifying the pixels into categories: lines, pavement or neither of the aforementioned. For this classification, color and shape features are considered. The proposed method only detects vehicles but not classifies them; it obtains a 90% of correct detection rate. We include this work to have a reference on the detection rate.

The work presented in [5] addresses the real-time vehicle classification based on eigenfaces [4]. The method implies taking a set of training images to learn the features of each class of interest: buses, cars, etc. The feature space is calculated using principal components analysis and then a nearest neighbor classifier is applied. The published results show a classification rate of 100% but using as the test set the same set used for the training consisting of 100 images of vehicles fronts. We apply the same methodology but considering complete images of the vehicles as shown in Figure 2. We also train and test the algorithm with independent sets to evaluate different classifiers to understand the potential of the method in a real scenario.

Finally in [3] an algorithm for detection and classification of vehicles is presented. The detection achieves a good detection rate of 90% but the recognition based on structural features only achieves 70% of correct classification.

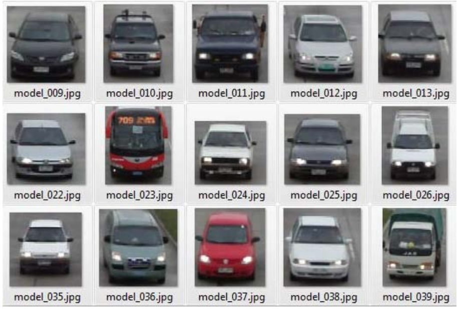


**Fig. 2.** Models uses for trainning

## 4 Proposed Method

Using a camera conveniently located we are able to capture traffic images that are processed in real time in order to detect the presence of buses. Before proceeding we shall mention that we assume the following working conditions: good visibility and normal weather conditions.

Our system is composed of the following modules: image acquisition, segmentation, feature extraction and classification. The first step of image acquisition also prepares the image for further processing, for example adding noise filtering operations. In the segmentation step vehicles in motion are extracted from the background. For this step we propose a probabilistic approach that facilitates the typical selection of thresholds for segmentation purposes. Once the regions of interest are obtained from the segmentation step, each region is expanded in the feature space given by the eigenfaces method [4]. Finally, based on the calculated features each moving region is assigned to a given class (car, bus, truck, etc.).

### 4.1 Segmentation

Given two consecutive frames at times $n$ and $n+1$, $I_n(x)$ and $I_{n+1}(x)$ we consider their difference $d(x) = |I_n(x) - I_{n+1}(x)|$ where $x$ indicates the pixel. If we assume that most of the pixels belong to the background and that differences among these pixels are assumed to be small, then most of the pixels will exhibit small values in $d(x)$. If we look at the histogram of $d(x)$ we will find that most of the pixels are concentrated at small values. If we view the image intensity differences as a random variable, whose magnitude represents the probability that a pixel belongs or not to a moving object, we can interpret the histogram of $d(x)$ as an empirical approximation to the density function, $f(y)$. Given a threshold $\alpha$ the probability that the difference falls in $[\alpha, 255]$ is:

$$P(\alpha \leq d \leq 255) = 1 - \sum_{y=0}^{\alpha} f(y).$$

Instead of fixing the threshold $\alpha$ which can vary due to lightning conditions and shadows, we fix the probability instead. That is, a pixel $x$ is declared as moving if the probability of its difference $d(x)$ is below a probability threshold $P_\alpha$: $P(\alpha \leq d(x) \leq 255) \leq P_\alpha$. We define the following probabilities image:

$$IMP_n(x) = 1 - \sum_{y=0}^{\alpha} f(y),$$

and with it a binary image with objects labeled as one: $IMS_n(x) = IMP_n(x) \leq P_\alpha$. In all the experiments of this paper we choose $P_\alpha = 0.08$. In Figure 3 the probability image and the final segmentation mask. Once we have the binary image with detected moving objects we apply mathematical morphology to simplify the detected regions followed by a labelling process which labels all connected components and extract their bounding boxes. In Figure 4 we show an image with all detected regions of interest. All bounding boxes are processed with some heuristic to join close bounding boxes and remove the ones that dont fulfill basic requirements of size and shape, see Figure 3. Finally we apply an optimization procedure to shrink the bounding boxes towards the real boundaries of the vehicles. To do this we use the integral image of $IMP_n(x)$ to move the bounding box inwards to minimize the mean probability inside the bounding box; the bounding box is reduced until no noticeable reduction is observed. The use of the integral image allows a fast implementation. Observe that the same procedure must be applied to each bounding box in the image. See Figure 4 for an example of the output of this procedure.

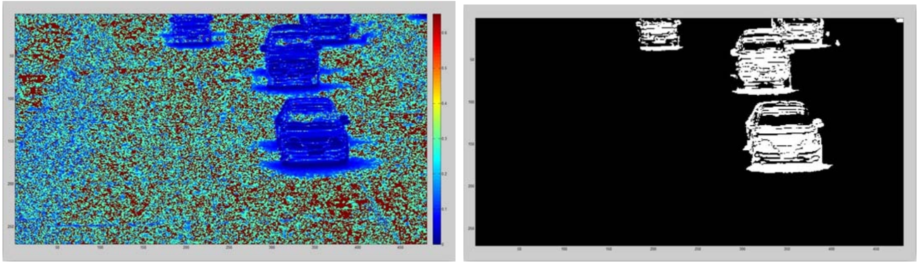


**Fig. 3.** Left: Probabilities image. Right: Segmentation of moving objects.

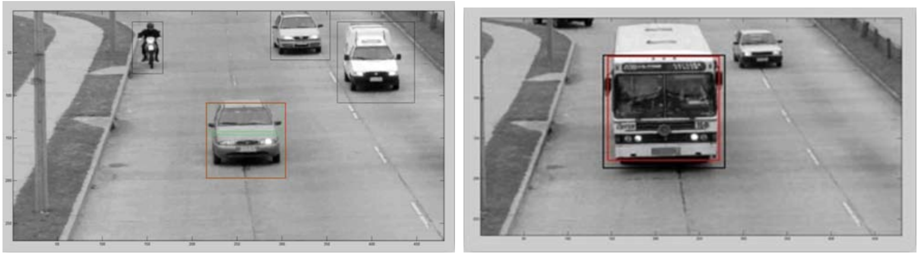


**Fig. 4.** Left: Probabilities image. Right: Segmentation of moving objects.

### 4.2 Classification

Once we have the bounding boxes of the regions of interest we use a classifier to decide the class of each of them. Starting from the greyscale values of regions of interest, see Figure 2, we apply the method of eigenfaces to extract the projection coefficients and use them as classification features. The method of eigenfaces [4], originally developed to recognize faces can be applied to our problem of vehicle recognition. In order to make this article self contained we are going to summarize the main concepts behind eigenfaces.

The training set of contains $M$ vehicles images of known class scanned in lexicographical order. We assume that all images are resized to have the same size $N \times N$ and therefore each sample $\Gamma_k$ in the training set has $N^2$ elements; $\Gamma = \{\Gamma_1, , \Gamma_M\}$ The eigenfaces method is based on principal component analysis (PCA). The first step for the application is to remove the mean of the samples:

$$\Phi_k = \Gamma_k - \frac{1}{M} \sum_{i=1}^{M} \Gamma_i$$

Then, the eigenvectors $u_i$ of the covariance matrix $C$ of the set $\{\Phi_1, ..., \Phi_M\}$ are computed. The matrix $C$ is calculated as:

$$C = \frac{1}{M} \sum_{i=1}^{m} \Phi_i \Phi_i^t = \frac{1}{M} AA^t,$$

where $A$ is a matrix with vectors $\Phi_k$ as columns.

Since this matrix $C$ size is $N^2 \times N^2$ we use the same idea of [4] and compute the eigenvectors of $A^tA$. It can be shown that if $v_i$ is an eigenvector of $A^tA$ then $Av_i$ is an eigenvector $AA^t$. In this way we obtain $M$ eigenvectors and eigenvalues of $AA^t$. Once we have the eigenvectors all the data points in the training set are projected into them to build the feature space. That is, the eigenvectors $u_i$ constitute the base of the feature space and the coordinates on each eigenvector are the features to be used for classification.

During recognition, each region of interest is resized to size $N \times N$, the mean of the training set is removed and then projected to $\{u_1, ..., u_L\}$ to obtain its coordinates in the feature space. In this space we can apply any supervised classifier. For this work we tested five standard classifiers to select the one with best performance and lowest computational cost.

The training set was build using samples for each of the target classes as shown in Table 1. In Figure 2 we show some of the image samples. As we can see the sample contain the object of interest but no fine tuning was imposed.

**Table 1.** Number of samples for each class in the testing set

| Class | #Samples | Class | #Samples |
|---|---|---|---|
| Bus | 76 | Truck | 46 |
| 4x4 | 54 | School Buses | 21 |
| Vans | 43 | Cars | 115 |
| Motorbikes | 30 | | |
| | Total | 385 | |

## 5 Classification Results

Five classifiers were trained following the procedure of previous sections and performance evaluated using 10-fold crossvalidation. Although our main goal is to detect buses we also evaluated the potential of the approach to correctly classify all the classes of vehicles contained in the training set. In Table 2 we show the results for the five selected classifiers.

**Table 2.** Bus classification performance for different classifiers. The last column contains the correct classification for the seven classes.

| Classifier | TP Rate | FP Rate | Precision | Recall | Global Correct Class. |
|---|---|---|---|---|---|
| Naive Bayes | 0,855 | 0,016 | 0,929 | 0,855 | 73,5% |
| Neural Network | 0,895 | 0,023 | 0,907 | 0,895 | 74,3% |
| Random Forest | 0,868 | 0,032 | 0,868 | 0,868 | 68,3% |
| k-NN (k=1) | 0,947 | 0,006 | 0,973 | 0,947 | 80,3% |
| k-NN (k=3) | 1,000 | 0,010 | 0,962 | 1,000 | 79,2% |

If we concentrate ourselves in the correct recognition of buses, that is we measure the correct classification into two metaclasses buses and non-buses, the performance of the five classifiers increases. Table 3 we show the percentage of correctly classified buses.

As we can see the best classifier is k-NN with $k = 3$; it achieves 100% of true positives and only 3 false positives. Although, the five classifiers obtain good classifier results, it is importante to note that the nearest neighbor classifiers are the ones with best performance in terms of true and false positives. This is important due to the difference between the number of samples in bus and non-bus classes.

**Table 3.** Classification performance for different classifiers into two classes bus and non-bus

| Classifier | Bus/Non-Bus Correct Class. | Bus TP Rate | Bus FP Rate |
|---|---|---|---|
| Naive Bayes | 95,8% | 85,5% (65/76) | 1,6% (5/309) |
| Neural Network | 96,1% | 89,4% (68/76) | 1,9% (7/309) |
| Random Forest | 94,8% | 86,8% (66/76) | 2,9% (9/309) |
| k-NN (k=1) | 98,4% | 94,7% (72/76) | 0,6% (2/309) |
| k-NN (k=3) | 99,2% | 100% (76/76) | 0,9% (3/309) |

## 6 Discussion and Conclusions

The complete sistem was tested with a set of videos where 14 buses where correclty classified among above 100 vehicles and only 1 false positive was produced. We note that a false positive is tolerable while missing a bus is not. From this point of view the proposed system achieves good performance in terms of recall and precission.

If we compare our results with the ones presented in [1] using inductive loops we can see that we we obtain 81,2 % in the classification in all classes while in the preovious work the author reports 98,4% over a smaller set consisting only in cars, vans and buses. If we consider both systems as bus detectors their performance is equivalent since we reach 100%.

When comparing our results to [5] the first obsevation is that in this work the authors recognize car categories and therefore a direct comparison is not posible. We showed that using the same basic algorithm, eigenfaces, buses can be correctly recognized among other vechiles. This shows the potential of this solution for real applications in the context of intelligent transport systems.

We also propose a simple and efficient method to detect moving regions based on the probability of a pixel to be part of the static background. This method simplifies the threshold selection and auto-adapts its value based on the probability distribution of the static pixels from the background.

## References

1. Derregibus, A.: Clasificación de vehículos con lazos inductivos. Master's thesis, Universidad Catolica del Uruguay (2012)
2. Tan, X.J., Li, J., Liu, C.L.: A video-based real-time vehicle detection method by classified background learning. World Transactions on Engineering and Technology Education 6, 189–192 (2007)
3. Mo, S., Liu, Z., Zhang, J., Wu, C.: Real-time vehicle classification method for multi-lanes roads. In: IEEE Conference on Industrial Electronics and Applications, pp. 960–964 (2009)
4. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
5. Wang, W., Shang, Y., Guo, J., Qian, Z.: Real-time vehicle classification based on eigenface. In: International Conference on Consumer Electronics, Communications and Networks, pp. 4292–4295 (2011)

# Music Genre Recognition Using Gabor Filters and LPQ Texture Descriptors

Yandre Costa[1,2], Luiz Oliveira[2],
Alessandro Koerich[2,3], and Fabien Gouyon[4]

[1] State University of Maringá (UEM), Maringá, PR, Brazil
[2] Federal University of Paraná (UFPR), Curitiba, PR, Brazil
[3] Pontifical Catholic University of Paraná (PUCPR), Curitiba, PR, Brazil
[4] Institute for Systems and Computer Engineering of Porto (INESC), Porto, Portugal
yandre@din.uem.br, lesoliveira@inf.ufpr.br,
alekoe@ppgia.pucpr.br, fgouyon@inescporto.pt

**Abstract.** This paper presents a novel approach for automatic music genre recognition in the visual domain that uses two texture descriptors. For this, the audio signal is converted into spectrograms and then textural features are extracted from this visual representation. Gabor filters and LPQ texture descriptors were used to capture the spectrogram content. In order to evaluate the performance of local feature extraction, some different zoning mechanisms were taken into account. The experiments were performed on the Latin Music Database. At the end, we have shown that the SVM classifier trained with LPQ is able to achieve a recognition rate above 80%. This rate is among the best results ever presented in the literature.

**Keywords:** Music genre, texture, image processing, pattern recognition.

## 1 Introduction

In recent years, a huge amount of data from different sources has become available online. In most cases, this information is not organized according to some predefined pattern. Thus, tasks related to automatic search, retrieval, indexing and summarization has become important questions, whose solutions could support a good and efficient access to this content. For some time, textual annotation was used to organize and classify multimedia data. However, this is not a good way to deal with this content efficiently. Textual annotation requires a large amount of human labor and, moreover, is subject to human perception subjectiveness.

Digital music is among the most common types of data distributed through the internet. There are a number of studies concerning to audio content analysis using different features and methods. Automatic music genre recognition is a crucial task for a content based music information retrieval system. As stated by Tzanetakis and Cook in [1], musical genres are categorical labels created by humans to characterize pieces of music. A musical genre is characterized by the

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 67–74, 2013.

common characteristics shared by its members. These characteristics typically are related to the instrumentation, rhythmic structure, and harmonic content of the music. In some studies it was found that genre is an important attribute which helps users in organizing and retrieving music files.

Costa et al. presented in [2] the first results obtained in music genre classification using features extracted from spectrograms. Spectrogram is a visual representation of the spectrum of frequencies in a sound [3]. In the most common representation, spectrogram is a graph with two geometric dimensions: the horizontal axis represents time, the vertical axis is frequency; a third dimension indicating the amplitude of a particular frequency at a particular time is represented by the intensity or color of each point in the image. As shown in Figure 2, texture is the most noticeable visual content in a spectrogram image. Taking this into account, we have explored different texture descriptors presented in the image processing literature in order to capture information to describe this content. In [2], we used the well-known Gray Level Co-occurrence Matrix (GLCM) to capture the textural content from the spectrogram images. By analyzing the spectrogram images, we have noticed that the textures are not uniform, so we decided to consider a local feature extraction beyond the global feature extraction. In that work, only one classifier was created even when a zoning strategy was used in order to preserve local information, and the final decision was done through majority voting among the results obtained with feature vectors extracted from different zones. In [4] and [5], the authors have evaluated the Local Binary Pattern (LBP) texture descriptor trying to capture the spectrogram image content. Furthermore, the authors introduced the creation of one classifier for each created zone, combining their outputs in order to get the final decision using fusion rules presented by Kittler *et al.* [6], like Product, Sum, Max and Min. The best obtained results on the ISMIR 2004 dataset are comparable to the best results described in the literature. Regarding LMD dataset, the best obtained result is the best ever obtained using artist filter.

In this work, we are interested in investigate the performance of LPQ and Gabor filters texture operators in music genre recognition using spectrogram images. The reason for choosing Gabor filters is that in our previous works, there is a lack of experiments using some spectral texture descriptor approach. With regard to LPQ, the choice was done because this is a novel operator which has shown good performance in many different works presented in the literature.

This paper is organized as follows: Section 2 describes the feature extraction performed in this work. Section 3 describes the classification while Section 4 reports the results and discussions about them. Section 5 concludes this work.

## 2 Feature Extraction

Before proceed the generation of the visual representation, we performed a time decomposition based on the idea presented by Costa et al. [7] in which an audio signal $S$ is decomposed into $n$ different sub-signals. Each sub-signal is simply a projection of $S$ on the interval $[p, q]$ of samples, or $S_{pq} =< s_p, \ldots, s_q >$.

In the generic case, one may extract $K$ (overlapping or non-overlapping) subsignals and obtain a sequence of spectrograms $\overline{\Upsilon}_1, \overline{\Upsilon}_2, \ldots, \overline{\Upsilon}_K$. We have used the same strategy used in [8], which considers three 10-second segments from the beginning ($\overline{\Upsilon}_{beg}$), middle ($\overline{\Upsilon}_{mid}$), and end ($\overline{\Upsilon}_{end}$) parts of the original music. In order to avoid segments that do not provide good discrimination among genres, we decided to ignore the first ten seconds and the last ten seconds of the music pieces. The rationale behind this strategy is that some common effects present in these parts of the music signal, like fade in and fade out, and some kinds of noise, like those produced by the audience, could turn these signal samples less discriminant than the others.

After the signal decomposition, the next step consists in converting the audio signal into a spectrogram. The spectrograms were created using a bit rate = 352kbps, audio sample size = 16 bits, one channel, and audio sample rate = 22.05 kHz. Figure 1 depicts the signal segmentation and spectrogram generation.

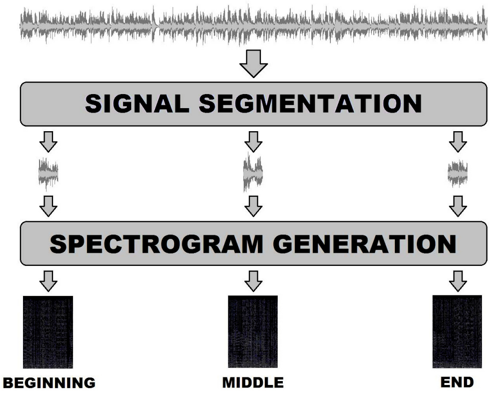


**Fig. 1.** Creating spectrograms using time decomposition

Once the spectrograms were generated we proceeded the texture feature extraction from these images. As stated before, the approach proposed in this work considers that the main visual content present in the spectrogram images is the texture. With this in mind, we used Gabor filters and LPQ texture operator to capture the image content.

In this work, before proceeding the feature extraction with Gabor filters, the spectrogram images were scaled to 64×64 pixels. Once it was done, the Gabor wavelet transform was applied on the scaled image with 5 different scale levels and 8 different orientations, which results in 40 subimages. For each subimage, 3 moments are calculated: mean, variance and skewness. So, a 120-dimensional vector is used for Gabor texture features. More details about Gabor filters can be found in [9].

Our experiments with LPQ were performed with the original implementation. The window size used to compute the short-term Fourier Transform was empirically adjusted to 7 × 7. Additional mathematical details about LPQ can be found in [10].

### 2.1 Global and Local Feature Extraction

The rationale behind the zoning and combining scheme is that music signals may include similar instruments and similar rhythmic patterns which leads to similar areas in the spectrogram images. By zoning the images we can extract local information and try to highlight the specificities of each music genre.

A positive side effect obtained with zoning strategy is that one can create a specific classifier to deal with the features extracted from each specific zone. Thus, we can naturally obtain several classifiers. Not by chance, the best results achieved in previous works were obtained by combining these classifiers outputs.

In order to proceed the local feature extraction, we have evaluated three different number of linear zones (1,5, and 10), which are applied to the spectrogram image before extracting textural features. Thus, considering that three spectrogram images were generated from each music piece, since we extracted three segments, the number of total zones and consequently the number of classifiers is $3n$, where $n$ is the number of zones per segment. Figure 2 shows a linear zoning scheme, with $n = 10$, superimposed over a spectrogram image extracted from 30 seconds signal (three segments of ten seconds).

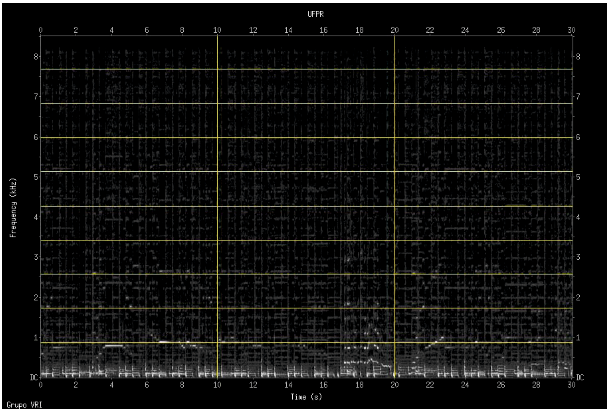


**Fig. 2.** Linear zoning used to extract local information

## 3 Classification

The classifier used in this work is Support Vector Machine (SVM), introduced by Vapnik in [11]. Normalization was performed by linearly scaling each attribute to the range [-1,+1]. The Gaussian kernel was used, with parameters $C$ and $\gamma$ tuned using a greedy search.

The classification process is done as follows: as aforementioned, the three 10-second segments of the music are converted to the spectrograms ($\overline{\Upsilon}_{beg}$, $\overline{\Upsilon}_{mid}$, and $\overline{\Upsilon}_{end}$). Each of them is divided into $n$ zones, according to the values of $n$ described in subsection 2.1. Then, a 120-dimensional Gabor filters feature vector and a 256-dimensional LPQ feature vector were extracted from each zone. Next, each one of these feature vectors is sent to a specific classifier, which assigns a prediction to each one of the ten possible classes. Training and classification were carried out using the 3-fold cross-validation. For each specific zoning scheme, we created $3n$ classifiers with 600 and 300 feature vectors for training and testing, respectively. With this amount of classifiers, we used estimation of probabilities to proceed the combination of outputs in order to get a final decision. In this situation, is very useful to have a classifier producing a posterior probability $P(class|input)$. Here, we are interested in estimation of probabilities because we want to try different fusion strategies like Max, Min, Product, and Sum.

## 4 Experimental Results and Discussion

Firstly, some details about the music database used in the experiments reported here are described. The Latin Music Database (LMD) is a digital music database created for support research in music information retrieval. The database was presented by Silla et al. [12]. It is composed of 3,227 full-length music samples in MP3 format originated from music pieces of 501 artists. The database is uniformly distributed along 10 music genres.

In our experiments we have used the artist filter [13] restriction when splitting the dataset to create folds. The use of the artist filter does not allow us to employ the whole dataset since the distribution of music pieces per artist is far from uniform. Thus, 900 music pieces from the LMD were selected, which are split into 3 folds of equal size (30 music pieces per class). In order to compare the results obtained here with those obtained in other works, the folds splitting taken was exactly the same used by Lopes et al. [14] and by Costa et al. [2] [4] [5]. The results described here refer to the average recognition rate considering the three folds aforementioned. In addition, the standard deviation between the three folds used in classification is presented.

Table 1 reports the results obtained when features extracted with Gabor filters were used with four different fusion rules and with the three different zoning configurations mentioned in section 2.1. As in the results presented in [4], the best result was obtained when five zones were created. Like in that work, one can see that increasing the number of zones up to a certain point we observe a noticeable performance improvement.

Table 2 presents results obtained using LPQ texture descriptor. Interestingly, the best result with LPQ, both in terms of recognition rates and standard deviation, were obtained when the global feature extraction (without zoning) was used. One can notice that the results obtained with global feature extraction and five linear zones are very close to each other. However, it is important to contrast that using global feature extraction, only three classifiers are created

**Table 1.** Average recognition rates (%) and standard deviation obtained between the three folds using different number of zones with Gabor filters

| Number of zones | Maximum rule | Minimum rule | Product rule | Sum rule |
|---|---|---|---|---|
| 1 | 55.89±9.94 | 56.67±11.60 | 59.78±9.91 | 58.78±9.08 |
| 5 | 66.22±2.22 | 69.67±2.33 | **74.67±3.79** | 74.11±2.69 |
| 10 | 60.56±1.02 | 65.33±2.85 | 71.78±1.84 | 71.00±0.58 |

whereas 15 are created when five linear zones are created. In addition, the best result obtained with LPQ is very close to, but sligtly better, the best result reported in [4], obtained with Local Binary Pattern (LBP) texture descriptor.

**Table 2.** Average recognition rates (%) and standard deviation obtained between the three folds using different number of zones with LPQ

| Number of zones | Maximum rule | Minimum rule | Product rule | Sum rule |
|---|---|---|---|---|
| 1 | 76.89±2.12 | 77.22±1.68 | **80.78±0.77** | 79.44±1.17 |
| 5 | 74.00±1.91 | 76.00±1.66 | 80.67±1.44 | 80.56±1.10 |
| 10 | 70.11±2.57 | 73.33±1.25 | 79.00±0.89 | 78.00±0.27 |

### 4.1 Discussion

Unlike the results obtained with Gabor filters and the texture descriptors used in [4], i.e. LBP and GLCM, the best result with LPQ was obtained using global feature extraction, as shown in table 2. This is very interesting, once with global feature extraction we create a smaller amount of classifiers, which decreases the overall system complexity. In addition, it is important to notice that the result obtained with LPQ is the best one ever obtained with linear zoning or global feature extraction taking into account all the texture descriptors already experimented on the LMD dataset.

**Table 3.** Recognition rates (%) with all the texture descriptors used here and in [4]

| Texture descriptor | Number of zones | Best result |
|---|---|---|
| GLCM [4] | 5 | 70.78±2.69 |
| LBP [4] | 5 | 80.33±1.67 |
| Gabor filters | 5 | 74.67±3.79 |
| LPQ | 1 (no zoning) | **80.78±0.77** |

Table 3 presents the best results obtained with four different texture operators on the LMD. We have evaluated if there are statistically significant differences between these results. For this, the Friedman test with post hoc Shaffer's static procedure was employed. The multiple camparison statistical test has shown that the $p$ value of the statistical test was higher than the critical value in all cases at 95% confidence level. Thus, we have not found statistically significant difference between these results. This is favourable to LPQ, once it is the only one operator which presented the best result using global feature extraction.

Table 4 shows some results recently obtained on the LMD dataset using artist filter. Some of the works shown in this table refer to results presented in MIREX (Music Information Retrieval Evaluation eXchange) contest. In [15,16,17] the authors used acousitc features, extracted directly from the audio signal. One can see that the best result obtained here is among the best results.

**Table 4.** Best recognition rates (%) obtained on the LMD with artist filter

| Work reference | Recognition rate (%) |
|---|---|
| Lopes et al. [14] | 59.67±13.5 |
| MIREX 2008 - LMD [15] | 65.17±10.72 |
| MIREX 2009 - LMD [16] | 74.67±11.03 |
| MIREX 2010 - LMD [17] | 79.86±5.20 |
| LBP (5 zones) [4] | 80.33±1.67 |
| LBP (Mel scale zoning) [5] | **82.33±1.45** |
| LPQ (this work) | **80.78±0.77** |

On the one hand, one can say that the best recognition rate obtained on the LMD using visual features is that described in [5]. On the other hand, it is important to note that in that work a much bigger amount of classifiers (45) was created, since a nonlinear zoning with much more zones was used.

## 5 Conclusion

In this work we follow the investigation of the use of features extracted from the visual representation (spectrogram) of the audio signal in music genre recognition. We have compared the use of two different texture descriptors to capture the content of spectrogram images, i.e. Gabor filters and LPQ. We have tried two different approaches to deal with the intra-class variability of the spectrogram images, a global feature extraction and a feature extraction taking into account a linear zoning to obtain local information of the images.

The results obtained with LPQ texture operator are better than those obtained with Gabor filters. Regarding to results obtained with other texture descriptors on the LMD with global feature extraction or linear zoning, the result obtained with LPQ is the best one ever obtained. Interestingly, the global feature extraction performed slightly better than zoning with LPQ, unlike with Gabor filters and the other texture descriptor already investigated in other works.

In future works, we intend to develop experiments using LPQ descriptors with feature selection. The rationale behind this strategy is that one can reduce the dimensionality of the features vector and improve the performance either in terms of recognition rate or in terms of time.

**Acknowledgments.** The authors thank to CNPq (Grants 301653/2011-9, 402357 /2009-4), CAPES (Grant BEX5779/11-1) and Araucaria fundation (Grant 16767).

## References

1. Tzanetakis, G., Cook, P.: Music Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing, 293–302 (2002)
2. Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F.: Music Genre Recognition Using Spectrograms. In: 18th International Conference on Systems, Signals and Image Processing. IEEE Press, Sarajevo (2011)
3. Haykin, S.: Advances in spectrum analysis and array processing, vol. 3. Prentice-Hall, NJ (1991)
4. Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F.: Comparing Textural Features for Music Genre Classification. In: WCCI 2012 IEEE World Congress on Computational Intelligence, Brisbane, Australia, pp. 1867–1872 (2012)
5. Costa, Y.M.G., Oliveira, L.E.S., Koerich, A.L., Gouyon, F., Martins, J.G.: Music genre classification using LBP textural features. Signal Processing 92(11), 2723–2737 (2012)
6. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239 (1998)
7. Costa, C.H.L., Valle Jr., J.D., Koerich, A.L.: Automatic classification of audio data. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 562–567 (2004)
8. Silla Jr., C.N., Kaestner, C.A.A., Koerich, A.L.: Classificação de Gêneros Musicais Utilizando Vetores de Característica Híbridos. In: 13o Simpósio Brasileiro de Computação Musical (SBCM 2011), pp. 32–44 (2011)
9. Gabor, D.: Theory of communications. Journal of Institution of Electrical Engineers 93, 429–457 (1946)
10. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
11. Vapnik, V.: The nature of statistical learning theory. Springer, New York (1995)
12. Silla Jr., C.N., Koerich, A.L., Kaestner, C.A.A.: The latin music database. In: Proceedings of the 9th International Conference on Music Information Retrieval, Philadelphia, USA, pp. 451–456 (2008)
13. Flexer, A.: A closer look on artist filter for musical genre classification. In: International Conference on Music Information Retrieval, pp. 341–344 (2007)
14. Lopes, M., Gouyon, F., Koerich, A.L., Oliveira, L.E.S.: Selection of Training Instances for Music Genre Classification. In: ICPR 2010 - 20th International Conference on Pattern Recognition, Istanbul, Turkey (2010)
15. MIREX: Music information retrieval evaluation exchange (2008), `http://www.music-ir.org/mirex/wiki/2008:Main_Page`
16. MIREX: Music information retrieval evaluation exchange (2009), `http://www.music-ir.org/mirex/wiki/2009:Main_Page`
17. MIREX: Music information retrieval evaluation exchange (2010), `http://www.music-ir.org/mirex/wiki/2010:Main_Page`

# Unseen Appliances Identification

Antonio Ridi[1,2], Christophe Gisler[1,2], and Jean Hennebert[1,2]

[1] University of Applied Sciences Western Switzerland
College of Engineering and Architecture of Fribourg, ICT Institute
{antonio.ridi,christophe.gisler,jean.hennebert}@hefr.ch
[2] University of Fribourg
Department of Informatics, Fribourg, Switzerland
{antonio.ridi,christophe.gisler,jean.hennebert}@unifr.ch

**Abstract.** We assess the feasibility of unseen appliance recognition through the analysis of their electrical signatures recorded using low-cost smart plugs. By unseen, we stress that our approach focuses on the identification of appliances that are of different brands or models than the one in training phase. We follow a strictly defined protocol in order to provide comparable results to the scientific community. We first evaluate the drop of performance when going from seen to unseen appliances. We then analyze the results of different machine learning algorithms, as the k-Nearest Neighbor (k-NN) and Gaussian Mixture Models (GMMs). Several tunings allow us to achieve 74% correct accuracy using GMMs which is our current best system.

**Keywords:** Intrusive Load Monitoring (ILM), appliance recognition, electric signatures, load identification.

## 1 Introduction

The automatic recognition of appliances from their electric signatures has several applications such as energy consumption understanding and appliance management for energy consumption optimization [1]. Other applications can also be envisioned such as an indirect activity detection in houses or monitoring of elderly people [2].

Due to the rising price of energy and an increased sensitivity from people to environmental matters, the field of energy consumption understanding and management is nowadays rising interests. In US, 51% of the electricity consumption in homes is due to appliances and lighting [3]. In this context, a system able to recognize appliance would allow to know which appliance is consuming how much, giving an explanation on their contribution to the electricity bill. This will allow householders to optimize their energy consumption. Appliance identification could also be very useful for Building Management Systems (BMS), allowing to implement smarter rules and optimizing the local production and consumption of electric energy.

An electric signature represents the time evolution of the electricity consumption which is summarized by the active and reactive power on AC networks.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 75–82, 2013.

Appliances can be categorized into 4 classes [4]: two-states on/off appliances (e.g. lamps, toasters); multi-states appliances, when a finite number of operating states exist (e.g. fridges, dishwashers); continuously variable devices, when the consumption varies continuously (e.g. battery chargers); permanent consumer devices, when the consumption is constant over a long period of time (e.g. telephone sets, smoke detectors). According to this, a more complex task than appliance identification could consist of recognizing in which state a given appliance is at a given time, allowing for example to automatically detect stand-by. Recent studies are estimating this consumption at about 10% of the residential electricity use [5,6].

Appliance identification can be done using two approaches: Non-Intrusive Load Monitoring (NILM) and Intrusive Load Monitoring (ILM) [7]. NILM monitors the total house electricity consumption at the smart meter, while the ILM refers to a distributed sensing approach, using one or more sensor per appliance. In the first case, the signals have to be decomposed to identify single appliance, i.e. performing a *disaggregation* [4]. NILM approaches are less expensive but more difficult while ILM approaches are more expensive but more precise [7]. We focus in this paper on ILM approaches.

As detailed in Section 2, several modeling approaches have been proposed for appliance identification, often based on machine learning principles. Given the differences among brands and models, a challenge for such approaches is in the necessity to have large training databases that represent all types of appliances for a given type, including as many brands and models as possible. In this paper we address the problem of identification systems that are tested with unseen appliance, i.e. appliance brands that are not available in the training set. In other words, we evaluate the generalization capacity of such machine learning systems when dealing with new appliances that are not yet observed in the training set. In this direction, the availability of large databases is important (Section 3). System description, results and discussions are presented in Section 4 and 5.

## 2 Related Works

Several ILM approaches have been proposed. In the work of F. K. Adeel Abbas Zaidi and P. Palensky [8], machine learning approaches such as Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs) are presented. Observations are sampled at $10^{-1}$ Hz on different appliances spread into 6 categories including fridges, microwaves, dishwashers, coffee machines, computers and printers. From the raw observations, features are extracted such as average energy consumption, edge counts, percentage energy consumption and discrete Fourier transform coefficients. The best feature sets are showing results up to 90% for five categories.

In the work of Reinhardt et al. [9], 33 appliance categories are used to build an identification system showing promising performance up to 95.5% accuracy. The system samples the current consumption at 1.6 kHz which brings much finer information on the time evolution. Their approach extracts numerous features from the signal leading to 517 feature vector representing the electricity trace. Different classification algorithms were also analyzed showing the best results with random committee approaches. Due to a pretty large database of signals, they could analyze the impact of using different features and types of classifiers.

In the work of Zufferey et al. [10], the objective was to categorize appliances into 6 categories. The system is based on low-cost smart plugs measuring the electricity consumption parameters at low frequency every 10 seconds. k-Nearest Neighbors and Gaussian Mixture Models were compared, showing similar accuracies up to 85%. Interestingly, the raw observations were simply normalized and used directly as features. A continuation of this work was presented by Ridi et al. in [11] where the signature database ACS-F1 was used, increasing the number of categories to 10 and showing a tuned up system performance of 93.8%. In the next Section this database will be presented.

To the best of our knowledge, all these related works have been evaluated on appliance types and brands that were also seen in the training database, i.e., according to so-called *intersession* protocols where the same appliances are producing the training and testing signature materials. In this paper, we investigate the recognition of categories using *unseen appliance* protocols, i.e., where the testing signatures come from new appliances that are not observed in the training set. The task is expected to be more complex due to the extra inter-brand and inter-model appliance variability.

## 3 ACS-F1 Database

We based our work on the Appliance Consumption Signature Fribourg 1 (ACS-F1) database [12]. This database contains appliance signatures acquired using low-cost smart plugs capturing the electricity parameters at low frequency with a sampling rate of $10^{-1}$ Hz. A signature is a sequence of raw measurements $O = \{o_1, \ldots, o_N\}$ where $o_n$ is a vector of 6 coefficients including real power (W), reactive power (var), RMS current (A), RMS voltage (V), frequency (Hz) and phase of voltage relative to current ($\varphi$). The database contains for each appliance two acquisition sessions of one hour. 100 appliances are recorded and spread uniformly into 10 categories: mobile phone chargers, coffee machines, computer workstations with monitor, fridges and freezers, Hi-Fi systems, lamp (CFL), laptops chargers, microwave ovens, printers, and televisions (LCD or LED).

Two protocols are proposed with the database: the *intersession* and *unseen instances* protocols. In the first protocol all the instances of the first session constitute the train set, whereas those of second session are used for testing. With this protocol, all the testing signatures come from appliances already seen in the training phase. At the time of writing this article, the best performances on the *intersession* protocol are reported in [11]. In this work, two classifiers are compared, namely k-NN and GMM systems showing respectively 88% and 93.8% correct category identification.

The second protocol aims at evaluating *unseen instances* configurations as illustrated in Figure 1. The goal is here to classify instances that are not seen

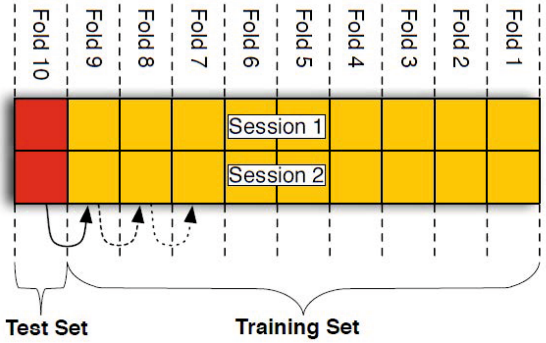


**Fig. 1.** *Unseen instances* protocol of the ACS-F1 database

before by the classifiers. This protocol also proposes to use a 10-cross fold procedure to smooth the evaluation results. The fold partitions are made available by the providers of the database. This protocol evaluates the system capability of generalizing to new brands or models of appliances.

# 4 System Description

## 4.1 Feature Extraction

In our procedure and as proposed in [10, 11], we use as baseline coefficients the raw observation $O$ as part of the features. We analyze here the impact of including information about the dynamics of the signal through the computation of the so-called *delta* and *delta-delta* or acceleration coefficients. These coefficients have been mainly used in speech recognition and have already been successfully used for appliance identification [11]. As explained in [13], the *delta coefficients* are computed with:

$$\Delta o_n = \sum_{w=-W}^{W} w \times o_{n-w} \tag{1}$$

where $K$ represents the window length. The value $W = 2$ has been retained after some tests, which corresponds to a window of 50 seconds. We then perform a z-normalization of the features, after which the mean is equal to zero and the variance is equal to one. The normalization is mainly useful for classifiers based on distance computation such as k-NN with a side effect of balancing each feature contribution. Our feature sequence $X = \{x_1, \ldots, x_N\}$ is therefore constituted of vectors composed of normalized observations and delta coefficients with $x_n = [c_{1n}, \ldots, c_{6n}, \Delta c_{1n}, \ldots, \Delta c_{6n}]$ and with $c_{in}$ the normalized value of the corresponding $o_{ik}$ observation. In a similar way, we also analyzed the extension of the features including the *acceleration* coefficients that are computed from the *delta* coefficients with:

$$\Delta\Delta o_n = \Delta o_{n+1} - \Delta o_{n-1} \tag{2}$$

In this work, we also evaluate the effect of applying power thresholding, eliminating from the sequence the observations where the value of the active power is below a given threshold $T_P$. Intuitively, this is related to the fact that appliances are difficult to discriminate when they are consuming a small quantity of energy, e.g. when they are off or in stand-by. After some pre-tests, the threshold $T_P$ is set to $0.5W$.

### 4.2 Classification

Two machine learning algorithms are analyzed in this work: k-NN and Gaussian Mixture Models (GMM). A k-NN classifier computes the $k$ closest features from the train set and then uses the labels of these features to perform the classification. In our case, we choose the winning class through a simple majority voting on the labels. In case of a tie, the class having the closest points is elected as winner. The normalized observation, *delta* and *acceleration* coefficients are representing different type of information. We then propose here to weight the distance computation with

$$dist(x_{ts}, x_{tr}) = \alpha \times d(c_{ts}, c_{tr}) + (1 - \alpha) \times d(\Delta c_{ts}, \Delta c_{tr}) \tag{3}$$

where $d$ is the euclidean distance, $x_{ts}$ a test feature vector and $x_{tr}$ a train feature vector. The coefficient $\alpha$ is tuned between 0 and 1 to give more or less weight to the *delta* versus the plain coefficients.

A GMM is a parametric probability density function estimating the likelihood $p(x_n|M_j)$ of a feature vector $x_n$ given a category $M_j$ as a weighted sum of Gaussian component densities. The model can be configured with the number of mixtures $I$. In our configuration, we used GMM with diagonal covariance matrices making the hypothesis of uncorrelated coefficient. This hypothesis is not true in practice but allows to reduce the number of parameters to estimate and to speed up the computations. The model is computed using the classical Expectation-Maximization (EM) algorithm [14]. The initial values of the Gaussian distributions are computed using the k-means algorithm. For testing, the likelihood $p(X|M_j)$ of an observation sequence $X$ given a model $M_j$ is computed by multiplying the local likelihoods $p(x_n|M_j)$ by making the observation independence hypothesis.

## 5 Result and Discussion

**Influence of the Delta Coefficients.** We observe that the inclusion of *delta* coefficients is beneficial for both k-NN and GMM models. Accuracy rates increase from 45% to 52.5% for the k-NN model when including the deltas. Similarly, accuracy rates increase from 62% to 69% for the GMM when including the deltas[1]. The dynamic information is bringing significant improvement to both systems.

[1] An optimization of $k$, the number of neighbors and $I$ the number of Gaussian is systematically performed in all reported results.

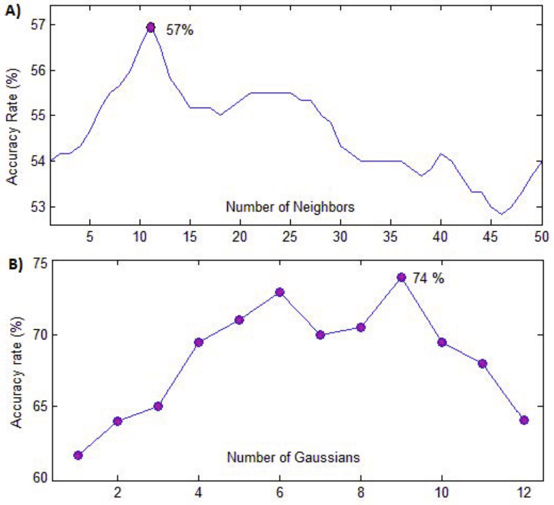


**Fig. 2.** Accuracy rate trend for a) k-NN varying the number of neighbors ($\alpha = 0.1$, with thresholding), B) GMMs varying the number of Gaussians (using delta and delta-delta coefficients, without thresholding)

**Influence of the Thresholding.** We observe that eliminating the feature vectors that show an active power below the threshold $T_P$ is beneficial for the k-NN system. Accuracy rates increase from 52.5% to 54.5% when applying the thresholding. This can be intuitively explained considering that close-to-zero power features are present in most signatures, corresponding to stret-ches of time where the appliances are not used. The training features corresponding to these stretches are independent to the categories and lead to noisy neighbors in the k-NN procedure. Also, as expected, we do not observe a benefit of the thresholding for the GMM models where the zero power stretches bring equivalent score contributions in all categories.

**Influence of Weighted Distance Computation.** We observe the benefit of applying a weighted distance computation as explained in Eq. 3. The performance improved from 54.5% to 57% with the k-NN system using thresholding and a value of $\alpha = 0.1$. As illustrated in the top part of Figure 2, we obtain this performance for an optimal value of $k = 11$.

**Influence of the Delta-Delta Coefficients.** Including further the acceleration coefficients, we could achieve an improvement of the GMM system from 69% to 74%. As illustrated on the bottom part of Figure 2, we also observe the effect of tuning the number of mixtures $I$, with the best performance obtained with $I = 9$ mixtures in the model. A slight improvement of 1.5% is also observed for the k-NN system by including the delta-delta coefficients.

Table 1 provides more details with the confusion matrix for our best GMM system. The categories printer, hifi and lamp are showing the largest error rates. Categories fridge and battery charger are showing the best performances.

**Table 1.** GMM Confusion Matrix with $I = 9$, without thresholding, using delta and delta-delta coefficients

| | Hifi | Television | Battery C. | Coffee M. | Computer | Fridge | Lamp | Laptop | Microwave | Printer |
|---|---|---|---|---|---|---|---|---|---|---|
| Hifi | .6 | .05 | 0 | 0 | 0 | .15 | .2 | 0 | 0 | 0 |
| Television | 0 | .7 | 0 | 0 | .05 | .05 | .05 | .05 | 0 | .1 |
| Battery C. | 0 | 0 | .9 | 0 | 0 | .05 | .05 | 0 | 0 | 0 |
| Coffee M. | 0 | 0 | 0 | .75 | 0 | 0 | 0 | 0 | .25 | 0 |
| Computer | 0 | .15 | 0 | 0 | .7 | 0 | 0 | .15 | 0 | 0 |
| Fridge | 0 | 0 | 0 | 0 | 0 | .9 | 0 | .05 | 0 | .05 |
| Lamp | 0 | .1 | .1 | 0 | .15 | .05 | .55 | 0 | .05 | 0 |
| Laptop | 0 | .05 | 0 | 0 | 0 | 0 | .05 | .85 | 0 | .05 |
| Microwave | 0 | 0 | 0 | .15 | 0 | 0 | 0 | 0 | .85 | 0 |
| Printer | .05 | .05 | 0 | 0 | 0 | 0 | .3 | 0 | 0 | .6 |

## 6 Conclusions

A first objective of this paper was to evaluate the feasibility of equipment identification using simple machine learning algorithms fed by low-frequency electricity consumption measurements. The answer to this question seems positive. We analyzed the performance of different algorithms for the task of identifying unseen appliance. A large database of electrical signatures was used with a total of 200 appliances. Our first conclusion is about the complexity of recognizing *unseen* appliances. When going from a *seen* appliance protocol to an *unseen* appliance protocol using the same database, we observe a drop of performance from 93.8% to 74% correct classification using the best GMM system for both protocols. The *unseen* task still shows acceptable performance but is much more difficult. Improving the performance could probably be reached by increasing the training data set which is still limited in the case of the experiments carried on here. A second conclusion is about the benefit to include dynamic coefficient that are, in our proposal, computed through simple delta and delta-delta coefficients. A third conclusion is about the tuning of some parameters including the weight $\alpha$ used to emphasize the information brought by the delta coefficient in k-NN systems and the number of Gaussians in the GMM model. Finally, as observed in previous works, we can also conclude on the superiority of the GMM over k-NN for signature modeling. Overall, our best accuracy has been raised up to 74% obtained with a GMM model using 9 Gaussians.

As future work, we plan to evaluate the use of state-based models such as HMMs, which should be particularly suitable for electrical signatures that intrinsically show a state nature. HMMs can also be seen as a generalization of GMMs. Comparison with discriminant approaches such as SVM and ANN will also be analyzed.

**Acknowledgment.** This work was supported by the research grant of the Hasler Foundation project *Green-Mod*, by the HES-SO and by the University of Fribourg in Switzerland.

## References

1. Chetty, M., Tran, D., Grinter, R.E.: Getting to green: understanding resource consumption in the home. In: Proc. UbiComp 2008, pp. 242–251 (2008)
2. Rahimi, S., Chan, A.D.C., Goubran, R.A.: Usage monitoring of electrical devices in a smart home. In: Proc. IEEE EMBS 2011 (2011)
3. U.S. Household Electricity Report, E.I.A. (2005), `http://www.eia.doe.gov/emeu/reps/enduse/er01_us.html`
4. Hart, G.W.: Nonintrusive appliance load monitoring. Proceedings of the IEEE 80, 1870–1891 (1992)
5. Clement, K., Pardon, I., Driesen, J.: Standby power consumption in belgium. In: Proc. EPQU 2007 (2007)
6. Guan, L., Berrill, T., Brown, R.J.: Measurement of standby power for selected electrical appliances in australia. Energy and Buildings 43, 485–490 (2011)
7. Zoha, A., Gluhak, A., Imran, M.A., Rajasegarar, S.: Non-intrusive load monitoring approaches for disaggregated energy sensing: A survey. Sensors, 16838–16866 (2012)
8. Adeel Abbas Zaidi, F.K., Palensky, P.: Load recognition for automated demand response in microgrids. In: Proc. IECON 2010 (2010)
9. Reinhardt, A., Baumann, P., Burgstahler, D., Hollick, M., Chonov, H., Werner, M., Steinmetz, R.: On the accuracy of appliance identification based on distributed load metering data. In: Proc. SustainIT 2012 (2012)
10. Zufferey, D., Gisler, C., Khaled, O.A., Hennebert, J.: Machine learning approaches for electric appliance classification. In: Proc. ISSPA 2012 (2012)
11. Ridi, A., Gisler, C., Hennebert, J.: Automatic identification of electrical appliances using smart plugs. In: Proc. Wosspa 2013 (2013)
12. Gisler, C., Ridi, A., Zufferey, D., Khaled, O.A., Hennebert, J.: Appliance consumption signature database and recognition test protocols. In: Proc. Wosspa 2013 (2013)
13. Hennebert, J.: Hidden Markov models and artificial neural networks for speech and speaker recognition. PhD thesis (1998)
14. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of Royal Statistical Society 39, 1–38 (1977)

# Multi-step-ahead, Short-Term Prediction of Wind Speed Using a Fusion Approach

Julian L. Cardenas-Barrera[1], Eduardo Castillo-Guerra[2],
Julian Meng[2], and Liuchen Chang[2]

[1] Center for Studies on Electronics and Information Technologies,
Universidad Central "Marta Abreu" de Las Villas. Santa Clara. V.C. Cuba.
[2] Department of Electrical and Computing Engineering. University of New Brunswick
Fredericton. N.B. Canada.
julian@uclv.edu.cu, {ecastil,jmeng,lchang}@unb.ca

**Abstract.** Wind power generation is a green solution to power generation that is receiving increasing interest worldwide. Wind speed forecasting is critical for this technology to succeed and remains today as a challenge to the research community. This paper presents a neural network fusion approach to multi-step-ahead, short-term forecasting of wind speed time-series. Wind speed forecasts are generated using a bank of neural networks that combine predictions from three different forecasters. The wind speed forecasters include a naïve model; a physical model and a custom designed artificial neural network model. Data used in the experiments are telemetric measurements of weather variables from wind farms in Eastern Canada, covering the period from November 2011 to October 2012. Our results show that the combination of three different forecasters leads to substantial performance improvements over recommended reference models.

**Keywords:** Short-term wind speed forecasting, artificial neural networks, forecast combination.

## 1 Introduction

Wind power generation is a green and cost-effective solution to power generation, which has grown substantially worldwide. It is among the most competitive renewable technologies, projected to account for significant shares of the global power market in the near future [1]. However, the inherent variability and uncertainty of wind can lead to unexpected and sometimes substantial mismatches between scheduled and actual wind power [2]. This fact has driven the search for improved wind power forecasting methods [1–8].

Forecasting means that future values of wind power or speed $a(t+k)$ will be estimated as $\hat{a}(t+k|t)$ by some function $\varphi$, given some predicting variables $X(t)$ up to time $t$. A forecasting error $\varepsilon(t+k|t)$ is introduced and expected to be bounded. The target time $(t+k)$ is called the forecasting horizon.

$$\hat{a}(t+k|t) = \varphi(X(t)) \tag{1}$$

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 83–91, 2013.

$$a(t+k) = \hat{a}(t+k|t) + \varepsilon(t+k|t), \tag{2}$$

Forecasting horizons for wind power integration and operation span time frames ranging from a few seconds to weeks ahead [8]. Accurate short-term forecasts are critical for effective power management and receive considerable research attention [1, 7, 9–13]. Forecasting methods typically work better when the prediction is focused on specific time frames. Complex physical models for example, use numeric weather prediction (NWP) and are known to have good performance for horizons beyond 3-6 hours ahead [10]. The simplest model called persistence offers accurate wind speed forecasts for immediate horizons, typically up to around one hour ahead but also depending on weather conditions, can be reliable for up to 6 hours [9, 10]. Statistical and machine learning approaches on the other hand, can generate reasonable forecasts for less than a day horizon [2, 8].

Forecast combination offers a way to reduce forecasting errors by combining the predictions from a number of forecasters [14–16]. Prediction gains are obtained by exploiting diversification if individual forecasters use different models of the underlying process they predict [17]. Combination success depends on how well mixing weights can be assigned to individual forecasters. This method is often preferred over finding one single best model. For wind speed forecasting, statistical and machine learning models are known to automatically reduce systematical errors due to their adaptation to the location of the wind farm; however as opposed to NWP, they have difficulties predicting rare atmospheric conditions [6, 8, 11]. Combining NWP based, statistical and machine learning models could help at getting the most of all methods.

This paper exploits this approach proposing a fusion method for multi-step-ahead, short-term (up to 6 hours ahead) prediction of wind speed. Neural networks (NN) are used to determine the best nonlinear combination of three predictors at different time leads. The predicting models include persistence, artificial neural network and NWP-based predictors. Performance is assessed using recommended error measures and reference models [1, 9, 10].

## 2 Case Study and Data Description

The data used in this study come from one-year-long recordings of several measured weather variables at four on-shore wind farms in Eastern Canada. The wind farms are named herein as WindFarm-1, WindFarm-2, WindFarm-3 and WindFarm-4. The observed variables include temperature, wind speed, and wind direction measured at the hub heights (80 meters), every five minutes. Table 1 shows mean wind speeds,

**Table 1.** Wind characteristics during the period under study

| Wind farm | Mean Wind Speed (scale, shape of Weibull dist.) | Wind Direction |
|---|---|---|
| **WindFarm-1** | 8.35 m/s (9.43m/s, 2.38) | 250° (std=78.5 °) |
| **WindFarm-2** | 8 m/s (9.01m/s, 2.25) | 210° (std=100 °) |
| **WindFarm-3** | 7.6 m/s (8.54m/s, 2.57) | 220° (std=80 °) |
| **WindFarm-4** | 7.6 m/s (8.60m/s, 2.48) | 202° (std=98 °) |

shape and scale parameters of the fitted Weibull distribution and, mean and standard deviation of wind directions of each wind farm. A total of 366 days (105408 samples), from November 1st 2011 until October 31st 2012, conform the recording period.

NWP-based forecasts from Environment Canada (EC) were available for all sites, offering wind speed estimates at heights H = {10, 40, 120, 216} m. EC delivers a set of forecasts values four times a day. The sets include forecasts for the next 3 to 48 hours with a 15 minutes time step. Linear interpolation is used to estimate wind speed at the hub height (80m).

WindFarm-1 is presented in this paper to demonstrate the effectiveness of the proposed technique. Fig. 1 shows plots of observed values and monthly averages of wind speed along with a windrose diagram of wind directions. It is worth noting that prevailing and strong winds mainly blow from West and North-West with slower winds occurring in the summer. Noon and nights exhibit slower wind speeds than early morning and evening, which is reported to affect the accuracy of some NWP based predictors [18]. These elements will help selecting input features for one of the forecasting techniques to be used in the proposed fusion.

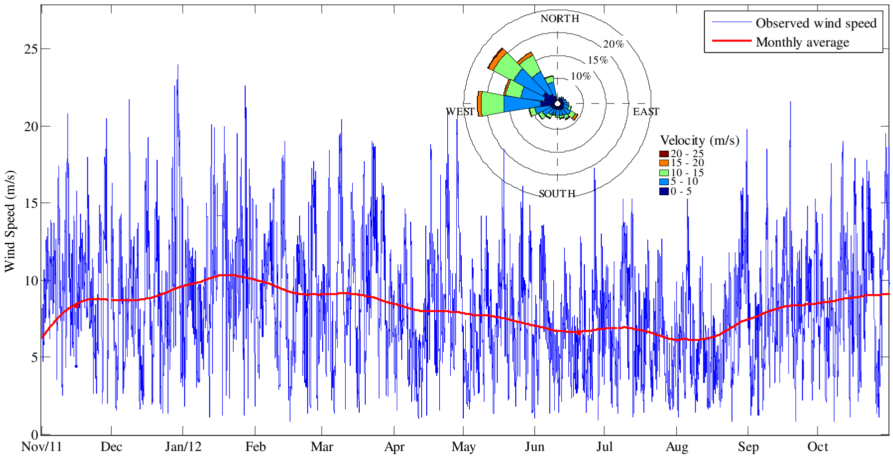


**Fig. 1.** Measured wind speeds and main wind directions at WindFarm-1

# 3 Performance Evaluation

An effective comparison of forecasting methods is not easily achieved as there are various performance criteria used by researchers [1, 9]. Recently, partners of the ANEMOS project have proposed a set of recommendations to establish a common ground for comparison [9, 10]. Some suggested error criteria include:

– mean absolute error (MAE),

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{a}_i - a_i| \tag{3}$$

— root-mean-squared error (RMSE),

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{a}_i - a_i)^2} \tag{4}$$

— mean absolute percentage error (MAPE).

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\frac{|\hat{a}_i - a_i|}{|a_i|} \tag{5}$$

N is the size of the sample used to calculate the measure, and $\hat{a}_i = \hat{a}_i(t + k|t)$ and $a_i = a_i(t + k)$ are the forecasted and the actual values respectively. MAE and RMSE express global errors in terms of wind speed; but RMSE accentuates the effect of larger errors. MAPE is a dimensionless, relative measure of global error. Its singularity problem is avoided here by ensuring the exclusion of zero-valued speeds, which have a very low probability of occurrence.

The recommendations also emphasize to build the models and validate them by using cross-validation. We reserved a set of data (15%) to test model predictions. Training (70%) and validation (15%) sets were used to create models and to avoid model to adapt too closely to the training sample (overfitting), respectively. Data for each set was randomly chosen in order to have training samples from any time of the year.

Finally, comparing farms with differently variable time series should use a skill score $\mathrm{SS}_{\gamma}^{\mathrm{ref}}(\mathrm{k})$. This is an objective indicator of the improvement a forecasting model has over a reference (ref) model at horizon k, using some metric γ [2].

$$SS_{\gamma}^{ref}(k) = \frac{\gamma^{ref}(k) - \gamma(k)}{\gamma^{ref}(k)} \times 100\% \tag{6}$$

# 4 Individual Wind Speed Forecast Techniques

## 4.1 Persistence or Naïve Model

This is a simplistic model where its k-step-ahead forecast is expressed as:

$$\hat{a}(t + k|t) = a(t), \tag{7}$$

i.e. future predicted values will be the same as the last observed value. Its accuracy is very high at the shortest horizons and cannot be easily outperformed by other, more complex models. For this reason it is almost universally used as a benchmark for short-term forecasting of less than six hours. Using persistence in combination with other models ensures high precision and simplicity at the shortest time leads.

## 4.2 The Artificial Neural Network Approach

Artificial neural networks are amongst the most successful machine learning approaches to short-term wind speed time-series forecasting. Improvement over other machine learning and statistical models has been extensively reported in the literature. Although several network architectures have been proposed in forecasting applications, the feed-forward network is one of the most popular [10–13].

In this study, we used a feed-forward multilayer perceptron (MLP) neural network architecture using the Levenberg-Marquardt backpropagation (LM) training algorithm. The selected training algorithm encompasses the Newton and gradient descent methods being one of the fastest algorithms. Mean squared errors (MSE) are minimized, which is appropriate for wind speed forecasting where prediction errors are assumed to have a Gaussian distribution [6]. For a multi-step-ahead prediction several approaches might be followed namely iterative, multi-model, or single-model-multivariate forecasting. Iterative forecasting predicts successive look-ahead times by aggregating previous forecasts to the input series. This solution might lead to decreasing accuracy as look-ahead time increases, due to accumulation of errors [12]. Single-model-multivariate forecasting uses output layers with as many neurons as the number of look-ahead times, leading to very complex systems. Finally we chose the multi-model forecasting. This approach builds a set of models for each time step; the individual networks are small, faster to train and less likely to be overfitted. A single-neuron linear output layer at each k$^{th}$ NN outputs $\hat{a}_i(t+k|t)$ wind speed estimates.

Exploratory analysis of the measured weather variables and a bootstrap aggregation of regression trees helped selecting the set of input features by providing measures of variable importance [19]. A final selection of eleven features includes lagged values of wind speed, wind direction, and temperature averages over the last three days prior to the forecasting date and the hour of the day. Having chosen the number of inputs, we addressed the problem of identifying an appropriate network topology. Two hidden layers with hyperbolic tangent sigmoid transfer functions with 20 neurons each resulted as the best configuration after running a set of trials. Fig. 2 depicts a plot of the global MAPE values obtained with these NNs.

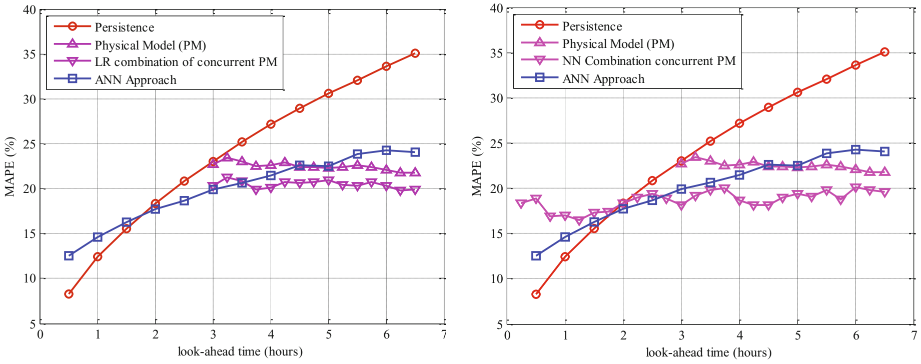


**Fig. 2.** Performance of individual forecasters

### 4.3 Forecast from a Physical Model Based on NWP

In order to include NWP-based forecasts as an aggregated approach in the final forecast, we decided to combine concurrent sets of EC's forecasts. EC's multi-step-ahead forecasts F(T) are delivered every D=6 hours, generating an overlap between current

F(T) and previous F(T-nD) forecasts. These overlaps can be used to extend the time span of the forecasts below 3 hours and to further improve the predictor's performance; particularly if concurrent forecasts can be properly combined.

When concurrent forecasts $\hat{a}_i$ are comparable, the combination can be reduced to simple average of competing forecasts. More complex combinations can also be exploited using more complex mixing functions. We tested several of these methods including simple averaging of concurrent forecasts, linear regression (LR), and NN. Simple averaging assigns equal weights to all, present and past, forecasts and does not eliminate any bias of the original forecasts. Regression and NN based approaches assign weights that act as "forgetting" factors indicating the influence of each previous forecast in the combined forecast. Fig. 2 graphically shows the improvement attained by linear regression and NN-based combinations. This figure also shows average MAPE performance for the other two independent forecasters. Individual models do not resemble competing models' forecasts at all k-step-ahead times. As expected, persistence performs best for look-ahead times below two hours. The relative performance between the NN approach and the combination of concurrent NWP-based forecasts depends on the combination approach employed.

## 5 Fusion Forecast Method

Independent forecasters are combined to generate a final forecast. The proposed combination should generate multi-step-ahead, short-term forecasts from five minutes, up to 6:00 hours ahead. The time step between time leads is set to five minutes. As aforesaid, although independent forecasters can be further improved, we decided to assign the task of performance improvement to the combination method. For this reason we selected simple averaging of concurrent NWP-based forecasts over NN or LR approaches as an independent model. This decision makes the overall system simpler.

Artificial neural networks have been successfully applied for forecast combination [14–16]. In this paper we investigated a NN architecture that optimizes the assigned

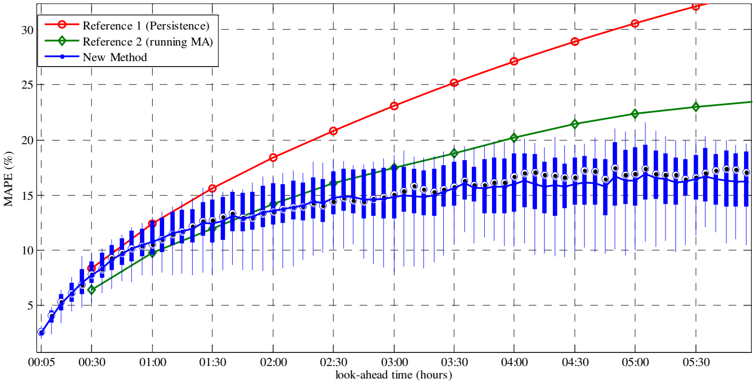


**Fig. 3.** Performance of the proposed hybrid method

weights to each forecaster at each time lead. Again, we decided to build a bank of simple neural networks, corresponding with our look-ahead times. Each network was designed as feed-forward with one hidden layer of 10 neurons and a hyperbolic tangent sigmoid transfer function. The output layer had only one neuron and a linear transfer function to produce one-step-ahead forecasts. The LM algorithm was used for training with the MSE distortion measure as the performance function.

Fig. 3 illustrates the monthly MAPEs for every prediction horizon and mean MAPEs are available for two reference models, namely persistence (Pers) and the moving average (MA) reference recommended by several authors [9, 10]. Box plots are also displayed for the proposed fusion method. From 10 minutes ahead, the proposed method is increasingly better than persistence. Similar behavior stands when comparing with the MA reference after 1:45 hours. The box plots show increment of error dispersion with increased horizons. The method also shows consistency after three hours ahead as error dispersion does not increase.

Table 2 shows some performance metric values at different horizons. It also shows how the proposed method holds significant gains over reference models as look-ahead times increase. Skill scores (SS) calculated using different error criteria show from 35% to more than 50% improvement over persistence, and from 16% to above 30% improvement over the recommended MA reference.

**Table 2.** Performance metrics and skill scores at different horizons

| *Horizon* | *MAE* | *$SS_{MAE}$ (%)* | | *RMSE* | *$SS_{RMSE}$ (%)* | | *MAPE* | *$SS_{MAPE}$ (%)* | |
|---|---|---|---|---|---|---|---|---|---|
| | *(m/s)* | **Pers** | **MA** | *(m/s)* | **Pers** | **MA** | *(%)* | **Pers** | **MA** |
| **0:30** | 0.543 | 9.46 | ***-8.31*** | 0.736 | 11.09 | ***-7.76*** | 7.75 | 6.89 | ***-9.47*** |
| **1:00** | 0.738 | 15.86 | ***-8.27*** | 0.978 | 17.58 | ***-6.73*** | 10.78 | 13.05 | ***-10.08*** |
| **2:00** | 0.915 | 28.35 | 8.08 | 1.200 | 29.53 | 8.44 | 13.61 | 25.90 | 6.20 |
| **3:00** | 0.984 | 38.86 | 17.91 | 1.266 | 40.50 | 19.30 | 14.79 | 35.87 | 15.59 |
| **4:00** | 1.063 | 42.95 | 21.56 | 1.356 | 44.45 | 23.13 | 15.99 | 41.09 | 21.43 |
| **5:00** | 1.077 | 48.37 | 28.02 | 1.381 | 49.20 | 28.66 | 16.31 | 46.61 | 28.61 |
| **6:00** | 1.082 | 52.75 | 31.25 | 1.386 | 53.54 | 32.05 | 16.24 | 51.69 | 32.61 |

## 6 Concluding Remarks

Wind speed forecasting is today an important research topic for a continued increase of wind power penetration into the global power market. In this paper, we presented an effective fusion-based wind speed prediction method that non-linearly combines the results of three different forecasters. Experimental results show that a Neural Network combination of forecasts improves performance over individual methods used in the combination. Error reductions up to more than 50% with respect to persistence and more than 30 % over the recommended MA reference predictors are obtained with different error measures. Even for horizons where only one independent method prevails over the rest, the fusion approach improves performance.

Although not explicitly shown in the paper, the application of this methodology to the other sites listed in Table 1 gave similar results.

We agree with other authors [13] in that better models would be obtained increasing the dataset to more than one year. Adaptation of current models as more data become available is another alternative for improving models.

Results from this paper can be extended in a number of directions. First, improved independent models could be used. This has the advantage of enabling the fusion approach to offer a rather good prediction when the other methods fail. Secondly, the adaptation to other wind farms should also be investigated without explicitly training the combination forecast to each wind farm. Finally, including probabilistic error bounds instead of point forecasting should also be investigated.

## References

1. Zhu, X., Genton, M.G.: Short-Term Wind Speed Forecasting for Power System Operations. International Statistical Review 80, 2–23 (2012)
2. Foley, A.M., Leahy, P.G., Marvuglia, A., McKeogh, E.J.: Current methods and advances in forecasting of wind power generation. Renewable Energy 37, 1–8 (2012)
3. Stathopoulos, C., Kaperoni, A., Galanis, G., Kallos, G.: Wind power prediction based on numerical and statistical models. Journal of Wind Engineering and Industrial Aerodynamics 112, 25–38 (2013)
4. Wang, X., Guo, P., Huang, X.: A Review of Wind Power Forecasting Models. Energy Procedia 12, 770–778 (2011)
5. Soman, S.S., Zareipour, H., Malik, O., Mandal, P.: A review of wind power and wind speed forecasting methods with different time horizons. In: North American Power Symposium (NAPS), pp. 1–8 (2010)
6. Lange, M., Focken, U.: New developments in wind energy forecasting. In: 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, pp. 1–8 (2008)
7. Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., Feitosa, E.: A review on the young history of the wind power short-term prediction. Renewable and Sustainable Energy Reviews 12, 1725–1744 (2008)
8. Ernst, B., Oakleaf, B., Ahlstrom, M.L., Lange, M., Moehrlen, C., Lange, B., Focken, U., Rohrig, K.: Predicting the Wind. IEEE Power and Energy Magazine 5, 78–89 (2007)
9. Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H.A., Nielsen, T.: Standardizing the Performance Evaluation of ShortTerm Wind Power Prediction Models. Wind Engineering 29, 475–489 (2005)
10. Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., Draxl, C.: The state-of-the-art in short-term prediction of wind power: A literature overview. ANEMOS. plus (2011)
11. Salcedo-Sanz, S., Pérez-Bellido, Á.M., Ortiz-García, E.G., Portilla-Figueras, A., Prieto, L., Correoso, F.: Accurate short-term wind speed prediction by exploiting diversity in input data using banks of artificial neural networks. Neurocomputing 72, 1336–1341 (2009)
12. Kusiak, A., Zheng, H., Song, Z.: Short-term prediction of wind farm power: a data mining approach. IEEE Transactions on Energy Conversion 24, 125–136 (2009)
13. Li, G., Shi, J.: On comparing three artificial neural networks for wind speed forecasting. Applied Energy 87, 2313–2320 (2010)
14. Donaldson, R.G., Kamstra, M.: Forecast combining with neural networks. Journal of Forecasting 15, 49–61 (1996)

15. Palit, A.K., Popovic, D.: Nonlinear combination of forecasts using artificial neural network, fuzzy logic and neuro-fuzzy approaches. In: The Ninth IEEE International Conference on Fuzzy Systems, FUZZ IEEE 2000, vol. 2, pp. 566–571 (2000)
16. Aladag, C.H., Egrioglu, E., Yolcu, U.: Forecast Combination by Using Artificial Neural Networks. Neural Process. Lett. 32, 9156:269–9156:276 (2010)
17. Timmermann, A.: Chapter 4 Forecast Combinations. In: Elliott, G., Granger, C.W.J., Timmermann, A. (eds.) Handbook of Economic Forecasting, pp. 135–196. Elsevier (2006)
18. Lange, M.: On the Uncertainty of Wind Power Predictions—Analysis of the Forecast Accuracy and Statistical Distribution of Errors. Journal of Solar Energy Engineering 127, 177 (2005)
19. Strobl, C., Malley, J., Tutz, G.: An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychol. Methods 14, 323–348 (2009)

# Green Coverage Detection on Sub-orbital Plantation Images Using Anomaly Detection

Gabriel B.P. Costa and Moacir Ponti

Instituto de Ciências Matemáticas e de Computação — Universidade de São Paulo
13566-590 São Carlos, SP, Brazil
{gpbcosta,moacir}@icmc.usp.br
http://www.icmc.usp.br/~moacir

**Abstract.** The green coverage region is a relevant information to be extracted from remote sensing agriculture images. Automatic methods based on threshold and vegetation indices are often applied to address this task. However, sub-orbital remote sensing images have elements that can hinder the automatic analysis. Also, supervised methods can suffer from imbalance since there is often many more green coverage samples available than regions of gaps, weed and degraded areas. We propose an anomaly detection approach to deal with these challenges. Parametric anomaly detection methods using the normal distribution were used and compared with vegetation indices, unsupervised and supervised learning methods. The results showed that anomaly detection algorithms can handle better the green coverage detection. The proposed methods showed similar or better accuracy when compared with the competing methods. It deals well with different images and with the imbalance problem, confirming the practical application of the approach.

**Keywords:** Anomaly, outlier, remote sensing.

## 1 Introduction

Precision agriculture can help small farmers in the management of plantations. One of the most important technologies in this context is remote sensing imagery. However satellite remote sensing can be expensive, while low-cost systems that acquire sub-orbital images can benefit developing countries and small properties[11].

A low-cost remote sensing system was proposed by Martins et al. [7] based on an image acquisition equipment attached to a balloon. This system acquires sub-orbital images that can be transmitted via radio frequency or processed offline. The advantages of this method includes the height control (often from 10 to 100 meters), the need of one or two persons to operate, and the low cost. The disadvantages are the limitation in regions with trees and electric wires, and a low load capability (from 2 to 4 kg).

One of the most relevant information to be extracted from the image is the green coverage region. By accessing a map of green coverage it is possible to

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 92–99, 2013.

locally adjust irrigation, application of fertilizers, and perform better weed control. To address this task, previous studies includes method based on threshold Otsu's method, histograms and vegetation indices such as ExG (excess green) [4] among others. A combination of vegetation indices and mean-shift segmentation improved the previous results [9].

Sub-orbital images suffer from illumination variation, shadows and other elements that can hinder the automatic analysis. For this reason, when using tools of satellite remote sensing, it is often difficult to improve the results using only unsupervised methods such as those based on threshold and vegetation indices. Also, supervised methods can also not perform well since there is often many more green coverage samples available than regions of soil, weed, gaps and degraded areas. Besides, it can be a hard task to label many samples before using the system. In order to deal with these challenge, we propose an anomaly detection approach.

Anomalies (or outliers, exceptions or deviations) are patterns with an unexpected behavior. Barnett and Lewis [1] defined anomaly as an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. Due to the nature of the problem, anomalies are often rare and dealing with it can help on applications such as fault detection, fraud detection, network intrusion, etc. An anomaly detection (AD) method take as input a sample or set of samples, and identify whether those samples are "normal" or "abnormal", according to what is expected to be found. On most applications the data is imbalanced, "normal" samples are widely available, while anomalies are scarce or not available [2].

The motivation to the application is that this approach needs mostly samples from normal data, that are abundant and easy to label, and few examples (and sometimes no examples) from anomalous data. We also organized a dataset based on sub-orbital images, available for download. Our contribution is to look at the green coverage detection as an anomaly detection process, so that green coverage will be considered normal behavior, while gaps, soil, degraded areas and others will be considered to be abnormal.

## 2 Low-Cost Remote Sensing System

A system built with a helium gas baloon model Skyhook Helikite was used to acquire the images. A digital camera with a 10 megapixel CCD sensor of size (1/2.3)-in was attached to the balloon with a radiofrequency controller board. It was build to provide an inexpensive solution for remote sensing in Brazil [7] [9].

For this study, a total of 12 images of plantations were obtained with an approximate height of 50 meters, from two different fields of common beans at 63 days after the emergence of the plants, in different days. The images were cropped to squared parcels, and resampled to $1024 \times 1024$ pixels, resulting in an approximate resolution of 3.1cm/pixel.

The original images were acquired in RGB color model. Figure 1 shows versions of six images used in the experiments, converted to grayscale. The difference between the two crops was the soil compaction, the second row of images were obtained from the crop with higher soil compaction.

Due to the different weather conditions, there are images with different contrast and bright characteristics, and some of the images have motion blur due to the balloon movement.

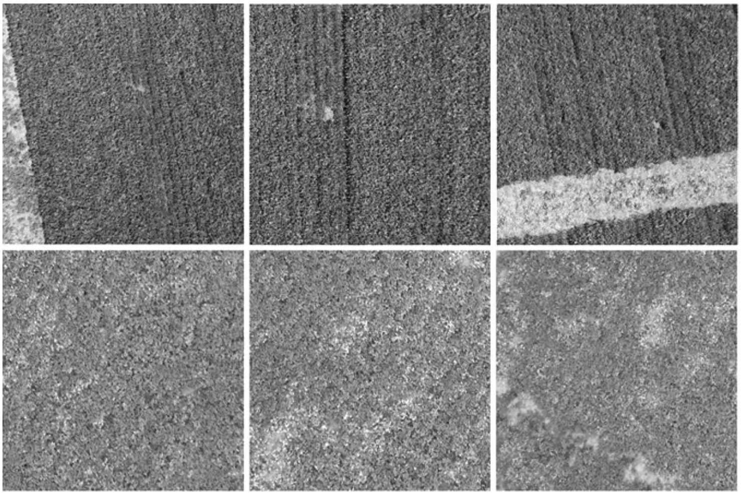


**Fig. 1.** Examples of images obtained from two different crops of beans (first and second row of images with different soil compaction) using a low-cost remote sensing system

### 2.1 Feature Extraction

In order to use the machine learning and anomaly detection methods, it is necessary to extract features from the image in order to build a feature vector. We selected a texture and a region-based color extractor.

*Haralick Texture Features* : after converting the image to a grayscale version using the composition $I = 0.2989 \cdot R + 0.587 \cdot G + 0.114 \cdot B$, the texture features were computed using 6 Haralick features [5] with a $(0, 1)$ co-occurrence matrix: entropy, maximum probability, homogeneity, uniformity, contrast and correlation.

*CCV Color Features* : the Color Coherence Vector method tries do codify how colors are organized in connected regions. It classifies each pixel as coherent or incoherent based on whether or not it is part of a large similarly-colored region [8]. The RGB image was quantized to 64 colors and a threshold of 25 was used to compute the CCV features.

## 3 Green Coverage Detection Methods

### 3.1 Vegetation Index

Vegetation index techniques uses arithmetic operations on the available bands (visible light, near-infrared, etc.). The aim is to to enhance some features, obtaining an image in which, for example, it is possible to visualize better the vegetation, with a better contrast between the response models in the available channels. These indices are often used in order to segment the green vegetation regions in agriculture remote sensing images. One of the most used ones, when only the visible light is available is the $ExG$, computed using $ExG = 2G - R - B$. After computing the index, a threshold method such as Otsu method is used to separate green coverage from other areas in the image, creating a binary image [9]. The user must interpret the results since the images can have zero or one values both for green coverage and without green coverage regions.

### 3.2 Unsupervised and Supervised Learning Methods

Any machine learning method can be used to detect regions in remote sensing images. Unsupervised learning methods can separate pixels or sub-images in groups by using distances between them. In this case there is no previous knowledge involved, and the user must interpret the results given the output. Supervised learning methods are able to build a model for each class, e.g. green coverage and lack of green coverage. For this reason, it is important to have enough labeled data so that all every model is well built.

In this study we use classic algorithms such as the $k$-Means, unsupervised method that minimizes the squared error with respect to samples and cluster centroids, and the Normal Bayes, a supervised probabilistic algorithm that assumes the data is normally distributed, but does not assumes independent features.

We also investigated the Optimum-Path Forest classifier, a classifier based on graph theory, since it obtained good results on imbalanced datasets [10].

### 3.3 Anomaly Detection

In this paper we used methods that models only the normal data, using few abnormal samples in order to obtain a threshold for the detector. According to Hodge and Austing [6], the advantages of these methods are: a) needs mostly data labeled as normal and just a few labeled as abnormal, b) it is suitable for static or dynamic data, as it only learns one class, c) most method are incremental, d) it does not assume any distribution for the abnormal data.

Three methods are proposed to the problem of detecting green coverage regions: the normal univariate and multivariate anomaly detectors [1], and our algorithm, based on the concatenation of features and detection in a normal parameter space [3].

- **Normal univariate and multivariate detectors:** uses the normal probability density function in order to learn with the normal data available. It can use a univariate model, defined in Equation 1, or a multivariate model, defined in Equation 2, that outputs the likelihood of a sample $x$ belonging to the same law of the samples used to estimate the parameters of the distribution.
$$p(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1}$$
$$p(x, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \tag{2}$$
The methods comprise three steps:
  1. Estimate the normal distribution parameters: mean and standard deviation (univariate) or mean vector and covariance matrix (multivariate), using the data available, i.e. green coverage samples;
  2. Find a threshold of anomaly detection: uses samples (normal and abnormal) from a validation set in order to find a threshold $T$ for the likelihood $p$ that maximizes the accuracy value.
  3. Detection: compute its likelihood using the estimated parameters, if the value is lower than $T$ it is considered an anomaly.

- **Parameter space anomaly detector:** selects randomly from the training set $M$ pairs of samples. Concatenates all features of each pair of normal samples, and computes mean and standard deviation for the whole concatenated vector. Each concatenated pair is a point in a parameter space $\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$, forming a point cloud, from which a convex hull is computed. This convex hull captures the normal behaviour.
The algorithm tries to detect abnormal samples by concatenating them with normal samples and observing the deviation from the normal point cloud convex hull. The method comprises the following steps [3]:
  1. Select pairs of normal instances, e.g. $a$ and $b$, concatenate the features of $a$ and $b$, and estimate the parameters $\mu$ and $\sigma$ for each pair,
  2. Compute a convex hull $\mathcal{H}_N$ from the 2D point cloud.
  3. Find a threshold of anomaly detection: uses samples (normal and abnormal) from a validation set in order to find a threshold $P$ that maximizes the accuracy value for the perturbation caused by concatenating normal with abnormal samples, forming a new convex hull $\mathcal{H}_T$. The perturbation is the distance of the created point to all points in the convex hull computer in the previous step.
  4. Detection: concatenate each point that contributed to the convex hull $\mathcal{H}_N$ with the unknown pattern $x$. Estimate the parameters $\mu$ and $\sigma$ and compute a new convex hull $\mathcal{H}_T$. If the intersection of $\mathcal{H}_T$ and $\mathcal{H}_N$ is lower than $P$, consider it an anomaly.

  This method captures data similarly to the normal univariate method. However it has more potential to be incremental, since new samples can be added in the normal point cloud in constant time.

## 4 Experiments

All images were manually labeled by three agronomists. These specialists segmented the images in two disjunct regions: i) green coverage and ii) vegetation gaps, soil, degraded areas and others. The agreement between the specialists was of 91.7%±5.2. The images labeled by the agronomist with the higher inter-agreement was used as ground truth.

For the classic vegetation index methods, each image pixel was used to detect green coverage since these methods need the whole image to process. In the other hand, sub-images of 100 × 100 pixels, also labeled by the agronomists, are used as observations for the other methods. The use of sub-images is feasible because the resolution is high when compared with satellite images. This high resolution is possible because the images were acquired by a sub-orbital equipment at just 50 meters as described in section 2. The six Haralick descriptors and the CCV feature vector, described in section 2.1, were computed for each one of the 230 sub-images. The dataset anomaly rate, i.e., the proportion of not normal samples, is $\sim 9\%$. The parameters for the CCV methods were found experimentally, after testing on a separate validation set of 20 images.

The settings for each methods used to detect green coverage are:

- *Unsupervised methods*:
    - Excess Green (ExG) and Mean-shift with Excess Green (MS-ExG): computed in the whole image, using each pixel as observation;
    - $k$-Means: computed using each feature vector extracted from the sub-images as an observation.
- *Supervised learning methods*: computed using each feature vector extracted from the sub-images as an observation. Uses 70% of both normal and abnormal samples for training, and 30% for testing.
    - Normal Bayes and Optimum-path Forest (OPF).
- *Anomaly detection (AD) methods*: computed using each feature vector extracted from the sub-images as an observation. Uses 55% of normal samples for training, %15 of both normal and abnormal samples for validation, and 30% for testing.
    - Normal univariate, normal multivariate and parameter space AD.

### 4.1 Evaluation

We used a repeated random sub-sampling validation, each experiment was repeated 100 times. The average and standard deviation were computed by these repetitions. The evaluation was based on the balanced accuracy value that takes into account the balance between the classes:

$$\text{Acc} = 1 - \frac{\sum_{i=1}^{c}[e_{i,1} + e_{i,2}]}{2c}, \quad e_{i,1} = \frac{FP(i)}{N - N(i)}, \quad e_{i,2} = \frac{FN(i)}{N(i)}, i = 1, ..., c,$$

where $c$ is the number of classes, $e_{i,1} + e_{i,2}]$ is the partial error of the class $i$, $FN(i)$ (false negatives) is the number of samples belonging to $i$ incorrectly classified as belonging to other classes, and $FP(i)$ (false positives) the samples $j \neq i$ that were assigned to $i$ [9].

## 5 Results and Discussion

The average accuracies (in percentages) for each method are presented in Table 1. The anomaly detection methods showed accuracies similar or better than the best previously proposed methods. Threshold methods used the ExG index, while the learning methods used texture or color features. The results shows that texture features have better discriminative potential when compared to the color features for this application.

**Table 1.** Average accuracy and standard deviation for the investigated methods

| Threshold Methods | | |
|---|---|---|
| ExG | 76.5±8.1 | — |
| MS+ExG | 81.1±7.3 | — |
| Learning Methods | Haralick-8 | CCV-64 |
| *k*-Means | 66.0±9.0 | 59.7±4.7 |
| Normal Bayes | 68.7±9.5 | 62.2±10.2 |
| OPF | 60.7±3.7 | 64.3±13.0 |
| Parameter space AD | 79.1±9.1 | 69.5±9.5 |
| Normal Univariate AD | 77.9±8.9 | 68.7±9.1 |
| Normal Multivariate AD | 89.7±6.9 | 70.1±6.8 |

The unsupervised methods based on vegetation indices, including the recently published MS-ExG, performed well, with results comparable with the proposed methods: parameter space AD and normal univariate AD. However, it is important to note that the unsupervised results must be interpreted after the algorithm outputs the processed image, while the anomaly detection algorithms already have a meaningful output.

Due to the scarce anomaly data available, the supervised learning methods (classifiers) produced mediocre results. The clustering algorithm, that used feature vectors to produce the results, performed worst than those based on vegetation indices. It is probably because the ExG and MS-ExG methods used each pixel value as an observation, while the *k*-Means used the feature vector computed over the 100×100 pixel sub-images.

## 6 Conclusions

This paper reports results of an anomaly detection methods applied to the green coverage detection problem. The main reasons for the success of this strategy is that: it does not assume any given distribution of the abnormal data, and does not require much abnormal samples to be trained. Besides, this approach carries most advantages of partially supervised algorithms, such as the incremental capability, in which new samples can be easily added to the model.

Whilst the multivariate method obtained the best result, the other methods showed good potential in this application. Future works can explore variations of the proposed parameter space, including multiple parameters that can capture correlations, exploring the use of the anomalous data in the training step, and improving the feature fusion, presently carried out by concatenation.

The experimental evidence showed that the green coverage detection can be successfully treated as an anomaly detection problem, benefiting applications in precision agriculture that uses low-cost sub-orbital images.

**Acknowledgment.** This work was supported by FAPESP (grants n.11/16411-4 and n.12/12524-1).

## References

1. Barnett, V., Lewis, T.: Outliers in statistical data. John Wiley & Sons (1994)
2. Chandola, V., Banerjee, A., Kumar, A.: Anomaly detection: A survey. ACM Computing Surveys 41(3), 15 (2009)
3. Costa, G., Ponti, M., Frery, A.: Partially supervised anomaly detection using convex hulls on a 2D parameter space. In: Partially Supervised Learning. LNCS (LNAI), vol. 8183, pp. 1–8 (2013)
4. Gée, C., Bossu, J., Jones, G.: Truchetet: Crop/weed discrimination in perspective agronomic images. Comput. Electron. Agr. 62, 49–59 (2008)
5. Haralick, R., Shanmugan, K., Dinstein, I.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics SMC-3(6), 610–621 (1973)
6. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85–126 (2004)
7. Martins, F.C.M.: Evaluation of compact areas in the bean culture using remote sensing techniques. Master's thesis, UFV, MG, Brazil (2010) (in Portuguese)
8. Pass, G., Zabih, R., Miller, J.: Comparing images using color coherence vectors. ACM Multimedia 96, 65–73 (1996)
9. Ponti, M.P.: Segmentation of low-cost remote sensing images combining vegetation indices and mean shift. IEEE Geoscience and Remote Sensing Letters 10(1), 67–70 (2013)
10. Ponti Jr., M.P., Papa, J.P., Levada, A.L.M.: A Markov Random Field Model for Combining Optimum-Path Forest Classifiers Using Decision Graphs and Game Strategy Approach. In: San Martin, C., Kim, S.-W. (eds.) CIARP 2011. LNCS, vol. 7042, pp. 581–590. Springer, Heidelberg (2011)
11. Swain, K., Thomson, S., Jayasuriya, H.: Adoption of an unmanned helicopter for low-altitude remote sensing to estimate yield and total biomass of a rice crop. Transactions of the ASABE 53(1), 21–27 (2010)

# A Comparison of Myoelectric Pattern Recognition Methods to Control an Upper Limb Active Exoskeleton

Alberto López-Delis[1], Andrés Felipe Ruiz-Olaya[2],
Teodiano Freire-Bastos[3], and Denis Delisle-Rodríguez[1]

[1] Medical Biophysics Center, University of Oriente, Santiago de Cuba, Cuba
{alberto.lopez,denis.delisle}@cbiomed.cu
[2] Faculty of Electronic and Biomedical Engineering, University Antonio Nariño, Colombia
andresru@uan.edu.co
[3] PPGEE, Federal University of Espirito Santo, Vitoria, Brazil
teodiano@gmail.com

**Abstract.** Physically impaired people may use Surface Electromyography (sEMG) signals to control assistive devices in an automatic way. sEMG signals directly reflect the human motion intention, they can be used as input information for active exoskeleton control. This paper proposes a set of myoelectric algorithms based on machine learning for detecting movement intention aimed at controlling an upper limb active exoskeleton. The algorithms use a feature extraction stage based on a combination of time and frequency domain features (mean absolute value – waveform length, and auto-regressive model, respectively). The pattern recognition stage uses Linear Discriminant Analysis, K-Nearest Neighbor, Support Vector Machine and Bayesian classifiers. Additionally, two post-processing techniques are incorporated: majority vote and transition removal. The performance of the algorithms is evaluated with parameters of sensitivity, specificity, positive predictive value, error rate and active error rate, under typical conditions. These evaluations allow identifying pattern recognition algorithms for real-time control of an active exoskeleton.

**Keywords:** Movement intention detection, myoelectric patterns recognition, machine learning, majority vote, surface electromyography, transition removal.

## 1 Introduction

Passive prostheses and orthoses are devices for functional compensation and physical rehabilitation of the human motor system. These are used on people suffering amputations and muscular disorders, but do not provide an intuitive reaction in its control to restore motor functions. On the other hand, active exoskeletons and myoelectric prostheses execute these functions in a natural way according to its learning process [1]. Surface Electromyography signal is the electrical manifestation of the neuromuscular activation associated with a contracting muscle [1]. sEMG pattern recognition based on control has emerged as a promising alternative in rehabilitation robotic devices [1]. Many studies have evaluated sEMG features in

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 100–107, 2013.

classification algorithms aiming to control active prostheses and robotic exoskeletons [2]. Different features extraction methods have been used in pattern recognition involving both time domain and time-frequency domain features. Some of these include mean absolute value [3], zero crossings (ZC) [3], slope sign changes (SSC) [3], auto-regressive (AR) model coefficients [3], cepstrum coefficients [3], waveform length (WL) [3] and wavelet packet transform [3]. Numerous studies have been proposed to classify the features extracted from the sEMG like Bayesian classifier (BYN) [4], linear discriminant analysis (LDA) [5], hidden Markov model [6], multi-layer perceptron (MLP) [4], fuzzy classifier [7], gaussian mixture model [8] and support vector machines (SVM) [9]. Most of the studies have been accomplished in health people to verify the feasibility of implemented algorithms for sEMG-based pattern recognition in human upper limbs.

This work is motivated by the ongoing development of a 4-Degree of Freedom (DoF) upper limb active exoskeleton for muscular rehabilitation therapies. The first stage of this work is related to the performance evaluation in off-line mode of myoelectric algorithms to control external devices. Next section describes the methodology utilized in the feature extraction methods, the myoelectric pattern classification process and the post-processing algorithms, supported on an experimental protocol. Also, the quantitative parameters used in the performance evaluation are here described. Later, the results and discussions are presented, based on the qualitative and quantitative parameters set. Finally, the conclusion about of this work is presented.

## 2 Methods

Figure 1 shows the blocks diagram of the different myoelectric algorithms. First, the sEMG data are segmented in windows of 256 ms, overlapped of 32 ms, taking into account that delays in myoelectric control must be inferior to 300 ms [1]. Later, each data segment is processed through a feature extraction method conformed from a combination of parameters in temporal and spectral domains aimed at extracting information from sEMG. Linear Discriminant Analysis, Support Vector Machine, K-Nearest Neighbor (KNN) and Bayesian classifier are employed for pattern recognition of seven classes, associated to upper limb movement. Finally, majority vote and transition removal algorithms are used to improve the pattern classification results.

### 2.1 Experimental Protocol Description

The stages of training and validation of the proposed algorithms were implemented using a set of signals from a sEMG database provided by the University of Carleton, Canada [8] from thirty healthy subjects. From this database, six sEMG recordings were taken for each subject, in four trials. Acquired recordings on eight channels with a sampling frequency of 3 kHz were provided through Ag-AgCl electrodes arranged at locations of the upper limb as shown in figure 2. Previous to the classification process, data were undersampled to 1 kHz. In each trial, subjects repeated four times, and in a random way, the following seven movements: hand open, hand close, wrist flexion, wrist extension, forearm pronation, forearm supination and resting. Each movement repetition lasted 3 s. A rest period of 5 s was introduced at beginning and ending of each trial, then the whole trial lasted 94 s [7]. The class identifiers for different movements are the following: 1- hand open; 2-hand close; 3-wrist flexion; 4-wrist extension; 5-forearm supination; 6-forearm pronation; 7-resting.

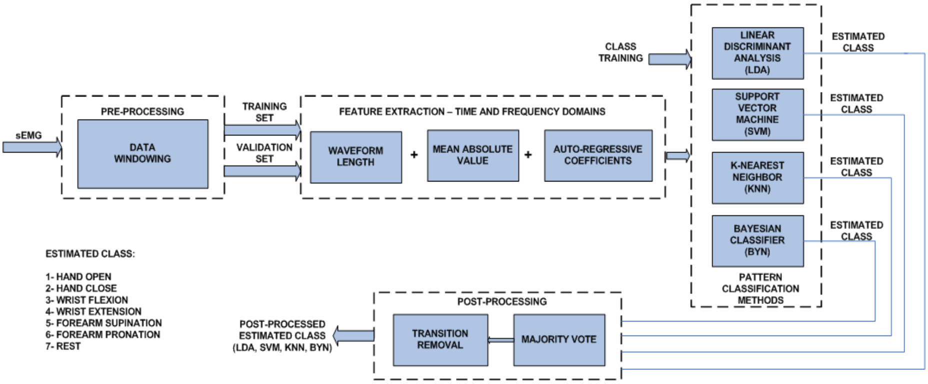


**Fig. 1.** Block diagram of the proposed myoelectric algorithms

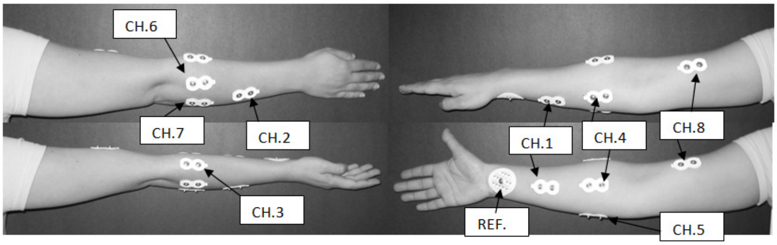


**Fig. 2.** Position of the bipolar electrodes associated to sEMG channels

## 2.2 Feature Extraction Methods

The feature extraction method includes a combination of time and frequency domain parameters. Recent researches have demonstrated that this mixture vectors is a functional and efficient configuration [2]. This configuration provides a good classification accuracy and, is computationally efficient, which facilitates its implementation on embedded systems. Furthermore, it is more robust to the displacement of the surface electrodes. In the temporal domain the mean absolute value (MAV) and the waveform length (WL) were used. The MAV provides the average amplitude of $x_i$ in the segment $i$ that is $N$ samples in length, see equation (1). The WL provides the cumulative length of the waveform over the time segment, see equation (2).

$$MAV = \frac{1}{N}\sum_{l=1}^{N}\left|x_i\right|, \tag{1}$$

$$WL = \sum_{l=1}^{N-1} \left| x_{i+1} - x_i \right|, \tag{2}$$

In the frequency domain, an Auto-Regressive (AR) model was implemented, basically expressed by follow expression:

$$x_i = \sum_{p=1}^{P} a_p \, x_{i-p} + w_i, \tag{3}$$

where $P$ is the order of the AR model and $w_i$ the white noise error. In the sEMG-based pattern recognition process, the coefficients of the AR model $a_p$ have been used as the feature vector. The AR model was based on Levinson–Durbin recursive method. This method is efficient at computation level in the calculus of the linear prediction coefficients, supported on the autocorrelation matrix [3]. Considering that any one of a four-order to six-order auto-regressive model is enough to represent the signal as a temporal series for the recursive method, a four-order model to obtain the linear prediction coefficients was defined [3]. Finally, in the feature extraction process, a concatenation of vectors of several parameters calculated from each sEMG channel was obtained: 1 MAV coefficients, 1 WL coefficients and 4 AR coefficients.

### 2.3 Myoelectric Pattern Classification

After extracting feature vectors, four classification methods (classifiers) were applied independently, according to the proposal myoelectric algorithms (LDA, SVM, KNN, BYN), see figure 2. Each sEMG channel and theirs characteristic vectors were concatenated from the first four auto-regressive coefficients, *MAV* and *WL* values, resulting in 48 coefficients (8 channels x 6 characteristic vectors/channel). Those feature vectors are the input to the different classifiers. The output of each classifier represents in each time anyone the seven motion class, see figure 2. Linear Discriminant Analysis technique [5] maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby guaranteeing maximal separability. This classification algorithm does not require iterative training, avoiding the problems with over-training that appear in artificial neural networks. Support Vector Machine constructs an optimal separating hyperplane in a high-dimension feature space of training data that are mapped using a nonlinear kernel function [9]. Therefore, although it uses a linear learning machine method with respect to the nonlinear kernel function, it is in effect a nonlinear classifier. The high generalization and classifying linearly-inseparable patterns with small computational complexity are capabilities of the SVM, which can be useful for classifying sEMG signal patterns whose features tend to change with time and can allow real-time motion classification, respectively [9]. K-nearest neighbor algorithm [10] is a non-parametric method for classifying objects based on closest training examples in the feature space. The k-nearest neighbor algorithm is one of the simplest of all machine learning algorithms. Bayesian classifier [4] is applied for use when features are independent of one another within each class, but it appears to work well in practice even when that independence assumption is not valid. The class-conditional independence assumption greatly simplifies the training step since it is possible to estimate the one-dimensional class-conditional density for each feature individually. The stages of training and validation of proposed algorithms were implemented using cross-validation technique evaluating the results based on partitioning the data into training and test sets. Specifically, $k$-fold cross-validation was used based on the partition the $k$ samples sub-conjunct. One subset is used as testing data and the rest ($k$-1) as training data. For this evaluation the $k$ value ($k$ = 6) is equal to the number of sEMG recordings acquired in one trial. For implementation of the four myoelectric pattern classifiers, the information of the classes during the training process was used.

### 2.4 Post-processing Techniques

The post-processing methods are designed to manage excessive outflows in the classification process and improve the system performance. The majority vote method (MV) uses the current classification result along with the $n$ previous classification (for this case, the eight previous classifications results) and makes a classification decision based on the class that appears more often [8]. The resulting effect is a smooth operation that removes spurious misclassification. The number of decisions that can be used in majority vote depends upon the length of the analysis window, the system processing delay, and the total system delay tolerable by the user for the exoskeleton control. On the other side, the errors that are present normally occur during transitional periods, which are expected as the system is in an undetermined state between contractions. Indeed, it is possible to remove them using transition removal algorithms [8].

## 3 Results

Feature extraction and patterns classification algorithms were implemented in an off-line mode using functions in Matlab (Mathworks Inc., Natick, MA). The performance of the proposed algorithms was evaluated based on quantitative measures that include sensitivity (SS), specificity (SP), predictive positive value (PPV), total error rate of classification (TER) and active error rate (AER). An active decision is a single output class from the classifier resulting in limb motion. Figure 3 presents the scatter plot based on the feature vectors and the representative motion class from the proposed myoelectric algorithms, for the eight myoelectric channels. From a qualitative evaluation, the four classifiers provide a good discrimination of the wrist flexion and extension motion class, based on *MAV*, *WL* and auto-regressive feature vectors. The others motion class (hand open, hand close, supination, pronation and rest) were grouped in homogenous and similar way from the four classifiers. Figure 4 shows the statistical dispersion based on the total error rate, sensitivity, specificity, active error rate and predictive positive value without post-processing techniques (majority vote and transition removal).

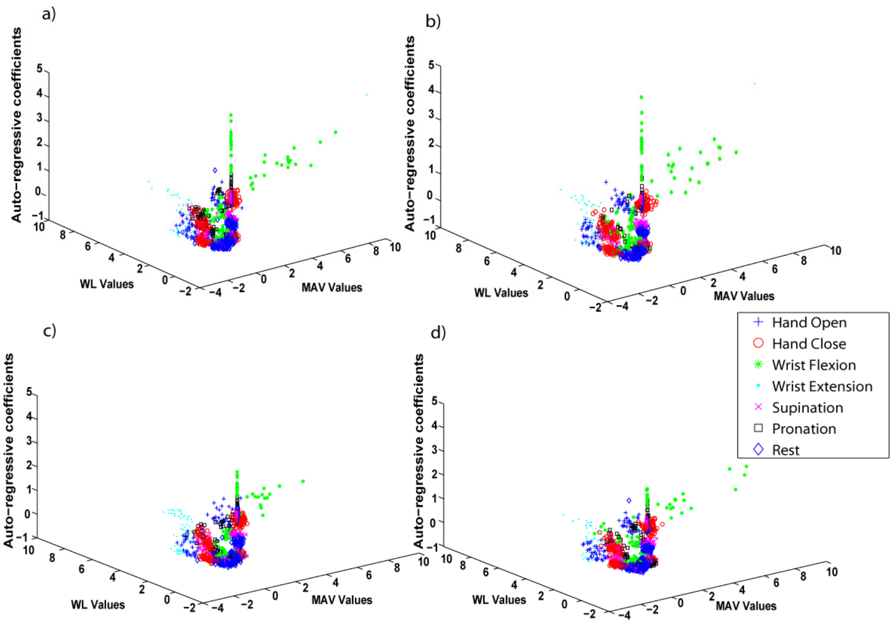


**Fig. 3.** Scatter plot the feature vectors and the motion class from the proposed myoelectric algorithms: a) LDA classifier; b) KNN classifier; c) SVM classifier; and d) BYN classifier

LDA, KNN and SVM classifiers (Fig.4a, b and c) present a similar performance from quantitative parameters. The total and active error rate (TER and AER) in the Bayesian classifier (Fig. 4d) is higher respecting to others classifiers, meaning lower accuracy during the movement action classification. Additionally, the specificity (SP) accuracy is lower, expressing that the movement actions proportion correctly rejected is lower respect to the previous classifiers. Therefore, the false positive number is higher during the classification process. From the above results and taking as example the Bayesian classifier, table 1 shows the confusion matrix from one working section the experimental protocol. Rows in the matrix represent the inputs related to classes that are required to obtain, and columns represent obtained patterns as classifier outputs. The main diagonal in both matrices represents the concordance between the true and obtained classes. Shared cells in the confusion matrix of the first table, under the main diagonal, present positive falses, i.e., a number of occurrences of motion class with the class that should be obtained. This is caused by the dispersion of the feature vectors and their relation with the motion class based on the assumption that not always is accurate, for Bayesian classifiers, that the predictor variables are independent. The second table shows the results obtained with the combinations of the majority vote and transition removal technique. The total removing of the positive falses with the combinations of these techniques is observed. Nevertheless, a considerable reduction of the motion class corrected classified from main diagonal is generated, as well as the motion class execution time. This is caused by removing the transition periods at the beginning and end of the motion class period while contractions occur.

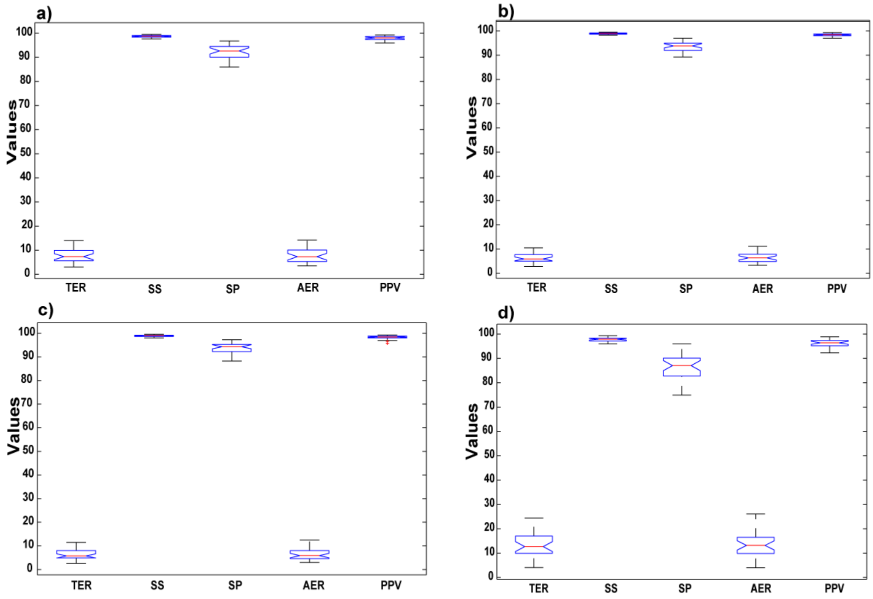


**Fig. 4.** Statistical results for a representative classification from the proposed myoelectric algorithms: a) LDA classifier; b) KNN classifier; c) SVM classifier; and d) BYN classifier.

**Table 1.** Confusion matrix of the Bayesian classifier

| | Bayesian Classifier without post-processing | | | | | | |
|---|---|---|---|---|---|---|---|
| | Hand Opened | Hand Close | Wrist Flexion | Wrist Extension | Forearm Pronation | Forearm Supination | Rest |
| Hand Opened | 293 | 2 | 1 | 0 | 1 | 0 | 4 |
| Hand Close | 2 | 292 | 2 | 0 | 0 | 1 | 0 |
| Wrist Flexion | 2 | 4 | 287 | 0 | 1 | 1 | 3 |
| Wrist Extension | 7 | 1 | 1 | 287 | 0 | 0 | 0 |
| Forearm Pronation | 5 | 3 | 14 | 3 | 280 | 0 | 1 |
| Forearm Supinatio | 5 | 3 | 8 | 1 | 4 | 280 | 3 |
| Rest | 3 | 5 | 3 | 1 | 3 | 0 | 506 |
| | Bayesian Classifier with majority vote and remove transitions | | | | | | |
| Hand Opened | 29 | 0 | 0 | 0 | 0 | 0 | 0 |
| Hand Close | 0 | 72 | 0 | 0 | 0 | 0 | 0 |
| Wrist Flexion | 0 | 0 | 27 | 0 | 0 | 0 | 0 |
| Wrist Extension | 0 | 0 | 0 | 23 | 0 | 0 | 0 |
| Forearm Pronation | 0 | 0 | 0 | 0 | 57 | 0 | 0 |
| Forearm Supinatio | 0 | 0 | 0 | 0 | 0 | 7 | 0 |
| Rest | 0 | 0 | 0 | 0 | 0 | 0 | 152 |

## 4 Conclusions

The control of exoskeletons working as an assistance or rehabilitation tools requires special considerations such as robustness, reliability and safe. These are mandatory requirements taking into account that the device must identify the user movement intention, analyze the information in real-time and compute the mechanical power to release in the right instant. This paper described the obtained results in a comparative study of four proposed algorithms to approach the detection of movement intentionality. Selected algorithms aim to control a robotic upper limb exoskeleton using sEMG signals. LDA, SVM and KNN have presented better accuracy than Bayesian classifier. Nevertheless, the execution time during the training and evaluation process of the Bayesian classifier (292 ms) is considerably lower than the other classifiers (LDA – 1.22 s, SVM – 700 ms and KNN – 428 ms). This result is an important parameter to be considered for its implementation in on-line mode. In this mode, the performance of the proposal algorithms could be improved using the post-processing techniques (majority vote and transition removal), but it is important to evaluate the number of decisions that can be used, as well as the length of the analysis window, taking into account that delays in myoelectric control. As future work, it is required to implement other algorithms and evaluate them under other conditions in order to obtain an optimal solution for myoelectric control.

**Acknowledgments.** Authors thank CAPES/Brazil (Project-018/2011 CAPES/MESCUBA), VCTI (University Antonio Nariño, Colombia) and the VLIR project (University of Oriente, Cuba) for the financial support of this research. Additionally, to Dr. C. Adrian Chan of Carleton University, Canada, for providing us the database of electromyography signals.

## References

1. Ho, S.L., Sheng, Q.X.: Exoskeleton robots for upper-limb rehabilitation: State of the art and future prospects. Med. Eng. Phys. (2011), doi:10.1016/j.medengphy.2011.10.004
2. Phinyomark, A., Phukpattaranont, P., Limsakul, C.: Feature reduction and selection for EMG signal classification. Expert Systems with Applications 39, 7420–7431 (2012)
3. Zecca, M., Micera, S., Carrozza, M.C., Dario, P.: Control of multifunctional prosthetic and by processing the electromyographic signal. Critical Reviews in Biomedical Engineering 30(4-6), 459–485 (2002)
4. Englehart, K., Hudgins, B., Parker, P.A., Stevenson, M.: Classification of the myoelectric signal using time–frequency based representations. Med. Eng. Phys. 21, 431–438 (1999)
5. Englehart, K., Hudgins, B., Parker, P.A.: A wavelet-based continuous classification scheme for multifunctionmyoelectric control. IEEE Trans. Biomed. Eng. 48(3), 302–310 (2001)
6. Chan, A.D.C., Englehart, K.: Continuous myoelectric control for powered prostheses using hidden Markov models. IEEE Trans. Biomed. Eng. 52(1), 121–124 (2005)
7. Ajiboye, A.B., Weir, R.F.: A heuristic fuzzy logic approach to EMG pattern recognition for multifunctional prosthesis control. IEEE Trans. Neural Syst. Rehabil. Eng. 13(3), 280–291 (2005)
8. Huang, Y.H., Englehart, K., Hudgins, B.S., Chan, A.D.C.: A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. IEEE Trans. Biomed. Eng. 52(11), 1801–1811 (2005)
9. Oskoei, M.A.: Support Vector Machine-Based Classification Scheme for Myoelectric Control Applied to Upper Limb. IEEE Transactions on Biomedical Engineering 55(8), 1956–1965 (2008)
10. Shakhnarovish, Darrell, Indyk: Nearest-Neighbor Methods in Learning and Vision. MIT Press (2005)

# An Arabic Optical Character Recognition System Using Restricted Boltzmann Machines

Abdullah M. Rashwan, Mohamed S. Kamel, and Fakhri Karray

University of Waterloo

**Abstract.** Most of the state-of-the-art Arabic Optical Character Recognition systems use Hidden Markov Models to model Arabic characters. Much of the attention is paid to provide the HMM system with new features, pre-processing, or post-processing modules to improve the performances. In this paper, we present an Arabic OCR system using Restricted Boltzmann Machines (RBMs) to model Arabic characters. The recently announced ALTEC dataset for typewritten OCR system is used to train and test the system. The results show a 26% increase in the average word accuracy rate and 8% increase in the average character accuracy rate compared to the HMM system.

## 1 Introduction

Digitizing information sources and making them available for the Internet users is taking much attention these days in both academic and industrial fields. Unlike typewritten Latin OCRs, typewritten OCRs for cursive scripted languages (ex. Arabic, Persian, etc.) still encounter a plethora of unsolved problems. The best average word accuracy rate achieved for large vocabulary Arabic OCR according to the rigorous tests done by ALTEC organization on 3 different commercial OCRs in 2011 was slightly below 75%. Providing a decent solution for recognizing cursive scripted language will allow millions of books to be available on the Internet, it will also push Arabic Document Management Systems (DMSs) steps forward.

There are few papers that tackled the typewritten Arabic OCR problems. The lack of dataset is one of the main reasons that directed the researchers away from tackling this problem. In 1999,[1] designed a complete Arabic OCR system, they used a good dataset to train and test their system. The dataset however is biased towards magazine documents. They reported a character error rate of 3.3%, such high accuracy is not only a result of using HMM classifier, but also because they supported the classifier with strong preprocessing and postprocessing modules to boost the accuracy. In 2007, [2] designed an Arabic multi-font OCR system using discrete HMMs along with intensity based features. Character models were implemented using mono and tri models. He achieved a character accuracy rate for the Simplified Arabic font of 77.8% using mono-models. In 2009, [3] introduced new features to use along with discrete HMMs to model Arabic characters. These features are less sensitive to font size and

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 108–115, 2013.

style. They reported character accuracy rate of 99.3% which is very high, such high accuracy is due to the use of a synthesized and very clean dataset, and using high quality images scanned at 600dpi. In 2012, [4] developed the previous system using a practical dataset created by ALTEC. The reported character accuracy rates for the HMM system using bi-gram and 4-gram character language model were almost 84% and 88% respectively. The paper introduced also an OCR system using Pseudo 2D HMM achieving a CER of 94% and gaining more than 6% over using HMM. Most of the work done so far to tackle the problem of recognizing Arabic letters, and cursive scripted languages in general, lacks either a good dataset to build a practical model or a good model to come up with a decent solution. In this work, a good model along with a practical dataset have been used to overcome the limitations and drawbacks of the previous systems.

In the following sections, Sec. 2 presents the system architecture and the feature extraction, Sec. 3 presents HMMs and how to apply them on the Arabic OCR problem, Sec. 4 presents the RBMs and how they can improve the performance of the HMM system, experimental results are presented in Sec. 5. The final conclusions are presented in Sec. 6.

## 2 System Architecture

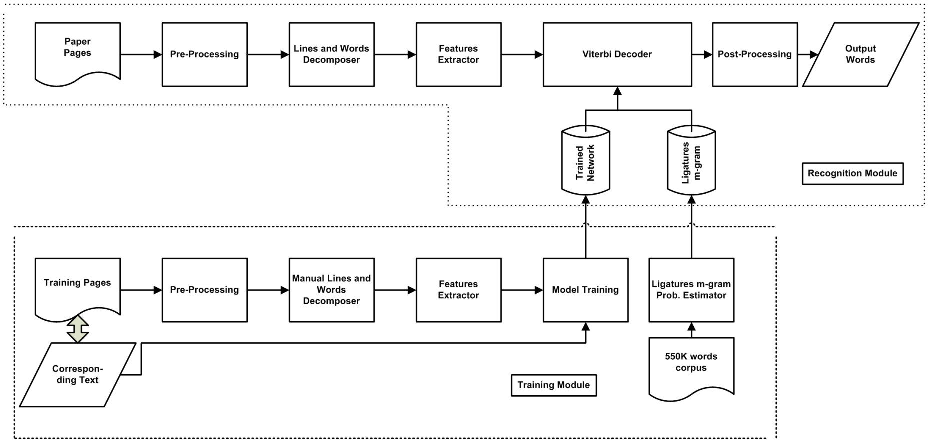


**Fig. 1.** System architecture of the Arabic OCR

The architecture of our OCR system is shown in Fig. 1. In the recognition phase, first, the printed text pages are scanned and digitized. Lines and words boundaries are specified automatically using histogram-based algorithm. After automatically extracting the words, each word is segmented to vertical frames and features are extracted from each frame. A moving window of width 11 pixels and a step size of 1 pixel, is used to extract the features for each frame.

The use of 11-pixels window size is due to the fact that the average character width is ~11 pixel. The features are simply the row pixels, the word height is resized to 20 pixels in order to form a fixed length feature vector. Using a Viterbi decoder, the most likely characters sequence is obtained using the feature vectors sequence, the classifier model, and n-gram language model probabilities that were constructed during the training phase. The decoder output is then processed to obtain the recognized words.

In the training phase, the printed text pages are photocopied using different photocopiers. Both pages, the original and the photocopied, are scanned and digitized. Different scanners and photocopiers are used in order to represent practical noise in the training database. Digitized images follow the same steps in the recognition mode to extract the frames. Frames are then concatenated in sequence and stored in order to be used in the parameters estimation. The extracted features along with the corresponding text are used in the parameters estimation for the classifier. In this paper, three different classifiers are used in the system; HMM classifier, neural network trained using backpropagation algorithm, and neural network pretrianed using RBMs. The power of the RBMs is that they can model any distribution without making assumptions that the distribution is Gaussian or discrete as in the classical HMMs. A training scheme is used to insure a good estimation of the parameters and different experiments are held to achieve the best performance. On the other hand, a corpus of ~0.5 Giga word is used in estimating the character-level language model. Both the language model and the classifier model are used in the recognition phase.

## 3 HMM and Arabic OCR

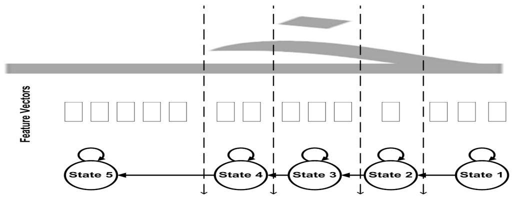


**Fig. 2.** The figure shows a five state HMM model of the 'خـ' character and the state aligment of extracted features. Zeros in the feature vectors represent the dummy segments inserted in order to form fixed size feature vectors.

First we have to define the number of character models that we want to use. The advantage of using large number of models is that the system will be able to distinguish between different shapes easily but that requires having sufficient number of training examples per each model. One should decide the number of different models based on the training database. After deciding the number of models, we have to decide the HMM model per each character. It is common to use first order HMM with right-to-left topology as shown in Fig. 2. The number of hidden states can be either the same for all models or each model can have different number of hidden states. Because we deal with discrete HMMs in this work, the feature vectors needs to be quantized.

The vector-quantizer serially maps the feature vectors to quantized observations. A codebook generated by the codebook maker module during the training phase is used for such mapping. The codebook size is a key factor that affects the recognition accuracy and is specified empirically to minimize the WER of the system.

Using numerous sample text images of the training set of documents evenly representing the various Arabic characters, the codebook maker creates a codebook using the LBG vector clustering algorithm. The LBG algorithm is chosen for its simplicity and relative insensitivity to the randomly chosen initial centroids compared with the basic K-means algorithm.

## 4 Restricted Boltzmann Machines

We use a Neural Network to replace the discrete distribution for the hidden HMM states. The Neural Network is first pre-trained as a multilayer generative model of a window of spectral feature vectors without making use of any discriminative information. Once the generative pre-training has designed the features, we perform discriminative fine-tuning using backpropagation to adjust the features slightly to make them better at predicting a probability distribution over the states of hidden Markov models.

### 4.1 Learning the Restricted Boltzmann Machines

As explained in [5], Restricted Boltzmann Machines (RBMs) are used to learn multi-layers of deep belief networks such that only one layer is learned at a time. An RBM is single layer and is restricted in the sense that no visible-visible or hidden-hidden connections are allowed. The learning works in a bottom-up scheme using a stack of RBM layers. The weights for the first layer are estimated using the input features, and then we fix the weights of the first layer and use its latent variables as an input to second layer. In such way, we can learn as many layers as we like.

In binary RBMs, both the visible and hidden units are binary and stochastic. The energy function of the binary RBMs is:

$$E(v,h|\theta) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j$$

Where $\theta = (w, a, b)$ and $w_{ij}$ represents the symmetric interaction term between visible unit $i$ and hidden unit $j$ while $b_i$ and $a_j$ are their bias terms. $V$ and $H$ is the number of the visible and hidden units.

The probability that an RBM assigns to a visible vector $v$ is:

$$p(v|\theta) = \frac{\sum_h e^{-E(v,h)}}{\sum_u \sum_h e^{-E(v,h)}} \tag{1}$$

Since there are no hidden-hidden connections and visible-visible connections, the conditional distribution $p(h|v,\theta)$ and $p(v|h,\theta)$ can be presented as:

$$p(h_j = 1|v,\theta) = \sigma(a_j + \sum_{i=1}^{V} w_{ij}v_i) \tag{2}$$

$$p(v_i = 1|h,\theta) = \sigma(b_i + \sum_{j=1}^{H} w_{ij}h_j) \tag{3}$$

Using "contrastive divergence" training procedure [6], the weights can be updated using the following update rule:

$$\Delta w_{ij} \ \propto < v_i h_i >_{data} - < v_i h_i >_{reconstruction} \tag{4}$$

### 4.2 Using RBMs in Arabic OCR

When using HMMs, the commonly used method for sequential data modeling, the observations are modeled using either discrete models or Gaussian Mixture Models (GMMs). Although these models are proven to be useful in many applications, they encounter serious drawbacks and limitations [5,7]. In this work, Neural Network (NN) will replace the discrete models or GMMs to model the inner states for the Arabic characters. The recognizer will use the posterior probabilities over the inner states from the NN combined with a Viterbi decoder in order to recognize an input Arabic word. A well-trained HMM system is used for state alignment of all training database in which feature vectors are assigned to states they belong to (Fig. 2). In the pre-training step, the Neural Network is pre-trained generatively using a stack of RBMs, this is essential for avoiding over fitting and to ensure a good convergence point in the classification step. In the classification step, a softmax layer is added to the pre-trained network then the network is trained using backpropagation algorithm. The trained Neural Network is used to predicting a probability distribution over the states of the HMM replacing the Gaussian Mixture Models (GMMs) or the Discrete Models.

## 5 Experimental Results and Evaluation

The system structure presented in Sec. 2 is trained using part of the ALTEC dataset [8]. Only 2 fonts, Simplified Arabic 14 and Arabic Transparent 14, are used to simplify the problem and to compare the RBMs to the HMMs. Around 4500 words scanned at 300dpi resolution are used to train both systems. We used 151 HMM models to model different Arabic character shapes. HTK toolkit is used for models training and recognition [9]. A manually modified version of the HTK is used to recognize the character sequences using the trained Neural Network. SRILM toolkit is used to create the character based n-gram language model [10]. The test database consists of 1790 words, these pages are scanned on different scanners than the ones used for the training dataset.

We analyzed the RBMs performance by conducting four experiments. The first experiment aims at testing the effect of using a stack of RBM layers instead of directly estimating the weights using backpropagation algorithm. The second expirment tests the effect of varying the number of hidden nodes on the system performance. The third experiment aims to test the effect of varying number of hidden layers on the performance. The last experiment compares the RBM-based system to the HMM-based system in terms of accuracy and speed.

For the HMM system, we used a discrete distribution to model the hidden states. We used codebook of size 1000, and the HTK toolkit was used for the training. A forced-alignment on the state level was performed using the HMM models, then the Neural Network, pre-trained generatively using a of RBMs, was trained using the state labeling along with the corresponding feature vectors.

### 5.1 Effect of Pretraining Using RBMs

**Table 1.** Word accuracy rates for the backpropagation training algorithm and the neural network pretrained using RBMs

| | Backpropagation | RBM Accuracy |
|---|---|---|
| Average WAR | 72% | 81% |

**Table 2.** Detailed word accuracy rates when varying the number of hidden units

| Number of Hidden Units | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|
| Average WAR | 61% | 72% | 79% | 81% |

**Table 3.** Detailed word accuracy rates when varying the number of hidden layers

| | 1 hidden layer | 2 hidden layers |
|---|---|---|
| Average WAR | 71.7% | 72.4% |

The goal of these experiments is to test the effect of using RBMs. We constructed two neural networks of 1 hidden layer and 1000 hidden nodes. We trained the first network using backpropagation directly, and we pretrained the second network using RBMs then we tuned it using backpropagation algorithm. Table 1 shows the word accuracy rates for both networks, we can clearly see a 9% increase in the accuracy which is very large. Using RBMs is even more useful when we deal with multi-layer neural networks[11].

### 5.2 Varying the Number of Hidden Units

In this experiment, we trained the neural network and varied the number of hidden units from 100 to 1000. The evaluation is based on the word accuracy rates. As shown in Table 2, we can see that the word accuracy rate increases as we increase the number of hidden units. The gain we obtained is not linear with increasing the number of hidden units, it takes the shape of the log curve where the accuracy saturates at certain level no matter the number of hidden units that you add.

### 5.3 Varying the Number of Hidden Layers

In this experiment, we compared the accuracy using 1 and 2 hidden layers. We fixed the number of hidden units to 200 and we compared a neural network with a 200 nodes layer to a neural network with a two stacked hidden layers with 100 hidden nodes each. Fig. 3 shows that we gained ~0.7% accuracy just by using 2-layers without changing the number of nodes. This is useful when we want to improve the accuracy without increasing the recognition time, we will however face the difficulties of tuning multi-layer network during the training phase.

### 5.4 Comparing the RBM and the HMM Based Systems

**Table 4.** Word and Character Accuracy Rates

| | HMM Accuracy | RBM Accuracy |
|---|---|---|
| Average WAR | 55% | 81% |
| Average CAR | 87% | 95.2% |
| Recognition Time (m.sec) | 26 | 198 |

In these experiments, we used a discrete HMM-based system with a codebook of size 1000. We compared the performance of this system to the RBM system where the number of hidden nodes is also 1000. For both systems we used the same features and the same bigram language model, and the codebook size for the HMM system equals to the number of hidden nodes for the Neural Network system. The system performance is evaluated using word accuracy rate. As shown in Table 4, the performance of the Neural Network system is much higher than the HMM system. This is due to the fact that the HMM is a generative model that tries to maximize the likelihood of the data. On the other hand, the Neural Network is pretrained using RBMs, then a fine tuning discriminative training is performed using the backpropagation algorithm. Although the word accuracy rate is more practical and intuitive, the Character Accuracy Rate (CAR) is also important and the results are shown in Table 4.

Of course there is something we should pay for such improvement, the computational complexity of the Neural Network based system is higher than the one for the discrete HMM. The discrete HMM simply stores the probabilities in lookup tables, that makes it too fast to evaluate the hidden state distributions. The Neural Network system has to evaluate the output of the 1000 hidden nodes plus the output of the 1384 output nodes. The recognition time per character for both systems is listed in Table. 4.

## 6 Conclusion and Future Work

Recently, research has been directed to improve the inputs and outputs (ex. preprocessing, features, and postprocessing) to the HMM model other than to improve the character modeling. In this paper, we have made use of RBMs to model Arabic characters. The experimental results looks very promising compared to a baseline HMM system. For future work, using more set of features can be used instead of the row pixels, also more font sizes and styles will be involved in the training.

## References

1. Bazzi, I., Schwartz, R., Makhoul, J.: An omnifont open-vocabulary ocr system for english and arabic. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 495–504 (1999)
2. Khorsheed, M.: Offline recognition of omnifont arabic text using the hmm toolkit (htk). Pattern Recognition Letters 28, 1563–1571 (2007)
3. Attia, M., Rashwan, M., El-Mahallawy, M.: Autonomously normalized horizontal differentials as features for hmm-based omni font-written ocr systems for cursively scripted languages. In: 2009 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 185–190 (2009)
4. Rashwan, A.M., Rashwan, M.A., Abdel-Hameed, A., Abdou, S., Khalil, A.H.: A robust omnifont open-vocabulary arabic ocr system using pseudo-2d-hmm. In: Proc. SPIE, vol. 8297 (2012)
5. Mohamed, A., Dahl, G., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing 20, 14–22 (2012)
6. Hinton, G.: Training products of experts by minimizing contrastive divergence. Neural Computation 14, 2002 (2000)
7. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine 29 (2012)
8. Altec database (2011), `http://www.altec-center.org/conference/?page_id=84`
9. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK (2006)
10. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proceedings of ICSLP, Denver, USA, vol. 2, pp. 901–904 (2002)
11. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313, 504–507 (2006)

# Astronomical Image Data Reduction for Moving Object Detection

Kevin Allekotte[1], Pablo De Cristóforis[1], Mario Melita[2], and Marta Mejail[1]

[1] Departamento de Computación, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires, Argentina
[2] Instituto de Astronomía y Física del Espacio, Consejo Nacional de Investigaciones
Científicas y Técnicas, Argentina
{kallekotte,pdecris,marta}@dc.uba.ar, melita@iafe.uba.ar

**Abstract.** In this work we present a system for autonomous discovery of asteroids, space trash and other moving objects. This system performs astronomical image data reduction based on an image processing pipeline. The processing steps of the pipeline include astrometric and photometric reduction, sequence alignment, moving object detection and astronomical analysis, making the system capable of discovering and monitoring previously unknown moving objects in the night sky.

**Keywords:** astronomical images, data reduction, moving object detection.

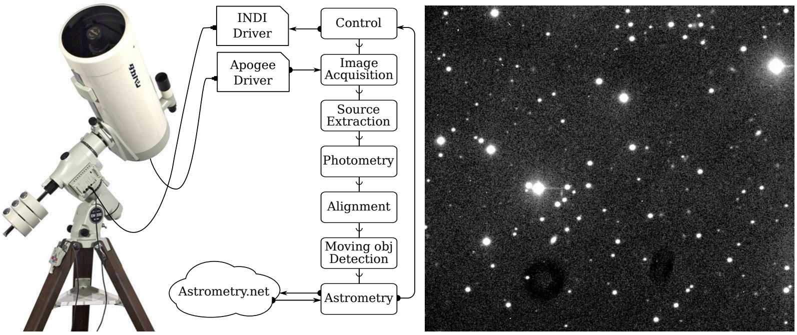
## 1 Introduction

One of the goals of modern astronomy is the exhaustive study and mapping of celestial bodies, in particular the moving bodies of the solar system. In the past, observations were made and analyzed manually, nowadays we have robotic telescopes with CCD cameras. This allows the development of systems that control the telescope and automate the data extraction and reduction from astronomical images. The replacement of rutinary tasks performed by astronomers using autonomous systems has multiple advantages and enables a continuous and long exploration (even in hostile environments) resulting in higher probabilities of achieving astronomical discoveries.

In this work we present a method for autonomous discovery of asteroids, space trash and other moving objects. This method performs astronomical image data reduction based on an image processing pipeline to find moving objects in the night sky and control the telescope for an automated tracking. The processing steps of the pipeline include astrometric and photometric reduction, sequence alignment, moving object detection and astronomical analysis, making the system capable of discovering and monitoring previously unknown moving objects. In Figure 1, the pipeline data flow and a sample of an astronomical image are shown.

Some of the previous related works include IRAF [1] (Image Reduction and Analysis Facility), a well known system that implements most methods for astronomical data reduction. The main drawback of this system is that it is designed

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 116–123, 2013.



(a) (b)

**Fig. 1.** (a) System's pipeline data flow and (b) sample of astronomical image captured with Takahashi Mewlon-210 telescope and Apogee Alta F16 CCD camera.

for manual analysis of the data, so it is rather complicated to use as a library with a programmable interface and, more importantly, has no moving object detection routines. The purpose of this paper is to address these issues.

The remainder of this paper is organized as follows: Section 2 details the proposed method, section 3 presents the results and section 4 summarizes our work.

## 2 Proposed Method

The images are captured with a CCD camera and stored in the FITS (Flexible Image Transport System) file format. The images usually contain a few hundred light sources, depending on the region of the sky and the angular field of the telescope. The profile of each light source can be approximated with a 2D Gaussian curve (if the camera is in focus). The one-dimensional profile of this Gaussian curve is characterized in astronomy by the FWHM (Full Width at Half Maximum) which depends on the dispersing process in the atmosphere and therefore is approximately constant for all point light sources of the image [2]. These images also contain background noise that is characterized by an additive and a multiplicative component. Furthermore, there are spurious detections caused by defective pixels or cosmic rays that randomly reach the sensor and cause peaks in the image. Figure 2 shows examples of light sources in astronomical images.

The moving objects we are interested in finding appear in the image with a very similar profile as the stars and satisfy the following conditions: a) $v \ll \frac{\mathrm{FWHM}}{\Delta t_{exp}}$, b) $\Delta t \gg \frac{FWHM}{v}$ and c) $\frac{\#D}{N \Delta t} > v$, where $v$ is the velocity of the object, $\Delta t_{exp}$ is the exposure time, $\Delta t$ is the time interval of the sequence, $\#D$ is the size of the captured image $D$ and $N$ is the number of images (frames) in the sequence (typically $N \geq 4$). The time interval between the images of the sequence, $\Delta t$, is chosen according to $v$.

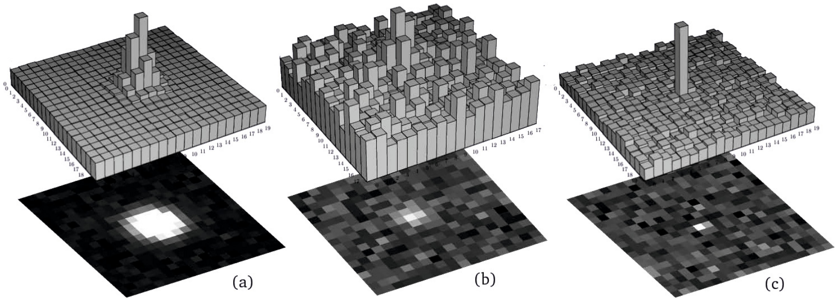


**Fig. 2.** Examples of light sources in a image: (a) bright star, (b) faint star and (c) a noise artifact (cosmic ray)

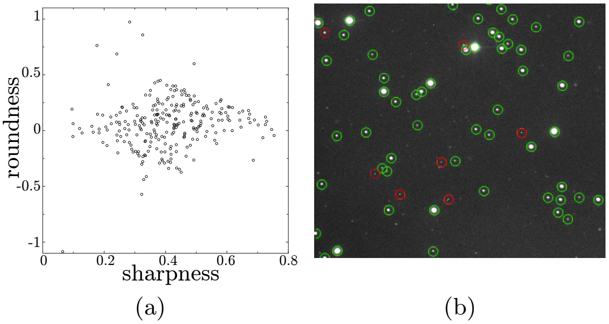
**Source Extraction.** The first step in the image processing pipeline is extracting the location of light sources, including stars and possible moving objects. In this work we explore two options: SExtractor (Source Extractor library) [3], which uses neural network algorithm to find point light sources and also other astronomic features; and DAOPHOT [4] a source extraction method which finds the location of sources by filtering the image with a gaussian convolution and then extracting the local maximums which values are above a threshold. A performance comparison between both methods is presented in section 3.

Some of the detected features in the images do not correspond to point light sources, but are rather artifacts of noise like cosmic rays or defective pixels in the CCD sensor. To filter out these spurious detections, we analyze each feature's roundness and sharpness parameters.

Let $\mathcal{N}(i_0, j_0)$ be a neighborhood around $(i_0, j_0)$. The sharpness of a feature centered in $(i_0, j_0)$ is defined as $\text{sharp}_{i_0,j_0} = \frac{D_{i_0,j_0} - \langle D_{i,j} \rangle}{H_{i_0,j_0}}$, where $\langle D_{i,j} \rangle$ is the mean of the $D_{i,j}$, $(i,j) \in \mathcal{N}(i_0, j_0) - \{(i_0, j_0)\}$ and $H_{i_0,j_0}$ is a filtered version of $D_{i_0,j_0}$. The roundness parameter is defined as $\text{round} = 2\frac{h_y - h_x}{h_y + h_x}$ where $h_y$ and $h_x$ are one-dimensional convolutions of $D$ corresponding to the gaussians: $g_x(\varDelta i; \sigma) = e^{-\frac{\varDelta i^2}{\sigma^2}}$ and $g_y(\varDelta j; \sigma) = e^{-\frac{\varDelta j^2}{\sigma^2}}$, respectively.

In Figure 3(a) the distribution of sharpness/roundness of detected features in the images can be seen. Figure 3(b) shows the detected light sources (green) and the spurious detections (red) filtered by sharpness/roundness parameters.

**Photometry.** The goal of the photometry reduction is to obtain a light flux estimation for each detected source, which represents the number of photons (amount of light) received in an area near the position of the source subtracting the background. This flux is stable over the frames for the same celestial body so it can be used as a descriptor of the source that distinguishes each other. Moreover, to eliminate the effect of the atmosphere the air mass has to be subtracted.



(a) (b)

**Fig. 3.** (a) Sharpness/roundness distribution of detected features and (b) detected light sources (green) and the spurious detections (red) filtered by sharpness/roundness

To estimate the background of the whole image we define $BG$ as an erosion filter of $D$, $BG_{i_0,j_0} = \min_{(i,j)\in\mathcal{N}(i_0,j_0)} D_{i,j}$. Then, the light flux of one source placed in $(i_0, j_0)$ can be calculated as follows: $FLUX_{i_0,j_0} = \sum_{(i,j)\in\mathcal{N}(i_0,j_0)} (D_{i,j} - BG_{i,j})$.

In sparsely populated areas, we can calculate the flux from a source without estimating the background, taking as reference baseline a ring around the source. In this case the $FLUX_{i_0,j_0}$ can be defined as:

$FLUX_{i_0,j_0} = \sum_{\substack{(i,j)\in\\ \text{circ}(i_0,j_0)}} D_{i,j} - \sum_{\substack{(i,j)\in\\ \text{ring}(i_0,j_0)}} D_{i,j} \times \frac{\#\text{circ}(i_0,j_0)}{\#\text{ring}(i_0,j_0)}$, where $\text{circ}(i_0, j_0)$ and $\text{ring}(i_0, j_0)$ are circular and a annular regions around $(i_0, j_0)$, respectively.

**Alignment.** The sequence of acquired images is obtained under the same sky coordinates. The robotic telescope corrects for the rotation of the earth as time passes, but still some mechanical errors accumulate. The result is that the light sources appear displaced a few pixels in both $X$ and $Y$ coordinates or even slightly rotated relative one image to another. Then, the sequence of images need to be aligned. The scale of the images does not change because the focal length of the telescope is fixed.

Since the misalignment of the images is given in three degrees of freedom (movement in $X$, $Y$ and rotation). The problem is reduced to finding a rigid transformation to bring all images to the same reference system. The method ICP (Iterative Closest Point) [5] assumes that the features are not very far from their original position, which can be assumed in our case because we want to correct alignment errors due to perturbations in the motion of the telescope. The idea of this method is to establish a correspondence between each feature of an image with the feature closest to it in the second image. Then the transformation that fits all correspondances best is computed using RANSAC (Random Sample and Consensus).

**Moving Object Detection.** Since the objects of interest move at an approximately constant speed and the field of vision of the telescope is small (long focal length), we can consider that the trajectory of the object describes a straight line. Then, the problem of finding moving objects is then reduced to detect a sequence of features that are collinear and are spacially distributed proportionally to the time intervals of the frames[1]. If we think of the features as points in a three dimensional space with coordinates $(x, y, t)$, we are looking for a sequence of collinear points in this space (see Figure 4).

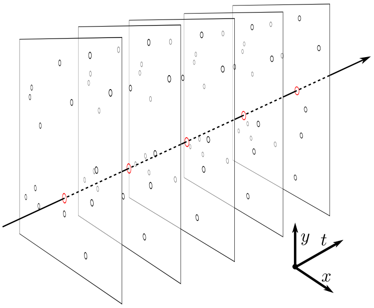


**Fig. 4.** Set of features aligned in space-time coordinates, corresponding to a moving object along the images sequence of frames.

The first step is discarding the features that appear repeatedly in the same position $(x, y)$ along the sequence of frames, i.e., corresponding to sources that do not move (mostly stars). Some of the features that remain correspond to noise and artifacts that randomly appear in a frame, and some of them might correspond to a moving object that appears at different positions along the sequence of frames. The challenge is to find which features are aligned in a trajectory, in an efficient way.

We also have to keep in mind that we might possibly detect the object only in a subset of the frames, so we have to search for collinearities in different subsets of frames. To achieve this we choose three frames $(f_1, f_2, f_3)$ randomly and seek for triplets of features which are collinear in $(x, y, t)$.

For each pair of features $(p_1, p_2)$, $p_1 \in f_1$ and $p_2 \in f_2$, we calculate the spatial displacement vector $p_1 \rightarrow p_2$ and estimate where we should find a feature $\hat{p}_3 \in f_3$ to complete the collinear triplet. Then we compare each estimated $\hat{p}_3$ with each real $p_3, \forall\, \hat{p}_3, p_3 \in f_3$, using nearest neighbor matching in high-dimensional spaces (implemented in the FLANN library [6]), to find which pairs $(p_1, p_2)$ have a $p_3$ in $f_3$ that completes a space-time collinear triplet.

Once we have these possible traces of three points we calculate the straight line that joins them. Then, we evaluate the remaining frames to find out whether they have features on this line. If the moving object is detected in all frames, a feature in each frame will be found that satisfies the equation of the line. In practice, it is very difficult to find the same feature in all frames, however, if a feature is found in a few frames we can be pretty sure that this feature

[1] When moving at constant speed, the distance traveled between two frames is proportional to the time difference of these frames.

corresponds to a moving object. Typically, four frames are enough to have very good certainty of a positive discovery. We repeat this process a many times (approximately the total number of frames) with random triplets of frames, and merge the found traces if they belong to the same object (see Algorithm 1).

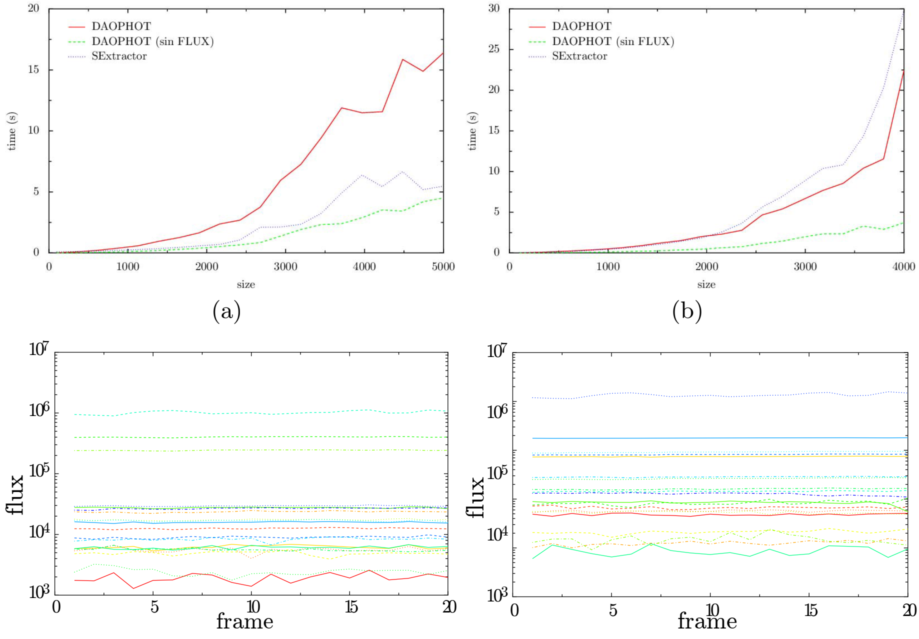
**Algorithm 1.** Pseudocode to find collinear traces of features.

```
$traces \leftarrow \emptyset$
for $i=0$ to $\#frames$ do
    $frame_1, frame_2, frame_3 \leftarrow$ random_sample($frames$, 3)
    $factor \leftarrow \frac{t_{f3} - t_{f2}}{t_{f2} - t_{f1}}$
    $triplets \leftarrow$ get_linear_triplets(features($frame_1$), features($frame_2$),
    features($frame_3$), factor)
    foreach $triplet \in triplets$ do
        $line \leftarrow$ fit_line($triplet$)
        $trace \leftarrow \{triplet[1], triplet[2], triplet[3]\}$
        foreach $frame \in frames \setminus \{frame_1, frame_2, frame_3\}$ do
            foreach $feature \in frame$ do
                if $feature \propto line$ then
                    trace $\leftarrow trace \cup \{feature\}$
            if $\exists\ trace' \in trace \cong trace'$ then
                merge($trace, trace'$)
            else
                $traces \leftarrow traces \cup \{trace\}$
return $traces$
```

As a result we obtain sets of traces, i.e. features that are aligned in space and time corresponding to objects moving in a straight line at constant speed. The equations for finding the line that best fits each trace of feaures form an overdetermined system, so we use the least squares technique. This line gives an estimation of the position and the velocity of the moving object.

## 3 Results

In this section we present some results of this work. First a comparison between DAOPHOT and SExtractor methods is considered. Figure 5 shows the performance of each method and the results of the photometry reduction (flux estimation). Figure 5(a) shows the execution time of each method as a function of the image size with a constant number of stars (300 aprox.). Figure 5(b) is the same with constant star density (20 stars for each 100×100 pixels). As can be seen, if the density of star is constant, the behaviour of the DAOPHOT method is better than that of SExtractor. Figures 5(c) and 5(d) show the photometry computation (flux estimation) using DAOPHOT and SExtractor, respectively. In both cases, flux estimation does not have large variations, allowing the detection of the point light sources in different frames.



(a) (b)

(c) (d)

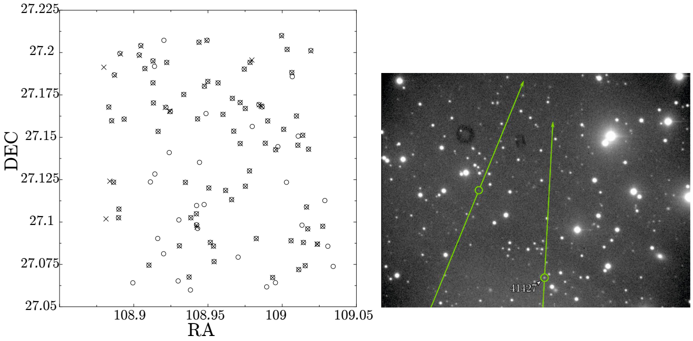
**Fig. 5.** Execution time comparison between DAOPHOT and SExtractor: (a) for different image size and constant amount of stars, and (b) for different image sizes and constant star density. Flux estimation using (c) DAOPHOT and (d) SExtractor.

For the case of real images acquired with a telescope, very satisfactory results are obtained. As test data set, we use image sequences containing moving objects from the CASLEO Observatory located in the Leoncito, San Juan, Argentina. Figure 6(b) shows the result of applying our proposed method to an image sequence within the data set corresponding to an observation of the asteroid 41427. This asteroid was correctly detected and another object (on the left side) was also found that is virtually imperceptible to the human eye.

To calculate the transformation between image coordinates and sky coordinates we use Astrometry.net [7]. Once we get the sky coordinates of the detected celestial bodies, we can compare the results with USNO-B1 catalog. In Figure 6(a) the positions of the detected and the catalogued celestial bodies, marked with circles and crosses, respectively, are plotted. In most cases they coincide, indicating a good alignment of the image. Then, we can conclude that mobile objects' sky coordinates were calculated accuratly.

## 4 Conclusions

In this work we present a method that automates the detection of moving objects in the night sky. To extract the locations of the point light sources we analyze



(a) (b)

**Fig. 6.** (a) Positions (in sky coordinates) of the detected and the catalogued celestial bodies, marked with circles and crosses, respectively. (b) Discovered moving objects in the image sequences from CASLEO data set, one of them is the asteroid 41427.

both DAOPHOT and SExtractor methods and compare them. We propose an algorithm for moving object detection in image sequences, which operates looking for collinearity in the sets of the detected point sources. Experimentation with synthetic and real images shows that this method is very effective for the detection of asteroids and even other faint objects. Morevover, it can be used with a robotic telescope to achive an autonomous system for astronomical discoveries.

## References

1. Tody, D.: The iraf data reduction and analysis system. In: 1986 Astronomy Conferences, International Society for Optics and Photonics, pp. 733–748 (1986)
2. Starck, J., Murtagh, F.: Astronomical image and data analysis. Astronomy and astrophysics library. Springer (2002)
3. Bertin, E., Arnouts, S.: SExtractor: Software for source extraction. Astronomy & Astrophysics, Supplement, 393–404 (June 1996)
4. Stetson, P.B.: Daophot: A computer program for crowded-field stellar photometry. Publications of the Astronomical Society of the Pacific, 191–222 (1987)
5. Zhang, Z.: Iterative point matching for registration of free-form curves and surfaces. International Journal of Computer Vision (2), 119–152 (1994)
6. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: International Conference on Computer Vision Theory and Application, VISSAPP 2009, pp. 331–340. INSTICC Press (2009)
7. Lang, D., Hogg, D.W., Mierle, K., Blanton, M., Roweis, S.: Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. The Astronomical Journal (5), 1782 (2010)

# Recognising Tabular Mathematical Expressions Using Graph Rewriting

Mohamed Alkalai

School of Computer Science, University of Birmingham
M.A.Alkalai@cs.bham.ac.uk

**Abstract.** While a number of techniques have been developed for table recognition in ordinary text documents, very little work has been done on tables that contain mathematical expressions. The latter problem is complicated by the fact that mathematical formulae often have a tabular layout themselves, thus not only blurring the distinction between table and content structure, but often leading to a number of possible, equally valid interpretations. However, a reliable understanding of the layout of a formula is often a necessary prerequisite to further semantic interpretation. In this paper, a graph representation for complex mathematical table structures is presented. A set of rewriting rules is applied to the graph allows for reliable re-composition of cells in order to identify several valid table interpretations. The effectiveness of the technique is demonstrated by applying it to a set of mathematical tables from standard text books that has been manually ground-truthed.

## 1 Introduction

The matrices of cells could be considered as the simplest tables: There are no spanning cells through columns or through rows. The borders of all the cells are marked by the ruling lines. This kind of tables is easy to recognise using the graphic ruling lines. However, due to the lack of standard convention of composing tables, not all tables follow such a distinction. As for the physical layout, it can be noted often the presence of cells that spread over several lines or several columns, and sometimes the borders of neighbouring cells are even misaligned. Also, in the majority of cases, the borders and the rules of a table are not marked by the graphic lines.

To characterize the table structure for various domains of documents, a flexible framework representation is necessary. The table's syntactic layout and the semantic structure must be depicted. While the information about the physical layout can contribute to table re-composition, the logical structure can be used to extract the table's content for re-use purposes.

Ramel et al. [6] analyse the most two well-known table representation systems (which are introduced by World Wide Web Consortium (W3C) and Advancement of Structured Information Standards (OASIS)) that are used to represent tables and find that they share the same deficiencies. First, the representation of irregular physical layouts are difficult. The poorly aligned borders of cells are

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 124–133, 2013.

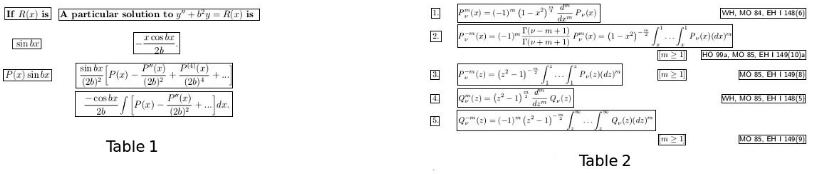


**Fig. 1.** Cell identification with tables containing multiline expressions that are taken from [5]

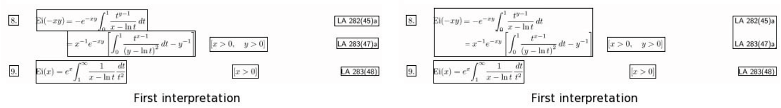


**Fig. 2.** Two different interpretations of a single table that is taken from [5]

not allowed and improvised solutions are provided for the spanning cells. Finally, limited means are supplied for the description of the logical structure of a table.

The aim of this work is to develop a table recognition algorithm that is particularly good for tables containing mathematical expressions. As the distinction between tables and complex typeset mathematical formulae spanning multiple lines is often difficult, the narrow definition of tables is forgone and instead consider a far wider range of expressions as tables as is usually the case in the literature.

The tabular form in which many mathematical expressions are being presented can often lead to ambiguities in the interpretation to what essentially constitutes a table component (i.e., a column, row or cell). While in table understanding of ordinary text tables [7], [8], the goal is generally to restrict a result to a single valid interpretation, for mathematical tables these ambiguities can lead to several possible valid interpretations. Therefore, the aim of the proposed recognition procedure is to produce as a result the set of possible valid interpretations.

Since, as mentioned above, that there is no standard convention of composing tables and that there is a need of building table representation framework which is flexible enough to deals with tables from various domains. Therefore, The framework that is illustrated in section 3 is constructed based on abstract concepts that allow for producing the maximum cells, columns and rows which can be extracted from table form. Graph rewriting rules are also introduced in this framework to selectively utilized for recomposing table's cells. The nature of the proposed framework gives the opportunity to used it on different table structures from various area of sciences like Literacy, Mathematics, Physic, Chemist...etc.

## 2 Interpretation of Mathematical Tables

While some tables, for example, Cayley tables in abstract algebra, are quite straight forward to recognise due to their easy tabular composition and sometimes clear separation of rows and columns with bars, this is generally not the case. In fact, the common absence of any vertical or horizontal bars as well as the complexity of formulae often spanning multiple lines make it not only difficult to identify the cell structure but can lead to a number of different interpretations for the same table, which are often equally valid.

Figure 1 presents two tables taken from [5] with a fairly conservative column and row layout. There is indeed a unique ideal interpretation for Table 1, consisting of two columns and three rows, where the cell in the lower right hand corner contains a math expression spanning two lines. In addition, given the difference in font weights one could even interpret the first line as a clear header row.

Table 2 on Figure 1 is less straightforward given the overlapping expressions in the second line. However, one can still come up with a unique interpretation of five rows and three columns. However, due to the overlap of the formulae which are effectively in a column of their own, it is difficult to obtain this interpretation automatically.

Figure 2 presents a clipping from a more complex table also taken from [5]. Here it is possible to see two different interpretations, both with their own merits. While both interpretations regard the basic table as consisting of four columns, the first interpretation results in three rows, using the formula names on the right hand side as header column. The second interpretation on the other hand uses the enumeration in the first column as header. Obviously there are still more interpretations: For example, one with three columns with the middle column containing complex formulae or even one with only two columns, where the right column contains named multiline formulae that possibly even be considered as subtables.

This gives not only rise to the problem of finding a method that can yield a number of possible valid interpretations, but also the need to finding an adequate grid structure to represent such tables holding the different interpretations and to give a means to re-compose the recognised cells.

## 3 Multi-interpretations of Table's Re-composition

In this section a description of the proposed method is given. The input of the technique is the bounding box of table cells which are extracted by the method presented in [1]. Using these cells, the algorithm first produces the maximum columns and also the maximum cells in each column that can be extracted from a table. This is further described in the preprocessing steps section below. Then an initial graph that represents the table grid and the relationship between their nodes (cells) is defined and built. Also a new set of graph rewriting rules that are used to produce all possible valid interpretations is illustrated. Finally, experimental results on 150 tables are shown and evaluated.

### 3.1 Preprocessing Steps

Several definitions are given below to formally describe the concepts of how the maximum columns and cells from tables are extracted. The first definition is for the Bounding Box of the cell component $C$:

**Definition 1 (Bounding Box).** *Let $c$ be a cell, then the borders of its* bounding box *are defined by $l(c), r(c), t(c), b(c)$ representing left, right, top and bottom limit respectively where $l < r$ and $t < b$*

Before building a graph to represent table structure, all cells $C$ are first sorted ascendantly using $l(c)$. Then, initial columns are constructed by splitting $C$ on the cell that does not horizontally overlap with all cells which are above it.

**Definition 2 (Initial Columns).** *Let $C = \{c_1, c_2, ....c_n\}$ be all cells ordered such that $l(c_1) < l(c_2)$ and col be a column of table. Then $col = \{c_1, c_2, ....c_m\}$ if one of the set $[r(c_1), r(c_2), r(c_3), .....r(c_m)] < l(c_{m+1})$ where $n = 1, 2, ...$ and $m < n$*

In case there is an absence of cell which should be beneath or above a cell that is being checked using the step described above, a virtual cell $c'$ is added. When the graph is built later, these virtual cells are represented as nodes. The goals of adding such nodes to the graph are firstly to avoid the complex relationship between nodes and secondly to use such nodes to detect actual rows. Some examples are shown in Figure 3

In order to locate the position of virtual cells, the definition 3 in [1] is recalled to calculate the borders of lines and then use them to detect if there is an existence of a line which has no corresponding cell (belong to a particular column) vertically overlapped with its borders. If so, a virtual cell is added.

**Definition 3 (Virtual cells).** *Let $col = \{c_1, c_2, ....c_n\}$ be a column and $l = \{g_1, g_2, ....g_n\}$ be a line such that if $b(c_1) =< t'(l_1) || t(c_1) >= b'(l_1)$ where $t'(l_1) = \min_{g \in l_1} t(g)$ and $b'(l_1) = \max_{g \in l_1} b(g)$ then add a virtual cell $c'_1$.*

### 3.2 Tabular Representation Using Graph Model

The total cells which are produced from the above steps are utilized to build an initial graph that represents the table structure. Each node $N$ in this graph

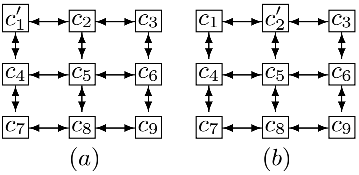


**Fig. 3.** Examples of virtual cells which their borders appear to be bigger

$G$ which corresponds to a cell $c$ has four edges $E$ with four directions where $l, r, t$ and $b$ are labelled left, right, top and bottom edge direction respectively (an exception is for border nodes which might have only two or three edges). Also, there must be an existence of all possible first degree connections between nodes $N$. The first degree connections mean here the edges that directly connect a node $n$ with its adjacent nodes.

**Definition 4 (Graph Specifications).** *Let $n$ be a node which represents a cell $c$, then the directions of its outgoing edges are defined by $l(n), r(n), t(n), b(n)$ representing left, right, top and bottom directions respectively. Let $n'$ be directly adjacent node to $n$ at any direction such that every $l(n)$ there exists of $r(n')$ and likewise for $r(n), t(n), b(n)$.*

**1) Type of Nodes:** When constructing the initial graph, one can divide the nodes to four types. The classification process is done by checking whether there is a horizontal overlap between columns or not. Table 1 shows the node types and how they are treated by the interpretor.

**Table 1.** Type of nodes

| Node Types | Definition |
|---|---|
| $R$ | is a real node which must not be merged with other nodes from other columns |
| $V$ | is a virtual node which must not be merged with other nodes from other columns |
| $R^*$ | is a real node which can be merged with other nodes to form one of the possible valid table interpretations |
| $V^*$ | is a virtual node which can be merged with other nodes to form one of the possible valid table interpretations |

**2) Type of Relationships between nodes (Edges):** To avoid having complex relationships between nodes, a graph which represents the maximum number of nodes for a table is constructed. This provides us with simple relationship between nodes which are horizontal and vertical edges.

### 3.3 Construct Initial Graph (Example)

The graph in Figure 4 represents Table 2 in Figure 1 which illustrates one possibly valid interpretation of the table. As can be seen, the proposed algorithm succeeded in distinguishing the second column that represents equations from the misaligned third column that represents the conditions associated with these equations and eventually splits them to two columns.

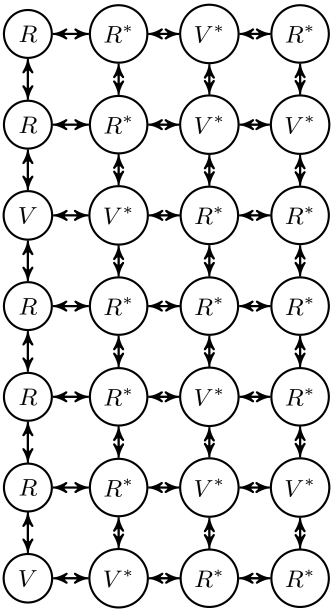


**Fig. 4.** A graph represents one of the possible interpretations of the table's columns shown in table 2 on the right of Fig 1

### 3.4 Graph Rewriting Rules

Graph rewriting rules are composed to represent structural information of table form. The graph defined above is used to represent them. Although, in [3] and [2] the authors have used graph rewriting rules before to analyse table layout, due to the complex structure of the table domain that are used in the experiments, new rules are produced.

Let $N$ and $E$ represent a specific set of nodes and a specific set of relationships between nodes called edges respectively. Then, A graph rewriting rule can

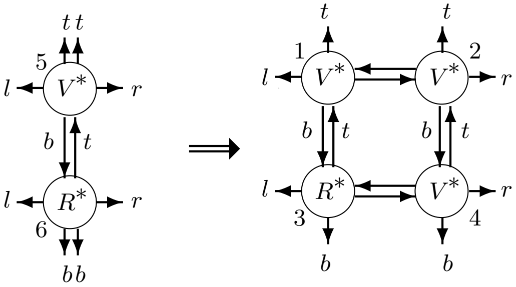


**Fig. 5.** Example of production rule

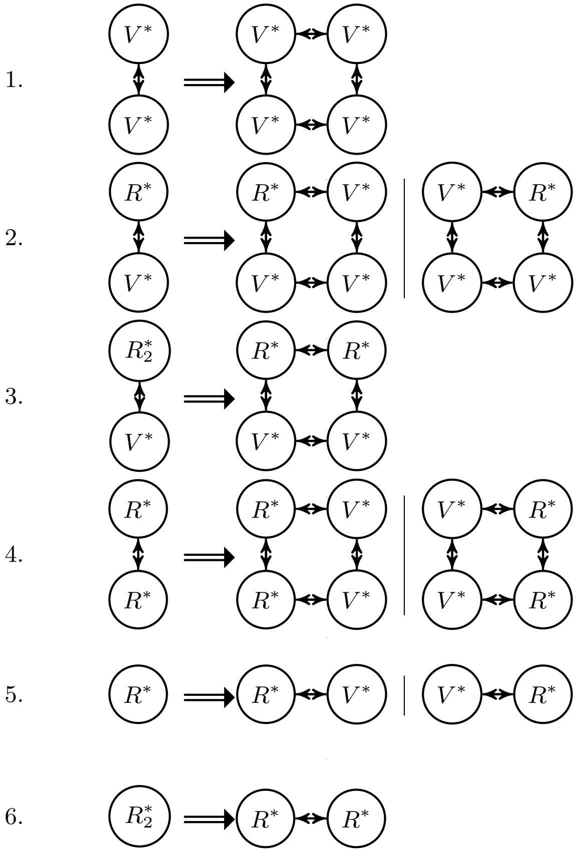


**Fig. 6.** Full production rules

be represented by the following tuple $g = \{N, E, P\}$ where $P$ are rewriting production rules which has the form $lhs \rightarrow rhs$, this specifies two graphs where the subgraph $rhs$ in a host graph $(G)$ can be replaced with a graph $lhs$. The embedding relations $ER$ associated with each rewrite rule $lhs \rightarrow rhs$ specify how the new subgraph $lhs$ is connected to the remainder graph of the host graph G, after $rhs$ is removed. The notation containing four-tuples of the form $\{(n_1,e_1,n_2,e_2)$: $n_1,n_2 \in N$; $e_1,e_2 \in E\}$ is used to represent embedding relations $ER$. Figure 5 shows an example of production rule.

Embedding rule $ER$ which tells edge label conversion from $rhs$ to $lhs$ for the production rule showed in Figure 5 is expressed as follows:

$$ER = ((1V^*, l, 5V^*, l), (1V^*, t, 5V^*, t), (2V^*, r, 5V^*, r),$$
$$(2V^*, t, 5V^*, t), (1V^*, b, 5V^*, b), (3R^*, t, 6R^*, t),$$
$$(3R^*, l, 6R^*, l), (4V^*, r, 6R^*, r), (3R^*, b, 6R^*, b),$$
$$(4V^*, b, 6R^*, b))$$

### 3.5 Full Table Production Rules

A sample of the production rules that are needed to represent all possible table interpretations are shown in Figure 6. Due to the pages number limitation, illustration of all rules which fully cover the different cases of node combinations is not possible. By using these rules, one can produce all possible table interpretations.

## 4 Evaluation and Experimental Results

To accomplish the table recognition evaluation, preparing table ground-truthing is usually needed. In [4] the author stated that, in some cases, the researchers who are ground-truthing tables might have different opinions about the right way of ground-truthing a table. Sometimes, several interpretations seem to be justifiable and appear to be equally valid. Taking into account this fact and for evaluation purposes, 150 tables were manually ground-truthed, such that, each table has all possible interpretations that can be extracted from it. To facilitate the comparison procedure, a visual technique is designed which allows us to visually assess the table recognition output. The technique draws rectangles around table cells. Each column's cells are given a unique colour to their rectangles. Experiments are done using 100 pages taken from [5] which contains 150 tables. Table 2 illustrates concise information about the experimental results. In this table, the 150 tables are categorised to three groups based on the number of possible interpretations that can be obtained from a table. This can be accomplished using the ground-truth tables. A comparison between outputted table possible interpretations and the corresponding ground-truth table is then manually done. The results of this comparison are classified into three categories. This is determined by observing how far one possible interpretation of the table from the proposed system matches one possible interpretation of the table according to the ground truth set. These three categories are: 1) Table interpretations that completely and correctly extracted. An output table is classified under this category if it 100% matches. 2) Table interpretations partially extracted. Here the

**Table 2.** Results of applying the proposed table interpretation technique on 150 tables

| No. of Tables | Ground Truth Table Dataset | Output Table Dataset | | |
|---|---|---|---|---|
| | Number of Possible Table Interpretations | No. of Tables Interpretations Completely and Correctly Extracted | No. of Tables Interpretations Partially Extracted | No. of Tables Interpretations That are missed |
| 82 | 2 | 124 | 26 | 14 |
| 65 | 4 | 141 | 56 | 63 |
| 3 | 7 | 6 | 7 | 8 |

matching rate is approximately between 75% and 95%. 3) Tables interpretations that are missed. In this cases, the matching rate is 0%.

### 4.1 Analysis of Table Interpretations That Are Partially Extracted or Missed

Although the experimental results — presented in Table 2 — show already a promisingly high accuracy rate, there is still a considerable problem with the mis-clustering of some cells to the wrong column (in the case of table interpretations that are partially extracted) and the failure of splitting two columns (in the case of table interpretations that are missed). An analysis of these cases yields that the majority of mis-clustering and failure cases are due to the preprocessing step when it fails to assemble the cells into their proper columns. One possible approach to tackle this kind of problem and eventually improve the accuracy of the proposed approach is to manually intervene by adding marks on tables which assist the proposed method in inferring the correct column and as a result, extract the all possible valid table interpretations. Figure 7 illustrates how adding marks on tables improves the accuracy rate.

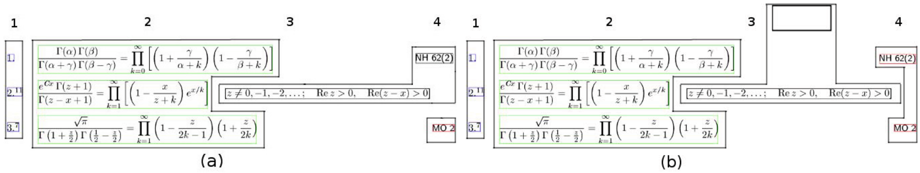


**Fig. 7.** Example of table re-composition (a) before manual intervention (b) after manual intervention

The Figure 7 shows tables before and after the manual intervention. Each column in this table is bordered and given a number. By observing table (a) it can be clearly seen that last cell in the first row is wrongly clustered to the third column where it should have been gathered with the cells in fourth column. Table (b) shows a solution of the problem by adding an empty rectangle over the third column to tell the proposed method that the last cell in the first row is not overlapped with all cells in third column and therefore, it should be gathered with cells in fourth column.

## 5 Conclusion

The proposed framework introduced in this paper was built accordingly upon the observation of a wide range of tabular forms which occur in many documents from different domains. The abstract components of this framework can be used as basis of wide range of other applications of document recognition.

The technique represented here is able to produce several interpretations of table form. Unlike other table representation techniques, the proposed approach has the capability to deal with misaligned columns that sometimes appear in tabular mathematical components. Adding virtual nodes to the initial graph prevents complex relationship between nodes and would contribute in deciding the actual table's rows. Using the described production rules allow for producing more than one possible valid interpretations of table structure. The experiments in 150 tables show promising results.

## References

1. Alkalai, M., Sorge, V.: Issues in mathematical table recognition. In: Conferences on Intelligent Computer Mathematics (CICM 2012), MIR Workshop (2012)
2. Amano, A., Asada, N.: Graph grammar based analysis system of complex table form document. In: ICDAR, pp. 916–920. IEEE Computer Society (2003)
3. Cooperman, R., Armon Rahgozar, M.: A graph-based table recognition system. In: SPIE Proc., pp. 192–203 (1996)
4. Hu, J., Kashi, R.S., Lopresti, D.P., Wilfong, G.T., Nagy, G.: Why table ground-truthing is hard. In: ICDAR, pp. 129–133 (2001)
5. Jeffrey, A., Zwillinger, D.: Table of Integrals, Series, and Products. Elsevier Inc. (2007)
6. Ramel, J., Crucianu, M., Vincent, N., Faure, C.: Detection, extraction and representation of tables. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. 1, pp. 374–378. IEEE Computer Society, Washington, DC (2003)
7. Costa Silva, A., Jorge, A.M., Torgo, L.: Design of an end-to-end method to extract information from tables. International Journal Document Analysis Research 8(2), 144–171 (2006)
8. Zanibbi, R., Blostein, D., Cordy, J.R.: A survey of table recognition: Models, observations, transformations, and inferences. Int. J. Doc. Anal. Recognit. 7(1), 1–16 (2004)

# Using Graph Theory to Identify Aberrant Hierarchical Patterns in Parkinsonian Brain Networks

Rafael Rodriguez-Rojas[1-2], Gretel Sanabria[3], Lester Melie[3], Juan-Miguel Morales[1], Maylen Carballo[1], David Garcia[4], Jose A. Obeso[4], and Maria C. Rodriguez-Oroz[4]

[1] International Centre for Neurological Restoration, Havana, Cuba
{rafael,jm,maylen.carballo}@neuro.ciren.cu
[2] "Abdus Salam" International Centre for Theoretical Physics, Trieste, Italy
[3] Cuban Neurosciences Center, Havana, Cuba
{lmelie,gretels}@cneuro.edu.cu
[4] Neuroscience Area, Applied Medical Research Centre, University of Navarra, Spain
{jobeso,dagarcia,mcoroz}@unav.es

**Abstract.** The topology of complex brain networks allows efficient dynamic interactions between spatially distinct regions. Neuroimaging studies have provided consistent evidence of dysfunctional connectivity among the cortical circuitry in Parkinson's disease; however, little is known about the topological properties of brain networks underlying these alterations. This paper introduces a methodology to explore aberrant changes in hierarchical patterns of nodal centrality through cortical networks, combining graph theoretical analysis and morphometric connectivity. The edges in graph were estimated by correlation analysis and thresholding between 148 nodes defined by cortical regions. Our findings demonstrated that the networks organization was disrupted in the patients with PD. We found a reconfiguration in hierarchical weighting of high degree hubs in structural networks associated with levels of cognitive decline, probably related to a system-wide compensatory mechanism. Simulated targeted attack on the network's nodes as measures of network resilience showed greater effects on information flow in advanced stages of disease.

**Keywords:** Brain networks, MRI, graph theory, morphometric connectivity.

## 1 Introduction

The human brain is considered to be one of the most complex systems in nature, structurally and functionally organized into complex and sparse networks. The topology of networks allows efficient dynamic interactions between spatially distinct brain regions, which are thought to provide the physiological basis for high-level information processing [1]. Efforts to understand its intricate wiring patterns and the way these give rise to normal functioning and connectivity abnormalities in neurological and psychiatric disorders, is one of the most challenging areas in modern science.

The mathematical framework of Graph Theory provides powerful tools to deal with intrinsic complexity of brain systems, allowing the extraction of global metrics

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 134–141, 2013.

that capture various aspects of the network's topological organization. However, graph theoretical approaches in neurosciences deals with large and complex neural systems that have revealed non-random but small-world architectures, providing regional specialization with more efficient rates of information transfer [1-3]. This hypothesis has been supported in structural and functional human brain networks studies, over a wide range of scales in space and time.

Small-world networks are characterized by the existence of a small number of nodes with higher connectivity degree, referred to as hub-nodes. Hubs are suggested to play an important role in the overall network organization and can be defined several possible measures of centrality, including degree (number of edges) and betweenness centrality [1]. Detecting hub-regions in a network helps to identify relevant structures subserving specific roles such as motor and cognitive processing, thus providing a link between structure and function [4]. Progress in Graph Theory, combined with advanced neuroimaging techniques like Magnetic Resonance Imaging (MRI), allow us to quantify topological properties of brain systems like basal ganglia – thalamus – cortical circuitry and disturbed functioning that give rise to movement disorders such as Parkinson's disease (PD). Previous functional brain network studies have demonstrated disruption of several large scale brain systems in PD [5-7]. Up to know remains unclear how the affected modular organization of brain network underlies motor and cognitive impairment in PD.

Morphometric-based connectivity has been recently introduced as a measure of structural association between brain regions [8-10]. This concept is defined as the covariance between two anatomical brain areas. Structural networks can then be constructed from morphometric correlations of anatomical metrics like cortical volume, thickness, and surface area. In the present study, we constructed structural networks using average cortical thickness of atlas-based regions, to explore the characteristics of the cortical networks in PD across subgroups at different stages of cognitive impairment, compared to healthy subjects. For the first time we applied graph theoretical approaches to investigate alterations in large-scale morphological brain networks, nodal centrality and network robustness in this neurological pathology.

# 2 Methods

## 2.1 Patients and Controls

This research was approved by the Ethics Committee for Medical Research at the Clinica Universidad de Navarra in Spain. All patients provided their written informed consent. All the participants underwent a neuro-psychological assessment, including the Mini-Mental State Examination (MMSE) for global cognitive functions and UPDRS-III scale for motor disabilities. Demographic and clinical data for the study groups are given in Table 1. PD patients were classified in three groups according to cognitive performance: cognitively normal (PDCN); PD with mild cognitive impairment (PDMCI), based on MCI criteria [11]; and with dementia (PDD), based on the DSM-IV-TR manual of mental disorders [12].

**Table 1.** Demographic and clinical characteristics of the study participants

| | HC | PDMCI | PDCN | PDD | test |
|---|---|---|---|---|---|
| No. | 20 | 22 | 28 | 18 | |
| Sex (M/F) | 11/9 | 15/7 | 15/13 | 7/11 | N.S[a] |
| UPDRS III | N.A | 32.3±8.5 | 35.0±12.2 | 50.0±10.0 | P < 0.01[b] |
| MMSE | 29.2±1.1 | 29.0±1.4 | 26.4±2.6 | 18.3±3.8 | P < 0.001[b] |

N.S: no significant; [a]Chi-square test; [b]Oneway analysis of variance

## 2.2 MRI Acquisition and Cortical Thickness Measurement

MRI examinations were performed on a 1.5 T Magnetom Symphony scanner (Siemens, Erlangen, Germany). All subjects were investigated with a whole brain T1-weighted coronal oriented Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence (repetition time TR = 13 ms; echo time TE = 10 ms; inversion time TI= 1100 ms; flip angle =15; 1 mm isotropic resolution; slice gap = 0 mm). Head motion was minimized with restraining foam pads provided by the manufacturer.

Reformatted T1-weighted MR images were processed using Freesurfer 5.0.0 software package (Massachusetts General Hospital, Harvard Medical School; freely available at http://surfer.nmr.mgh.harvard.edu). Figure 1 (1 to 4) summarizes Freesurfer pipeline, whose technical details have been previously described [13]. After segmentation into gray and white matter, the gray/white and the gray/pial interfaces were tessellated and labelled according to Destrieux sulcogyral-based atlas, which includes 74 regions per brain hemisphere [14]. Cortical thickness, defined as the shortest distance between white and corresponding pial surfaces, was computed for every region. A linear regression was performed at every region to remove the effects of age, gender, age–gender interaction, and mean cortical thickness. The residuals of this regression were then substituted for the raw cortical thickness values.

## 2.3 Graph Theoretical Approaches

The morphometric network is modeled as an undirected graph, $G_{brain} = [N, W]$ (figure 1.5). $N$ is a set of n=148 nodes determined by the anatomical parcellation and represents the voxels having a non-zero probability of belonging to the cortical tissue. $W$ is the set of $w_{ij}$ edges between each pair of regions $i$ and $j$. We computed $w_{ij}$ values as the Pearson's product-moment correlation coefficient in corrected thickness values across subjects, removing the influence of all other regions $n \neq (i, j)$. This resulted in a pair of {74 x 74} correlation matrices. Pearson's correlation was adopted instead of partial correlation analysis because the number of nodes exceeds the number of patients. Unweighted binary graphs were generated by thresholding the $w_{ij}$ values based upon the significance of the correlations. Bootstrapping samples (Nboot = 300 samples) of the connectivity matrix were obtained by selecting a random subset of the total number of subjects with replacement to compute the correlation coefficient.

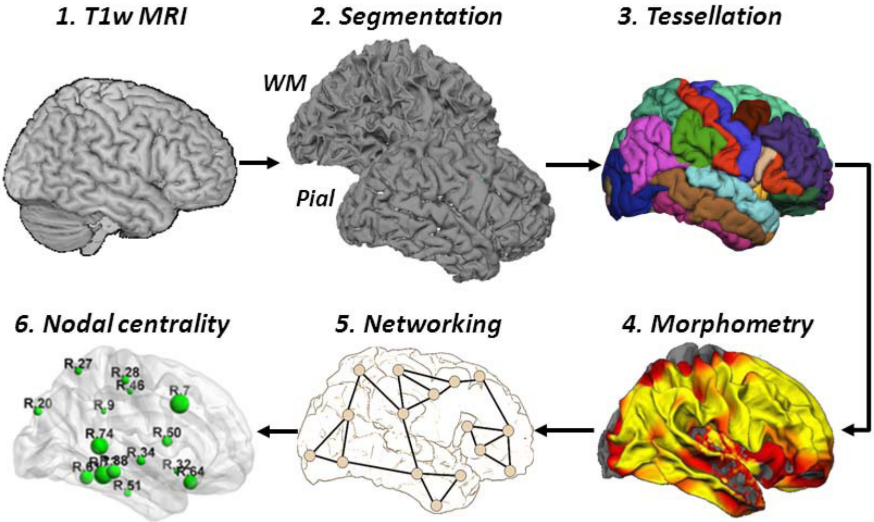


**Fig. 1.** Pipeline for morphometric-based graph analysis. 1. Acquisition of T1-weighted high resolution MRI; 2. Surface-based segmentation; 3. Atlas-based tessellation and labeling; 4. Calculation of corrected cortical thickness; 5. Schematic representation of the brain network in the form of a graph; 6 Definition of higher-degree connector hubs.

## 2.4 Nodal Centrality and Network Robustness

The shortest path $d_{ij}$ between any two vertices *i* and *j* is defined as the number of edges along the geodesic length connecting them [3]. Degree centrality of a given node *n(i)* is defined as the number of edges incident to the node. The 'betweenness centrality' *B(i)* of a *n(i)* is a global centrality measure of the influence of a node over information flow between other nodes in the network [3]. We measured the normalized betweenness as:

$$B(i) = \Sigma_{j\neq k} \{ n_{jk}(i)/ n_{jk} \} \qquad (1)$$

where $n_{jk}$ is the number of shortest paths connecting *j* and *k*, and $n_{jk}(i)$ is the number of these paths passing through i. The hubs are the regions with higher values of *B(i)* as seen in figure 1.6. To test differences between groups a nonparametric Kruskal Wallis (KW) statistical test was used, with Bonferroni correction for multiple comparisons.

Small-world networks show a high robustness to random failure of nodes, but are known to be vulnerable to target attack on the hubs [1]. A fault in the system is the removal of any n nodes and all edges connected to these nodes from $G_{brain}$. To evaluate the attack tolerance of each of the four networks, we removed the nodes and edges from the graph in decreasing order of their betweenness and then measured the changes in the size of the largest connected component.

# 3 Results

## 3.1 Nodal Characteristics

Figure 2 shows the strongest hubs in the four sets of undirected graphs, corresponded to healthy volunteers and patients with different levels of cognitive decline. In the control group, regions with *B(i)* > 2 (meaning that these hub regions have at least 2 times the network's average betweenness centrality) included right primary sensorimotor and posterior cingulate areas, and associative temporal regions. Compared with controls, the PD patients showed significant centrality decreases in primary motor cortex, while increases in associative and limbic frontal and occipito-temporal areas are observed (KW test, p<0.01). PDD's hubs were predominant in the occipital and parietal regions, with tendency to lose involvement of fronto-temporal areas. Nomenclature of human cortical gyri and sulci can be found in Destrieux et al [14]. Full list of anatomical regions with respective betweenness centrality values are available under request.

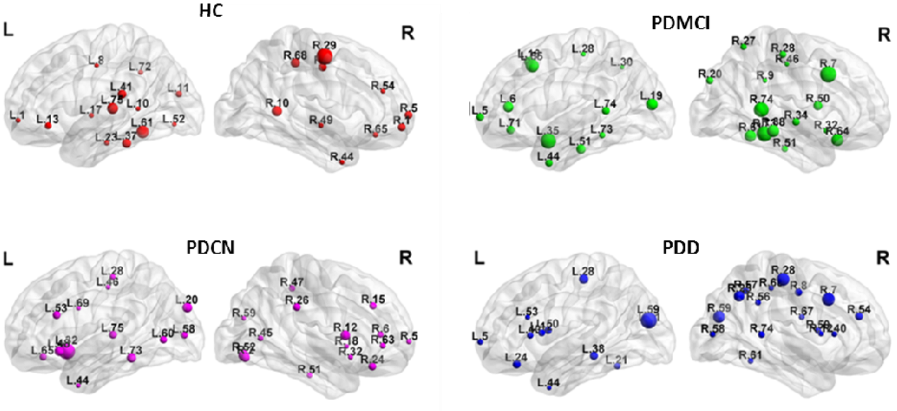


**Fig. 2.** The structural network cores for each group. Size of spheres indicates normalized betweenness centrality values of each region.

## 3.2 Reduced Network Robustness in PD

We find that the deletion of connector hubs have distinctly effects on the small-world attributes as a consequence of pathological stages. Figure 3 shows the networks robustness in response to the targeted attack. PDD group was considerably more vulnerable to hubs deletion, with reduction of the largest connected component when at least 15% of the most central nodes and links were removed, and remains noticeably reduced for all thresholds. The structural networks of patients without dementia (PDCN and PDMCI) were as robust as that of controls until the 57% of the most central nodes were attacked. In the range when 57 to approximately 70% of nodes are deleted, these three networks show a cross-linked behavior against attacks. Since that sparsity threshold, resilience to targeted failures are consistent with cognitive decline (HC > PDCN > PDMCI > PDD).

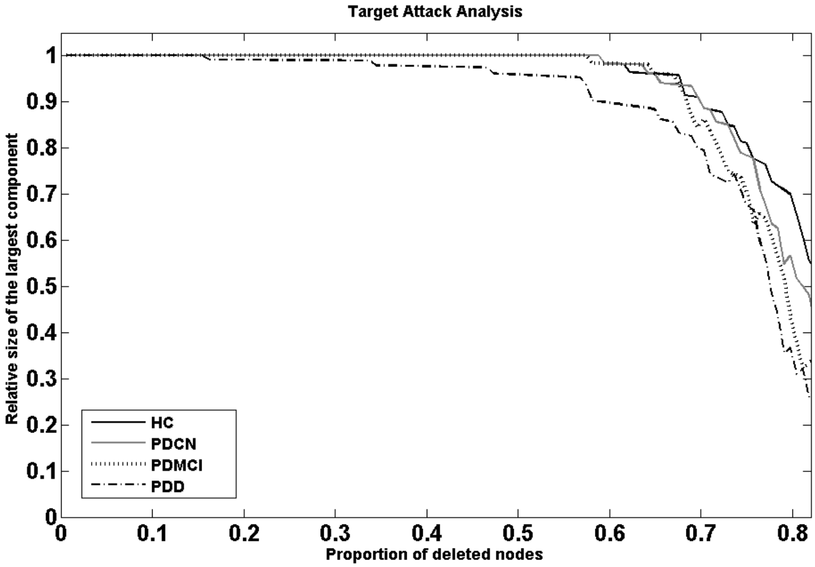


**Fig. 3.** The graph shows the largest component size of the networks for every group as a function of sparsity threshold. As the proportion of removed nodes increases, the largest component sizes of all groups tend to decrease. The arrow indicates the lowest sparsity threshold (15%) in which all the networks included all connected nodes.

## 4 Discussion

To our knowledge, this is the first time that graph theory is used to explore the morphological networks in PD and its relation with cognitive decline. We have considered the hypothesis that these covariation patterns reveal information about the dynamics of the brain networks in response to degenerative processes in PD. We have also modeled the vulnerability to targeted attack on the network's hubs in relation to cortical thinning and cognitive impairment.

### 4.1 Altered Nodal Centrality

Our results point out the degree and distribution of network's hubs as possible biological markers of deficits in cognitive and behavioral functions in PD. The loss of integrative capacities of the precentral regions may reflect altered output through basal ganglia-thalamo-cortical loops, which is consubstantial with PD [15]. The selective damage to high-degree hubs in structural networks should have an outsized impact on the capacity of the network for efficient high-level processing. This could explain the early emergence of motor and cognitive symptoms in the course of PD. During the course to more advanced phases of cognitive impairment, clustering of connector hubs shift to posterior parietal, temporal and occipital regions, including visual and auditory cortices, and to associative and limbic frontal areas (figure 2). This observation fits with the heavy reliance of PD patients on sensory modalities to guide their actions. Such reconfiguration leads us to speculate that alteration in degree centrality across the brain circuitry may be indicative of system wide compensatory mechanism, in response to the basal ganglia altered output arising from imbalances of dopamine.

In terms of network dynamics, the shift in *B(i)* suggests a reordering in the control of flow of information. However, it is difficult to differentiate between changes resulting from the disease itself as opposed to those that arise as part of a compensatory response. On the other hand, betweenness only takes into account shortest edges, while long range network connections also contribute to global communication patterns. Future studies are necessary to address network-wide integration and its effect over network's efficiency. Our results are in line with recent studies suggesting reduced sensorimotor connectivity and increased functional connectivity in associative and limbic circuits in PD [5-7, 16].

### 4.2 Topological Vulnerability in PD

Measures of network resilience may be computationally simulated by targeted attack on the network highest-degree nodes. The vulnerability of the network in different stages of disease may then be quantified by comparing its topological or dynamical behavior after the "lesioning". Our observations suggest that pathological attacks on high-centrality nodes have greater effects on information flow in advanced stages of PD than attacks on early phases and healthy controls. These results are consistent with recent inferences about the association between disease stages and thinning of core prefrontal, cingulate, temporal and parieto-occipital regions in PD [17]. More importantly, graphs corresponding to normal or middle cognitive impairment show a tendency to recover resilience capacity after an attack to a high percent of connector hubs, similar to healthy controls. Therefore, this PD related changes in centrality parameters may reflect a less optimal reconfiguration in hierarchical network topology in response to alteration of primary motor and cognitive circuits. Thus, topological organization of network's hubs could provide associations for the understanding of the relationship between network topology and neuropathological state of disease.

## 5 Conclusions

In the present paper we have shown that combining graph theory and MRI data allows studying the organizational properties of the morphological networks in Parkinson Disease at different stages of cognitive decline. This approach should yield more comprehensive understanding of how structural disruptions in the brain network architecture are associated with functional deficits in PD. Our findings are compatible with the notion that cognitive impairment in PD is associated with disruptions in the integrity of large-scale interconnected brain systems. The graph theory analysis also provides a new way to understand the pathophysiology of specific functional deficits and, possibly, to evaluate disease progression. In a near future, the combination of functional and morphometric-based connectivity in a graph theory framework could explain the nature of dynamical processes taking place on the parkinsonian networks, as well as the causality between network topology and network dynamics.

## References

1. Bullmore, E., Sporns, O.: The economy of brain network organization. Nat. Rev. Neurosci. 13, 336–349 (2013)
2. He, Y., Evans, A.: Graph theoretical modeling of brain connectivity. Curr. Opin. Neurol. 23, 341–350 (2010)
3. Stam, C., Reijneveld, J.: Graph theoretical analysis of complex networks in the brain. Nonlinear Biomedical Physics 1, 3 (2007)
4. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. USA 104, 36–41 (2007)
5. Tessitore, A., Esposito, F., Vitale, C., Santangelo, G., Amboni, M., Russo, A., Corbo, D., Cirillo, G., Barone, P., Tedeschi, G.: Default-mode network connectivity in cognitively unimpaired patients with Parkinson disease. Neurology 79, 2226–2232 (2012)
6. Wu, T., Long, X., Wang, L., Hallett, M., Zang, Y., Li, K., Chan, P.: Functional connectivity of cortical motor areas in the resting state in Parkinson’s disease. Hum. Brain Mapp. 32, 1443–1457 (2011)
7. Stoffers, D., Bosboom, J.L., Deijen, J.B., Wolters, E.C., Stam, C.J., Berendse, H.W.: Increased cortico-cortical functional connectivity in early-stage Parkinson’s disease: An MEG study. Neuroimage 41, 212–222 (2008)
8. He, Y., Chen, Z., Evans, A.: Small-world anatomical networks in the human brain revealed by cortical thickness from MRI. Cerebral Cortex 17, 2407–2419 (2007)
9. Sanabria-Diaz, G., Melie-Garcia, L., Iturria, Y., Aleman, Y., Hernandez, G., Valdes, L., Galan, L., Valdes-Sosa, P.: Surface area and cortical thickness descriptors reveal different attributes of the structural human brain networks. Neuroimage 50, 1497–1510 (2010)
10. Yao, Z., Zhang, Y., Lin, L., Zhou, Y., Xu, C., Jiang, T.: Abnormal cortical networks in mild cognitive impairment and Alzheimer’s disease. PLoS Comput. Biol. 6, e1001006 (2013)
11. Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.: Mild cognitive impairment- beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J. Intern. Med. 256, 240–246 (2004)
12. American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, Washington, DC (1994)
13. Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D.: Automatically parcellating the human cerebral cortex. Cereb Cortex 14, 11–22 (2004)
14. Destrieux, C., Fischl, B., Dale, A., Halgren, E.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15 (2010)
15. Obeso, J.A., Rodriguez-Oroz, M.C., Benitez-Temino, B., Blesa, F.J., Guridi, J., Marin, C., Rodriguez, M.: Functional organization of the basal ganglia: therapeutic implications for Parkinson’s disease. Mov. Disord. 23, S548–S559 (2008)
16. Sharman, M., Valabregue, R., Perlbarg, V., Marrakchi-Kacem, L., Vidailhet, M., Benali, H., Brice, A., Lehericy, S.: Parkinson’s Disease Patients Show Reduced Cortical-Subcortical Sensorimotor Connectivity. Movement Disorders 28, 447–454 (2013)
17. Zarei, M., Ibarretxe, N., Compta, Y., Hough, M., Junque, C., Bargallo, N., Tolosa, E., Martí, M.J.: Cortical thinning is associated with disease stages and dementia in Parkinson’s disease. J. Neurol. Neurosurg. Psychiatry 84, 875–881 (2013)

# Crack's Detection, Measuring and Counting for Resistance's Tests Using Images

Carlos Briceño, Jorge Rivera-Rovelo, and Narciso Acuña

Universidad Anahuac Mayab
krlozgod@gmail.com, {jorge.rivera,narciso.acuna}@anahuac.mx

**Abstract.** Currently, material resistance research is looking for biomaterials where mechanical properties (like fatigure resistance) and biocompatibility are the main characteristics to take into account. To understand the behavior of materials subject to fatigue, usually we analyze how the material responds to cyclic forces. Failures due to fatigue are the first cause of cracks in materials. Normally, failures start with a superficial deficiency and produce micro cracks, which grow until a total break of the material. In this work we deal with the early detection of micro cracks on the surface of bone cement, while they are under fatigue tests, in order to characterize the material and design better and more resistant materials according to where they would be applied. The method presented for crack detection consists in several stages: noise reduction, shadow elimination, image segmentation and path detection for crack analysis. At the end of the analysis of one image, the number of cracks and the length of each one can be obtained (based on the maximum length of crack candidates). If a video is analyzed, the evolution of cracks in the material can be observed.

## 1 Introduction

Tests with new biomaterials like bone cements with monomers of amino group, should be conducted in similar conditions to the real use; in the case of bone cements, it is very important because they are used inside the human body (implants, prosthesis). Particularly important is the analysis of the resistance of the material, and the study of the material under stress (by external forces applied on it).

To understand the behavior of materials subject to fatigue, usually we analyze how the material responds to cyclic forces. Such forces can cause cumulative damage in the material, and depending on the intrinsic properties of the material, as well as to external factors which can be under control in laboratory experiments, the service life of the material could be reduced.

Fatigue is a kind of failure observed in materials under dynamic and fluctuant forces. Such failure can be observed even in cases where the force is below the resistance threshold; they can appear suddenly and can be catastrophic. Failures due to fatigue are the first cause of cracks in materials.

Normally, failures start with a superficial deficiency and in conditions where the local force induced is greater than the resistance value of the weaker grain,

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 142–149, 2013.

or microstructural barriers. In most cases, the superficial fault results in one or more micro cracks, which can be observed with a microscope. The micro crack start growing by discontinuity points in the material, which concentrates the efforts.

Crack detection and tracking of its growth in materials like bone cements, gives useful information about early stages in fatigue damage; this kind of damage is similar to the one the bones suffer in daily activities. Such information can be used to develop better materials (ie. more resistant materials).

Prosthetic bone cement can be used in orthopedics and odontology; it is an acrylic resin used to fix the prosthesis to the bone [1]. This kind of cement is used in orthopedics for hip, knee or shoulder surgery (for example, to replace by a prosthesis), as well as in spinal surgery and dental prosthesis. In such surgery, the bone cement is used to fill the spaces or holes between the (metal) prosthesis and the bone cavity where it should be fixed. Currently, we can find commercially bone cements with different characteristics like viscosity (high, low, extra-low), or concentration (20g, 40g, 50g, 60g), and we choose among them depending of the application.

According to the norm ASTM E206 described in [2], fatigue is a structural and progressive change, located and persistent, which occurs in materials subject to efforts and fluctuating deformations, which can produce micro-cracks or even total rupture of the material after a sufficiently large number of fluctuations. Fatigue can also be described as a progressive fail which occurs due to crack propagation until they reach an unstable size. For this reason, we should put attention to the materials used in the bone cement and also to its applications, particularly if it implies repeated and fluctuating forces. Fatigue causes failures because of the simultaneous action of cyclic and strain (tension) stress, as well as plastic deformation.

The goal of the analysis of the growth of (micro) cracks, is to understand the mechanisms of the beginning and growing of cracks governing early stages of serious damage in bone cement, which are manufactures with monomers of amino group in a matrix of methyl methacrylate.

In this work we deal with the early detection of micro cracks on the surface of bone cements, while they are under fatigue tests, in order to characterize the material and design better and more resistant materials according to where they would be applied.

We use a microscope and a conventional camera in order to obtain some images, which are analyzed to detect crack clusters, identify crack paths, and to count the number of cracks in the image.

## 2 Detection of Cracks

We can deal with crack detection by means of several approaches; for example, we can use probabilistic or stochastic theory [3,4], continuous models [5] or deterministic Markov processes [6]. However all of them deal only with crack detection and does not analyze the growth of the crack, which needs to follow the crack paths during time.

Some characteristics of the cracks are their color and their width; cracks have a darker color than plastic deformations, scratches and grain boundaries. Therefore the threshold calculated assure that only cracks are detected, also cracks are wider than grain boundaries so if a grain boundary is detected as a crack, the difference can be observed as it would be a discontinuous line (dotted line).

The method presented in this paper allows the analysis of crack growing or crack evolution due to its ability to get not only the crack clusters, but also the number and lengths of paths in the image. Figure 1 shows the general scheme of the method.

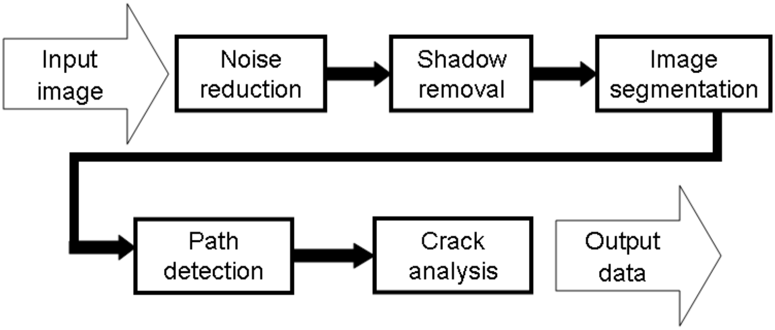


**Fig. 1.** General scheme of the method

A low pass frequency filter is used for noise reduction/elimination (Gaussian low pass filter), together with a median filter for elimination of salt and pepper noise (if it is present). Once the noise on the image has been reduced, it could be applied a method to reduce some effect on the boundaries of the images; this effect is called *shadow*. The shadows can result in erroneous detection of cracks on the boundaries of the image. To eliminate shadows, a histogram smoothed by a Gaussian kernel with bandwidth $B$ can be used to calculate a threshold, and pixels with gray values over the thresholds are changed for such value.

After the stage of preprocessing the image, we need to classify the image pixel into regions; that is, we segment the image (assign every pixel to a particular segment). Given a pixel, we can determine if it belongs to a segment or to other one by comparing its gray value with a threshold. The threshold value in step 2 is calculated in such a way that the resulting value can minimize the variance of every segment, and at the same time maximize the variance between segments [7]; that is, we compute the ratio between the two variances and choose as the threshold the value which maximizes that ratio. The weighted within-class variance is given by Eq. (1), while the the class variances are given by Eq. (2), the class probabilities are given by Eq. (3) and the means are given by Eq. (4). $P(i)$ is the probability of the gray value $i$.

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \tag{1}$$

$$\sigma_1^2(t) = \sum_{i=1}^{t}[i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)} \qquad \sigma_2^2(t) = \sum_{i=t+1}^{I}[i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)} \tag{2}$$

$$q_1(t) = \sum_{i=1}^{t} P(i) \qquad q_2(t) = \sum_{i=t+1}^{I} P(i) \tag{3}$$

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{q_1(t)} \qquad \mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{q_2(t)} \tag{4}$$

Once we have the segmentation, the crack clusters are detected as neighbor pixels with values under certain threshold; such pixels are considered as vertex of a directed graph. The adjacency of two vertex is determined with the adjacency of the pixels: if they are horizontal or vertical neighbors, they are connected with an arrow of length 1; if they are diagonal neighbors, they are connected with an arrow of length $sqrt(2)$ (see fig. 2). Then, the arrow lengths are modified adding a factor equal to the difference in gray values of the adjacent vertex (connected pixels). Finally, a method to find minimum length paths is used in order to build the paths in each crack cluster (considering that the cracks are associated with the darkest gray values of the pixels).

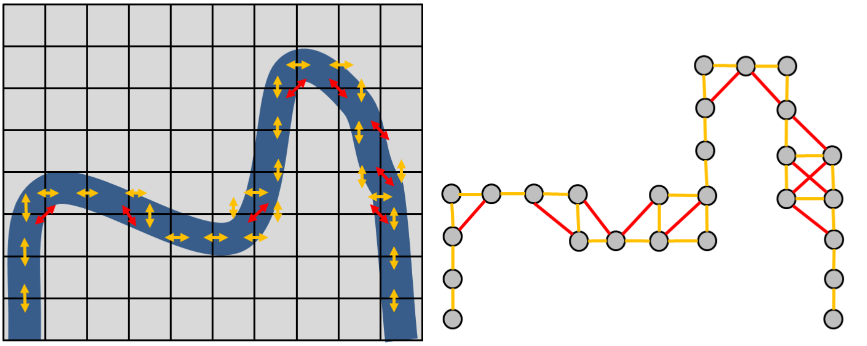


**Fig. 2.** Graph creation example. Supose there is a curve (crack pixels) like the blue one in the rigth image. The length of arrows is first assigned according to adjacency (yellow arrows are of length 1, while red arrows are of length $sqrt(2)$).

## 3 Experimental Results

Figure 3 shows an original image of a section of the surface of the material subject to strain efforts (obtained with a microscope with 200x of amplification), as well as the detected cracks. The threshold used in this case for the shadow elimination was 132 (gray value), which was obtained as the maximum increment of the blurred histogram of the image, with a bandwidth of $B = 30$. The number of cracks detected is 762, and the length of the longest crack is 179.41 pixels.

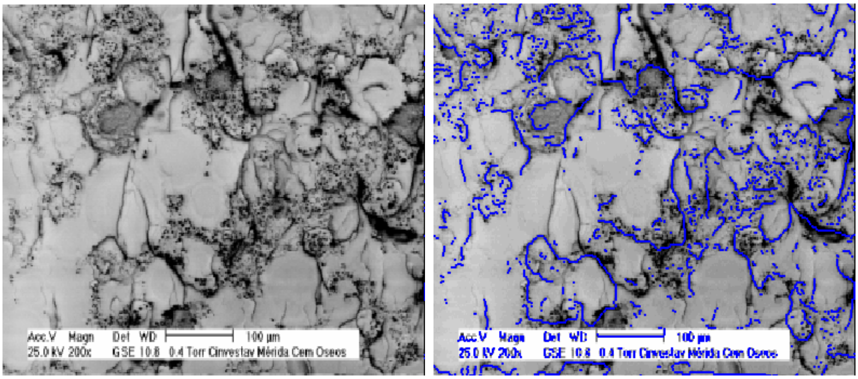

Acc.V Magn Det WD 100 µm
25.0 kV 200x GSE 10.8 0.4 Torr Cinvestav Mérida Cem Oseos
Acc.V Magn Det WD 100 µm
25.0 kV 200x GSE 10.8 0.4 Torr Cinvestav Mérida Cem Oseos

**Fig. 3.** Crack detection in a section of the surface of the material subject to assay under strain. a) Original image; b) Cracks detected (blue pixels).

Figure 4 shows the results with different bandwidth for the Gaussian filter. According to our experiments, the best value for the bandwidth of the Gaussian kernel used is $B = 35$, because in average it produces a threshold which allows a better identification of the cracks in the images.

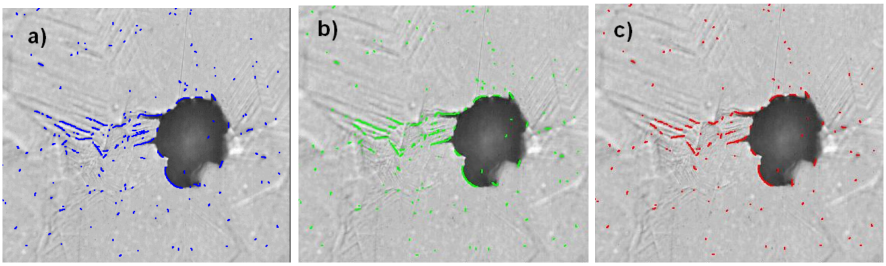


**Fig. 4.** Results obtained using different bandwidth for the Gaussian kernel used in filtering. Bandwidth: a) 30, b) 35, c) 40.

The method was applied to a set of 200 images (some of bone cement and others of steel). Two methods for removing the shadows in the images were applied. The first one applies an exponential decreasing value to the pixels of the border of an image if their mean gray level value is 10 units greater than the mean gray level value of the inner pixels of the image (this method is called shadow removal). The second one is the median filter. Figure 4 shows the number and size of the cracks detected (given in pixels) for a set of 40 images, comparing the results using the shadow removal method and using a median filter. According to the results, the median filter gives more accurate detection of cracks, in comparison with the crack detection using the shadow remove method instead; on average, if the shadow remove method is applied, fewer cracks are detected than

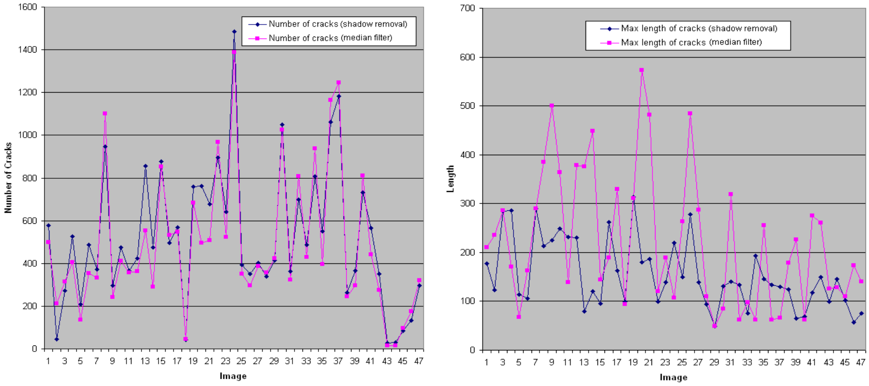


**Fig. 5.** Results obtained with some bone cement images. Observe that the number of cracks is bigger with the median filter, and the lengths are bigger with shadow remove (because some cracks are joined together).

using the median filter but that is because that method sometimes erroneously joins two or more cracks and because of that larger cracks are detected too.

Figure 6 shows why the median filter was the best option, you can observe than the median filter almost detect the complete hole in the middle of the image while shadow remove joins one crack at the left of the image with the crack at the bottom creating a big crack that goes through the hole. Using the median filter a threshold of 84 was calculated, 133 cracks were detected and the length of the largest cracks detected was 56.698. Using shadow remove a threshold of 158 was calculated, 155 were detected and the length of the largest crack detected was 173.509.

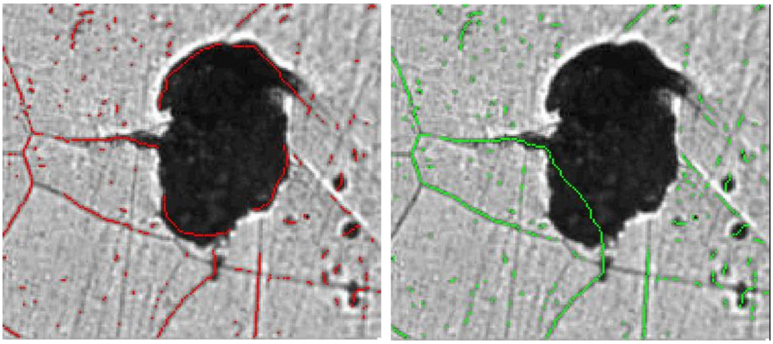


**Fig. 6.** Left: results using the median filter for removing the shadows (the cracks detected are in red color). Rigth: results using the method *shadow remove*, that is exponential decreasing gray value assignment (the cracks detected are in green color).

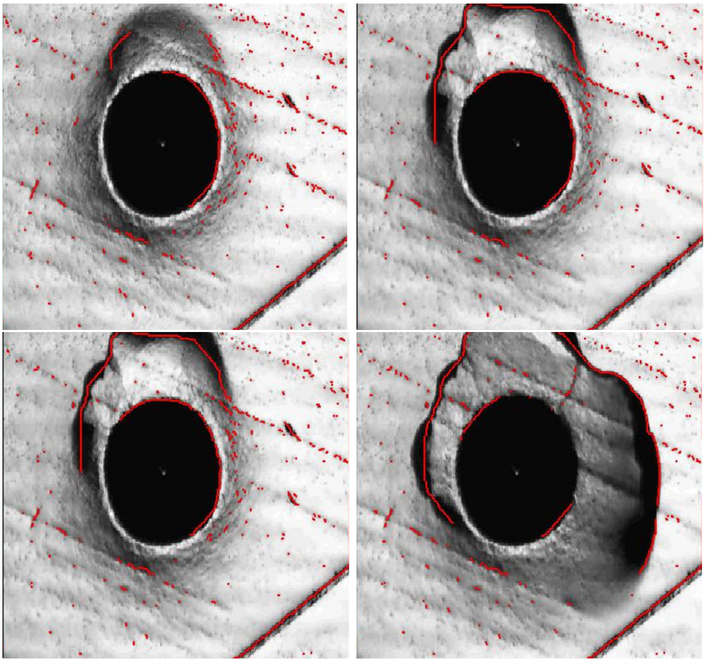


**Fig. 7.** Evolution of micro cracks detected in steel assay images

Figure 7 shows the detection of cracks in a steel assay subject to strain efforts; the method is applied to the set of images taken from the video of the microscope, and we can track the evolution of the cracks.

## 4 Conclusion

The method described can detect micro cracks in images of materials like bone cement under fatigue efforts. To accomplish such task several steps are needed, from image denoising to crack path calculation. The noise reduction and the shadow elimination, are of particular importance because otherwise, misclassification of pixels occurs.

The method presented has some limitations. One of them is that it cannot detect cracks in the form of trees because the crack are detected as one continuous line, and some cracks can be erroneously detected at the begging of the cracking process because the cracks are not dark enough yet.

Even that in about 93% of the images analized we were able to eliminate shadows correctly, there are some cases where different cracks are detected as one crack (they are erroneously joined), and other cases where one crack is divided into two cracks. We are analyzing how can we improve the accuracy of the method when detecting cracks by improving the removing of the shadows.

Another thing to work on, is that grain boundaries are sometimes confused with cracks; however this can be easily identified because the cracks are small in length, have a darker gray value and are thicker than grain boundaries; that is, visually the boundaries can be observed like a discontinuous crack (like a dotted line).

**Acknowledgment.** The authors would like to thank to Dr. Narciso Acuña, researcher of the materials' properties, for his valuable contributions for this work. We also would like to thank to Universidad Anahuac-Mayab for supporting this work.

## References

1. Rosell, G., Mendez, J.: Bone cement: prevention of exposure to its components. Technical notes. National Institute of Health and Safety at Work. Spain (2009)
2. Quesada, F., Charris, J., Perez, J.: Ensayos de fatiga en viga rotativa para determinar la Constante de Miner del acero Aisi 1045. Prospectiva 6(2) (2008)
3. Nicholson, D., Ni, P., Ahn, Y.: Probabilistic theory for mixed mode fatigue crack growth in brittle plazes with random cracks. Engineering Fracture Mechanics 66, 305–320 (2000)
4. Meyer, S., Bruckner-Foit, A., Moslang, A., Diegele, E.: Stochastic simulation model for microcracks in a martensitic steel. Computational Materials Science 26, 102–110 (2003)
5. Heron, E., Walsh, C.: A continuous latent spatial model for crack initiation in bone cement. Applied Statistics 57, 25–42 (2008)
6. Chiquet, J., Limnios, N., Eid, M.: Piecewise deterministic Markov processes applied to fatigue crack growth modelling. Journal of Statistical Planning and Inference 139(5), 1657–1667 (2009)
7. Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on System, Man, and Cybernetics SMC-9(1) (1979)

# Accuracy to Differentiate Mild Cognitive Impairment in Parkinson's Disease Using Cortical Features

Juan-Miguel Morales, Rafael Rodriguez,
Maylen Carballo, and Karla Batista

Neuroimages Processing Group, International Center for Neurological Restauration

**Abstract.** Mild cognitive impairment (MCI) is common in Parkinson's Disease (PD) patients and it is key to predict the development of dementia. There is not report of discriminant accuracy for MCI using based-surface cortical morphometry. This study used Cortical-Thickness (CT) combined to Local-Gyrification-Index (LGI) to assess discriminant accuracy for MCI stages in PD. Sixty-four patients with idiopathic PD and nineteen healthy controls (HC) were analyzed. CT and LGI were estimated using Freesurfer software. Principal Component Analysis and Lineal Discriminant Analysis (LDA) assuming a common diagonal covariance matrix (or Naive-Bayes classifier) was used with cross-validation leave-one-subject-out scheme. Accuracy, sensibility and specificity were reported to different classification analysis. CT combined to LGI limited revealed the best discrimination with accuracy of 82,98%, sensitivity of 85.71% and specificity of 80.77%. A validation process using independent and more heterogeneous data set and further longitudinal studies, are necessary to confirm our results.

**Keywords:** Naive-Bayes classifier, PCA, Accuracy, Parkinson's disease, MCI, Cortical Thickness, Cortical Folding, LGI, MRI, Surface-based morphometry.

## 1 Introduction

In Parkinson's disease (PD) exist a spectrum of cognitive dysfunction, ranging from mild cognitive impairment (MCI) to dementia (PDD). MCI is common in non-demented PD patients and predicts the development of dementia in PD patients over a long period of time [1,2]. Specific patterns of gray matter atrophy occur across all stages of PD and functional and metabolic changes also are measurable, but it is too early to determine their utility as biomarkers for cognitive impairment in PD [3,4,5]. Therefore, additional evidence is necessary and validation of biomarker candidate as an objective method of diagnosis and prognosis is an active research field nowadays.

Medical imaging is widely used for above purpose and a general approach is to detect subtle differences in the composition, morphology or other behavior in organs and relating these differences to clinical phenomena of interest[6]. In

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 150–157, 2013.

particular, surface-based morphometry has been used to identify pattern of atrophy associated to cognitive decline in PD patients[3]. However, only a few of then have considered PDMCI stage [7,8]. A recently research found that disease stage in PD was associated with thinning of the medial frontal region and discriminant analysis showed that mean cortical thickness and hippocampus volume have 80% accuracy in identifying PD patients with dementia [8]. However, it remains unclear how cortical changes is related to cognitive impairment and disease stage in PD, in addition, as far as we know, not any study report accuracy of cortical folding and cortical thickness for identifying PDMCI stage. In this study we used based-surface morphometry for contributing with additional evidence about associated cortical regions to cognitive dysfunction and to assess accuracy of cortical thickness combined with cortical folding for discriminating PDNC and PDMCI stages.

## 2 Methods

### 2.1 Patients and Controls

This study enrolled 64 patients with idiopathic PD and 19 healthy controls (HC). All the participants underwent an extensive neuropsychological assessment, including the Mini-Mental State Examination (MMSE) and Blessed Dementia scale for global cognitive functions. In order to evaluate motor disabilities at PD patients, the motor subset of the Unified Parkinson Disease Rating Scale (UPDRS-III) and the Hoehn and Yahr scale were applied. Significant co-morbidity at PD patients and controls were excluded by neurological and psychiatric evaluation, imaging and laboratory tests. Demographic and clinical data for the study groups are given in Table 1. PD patients were classified in three groups according to cognitive performance: cognitively normal PD patients (PDCN), PD with mild cognitive impairment (PDMCI), based on established MCI criteria[9] and PD with dementia (PDD); based on the criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR)[10]. All the participants provided informed consent for the study in accordance with Helsinki Declaration.

### 2.2 MRI Acquisition

MRI examinations were performed on a 1.5 T Magnetom Symphony MRI scanner (Siemens, Erlangen, Germany). All subjects were investigated with a whole brain T1-weighted coronal oriented Magnetization Prepared Rapid Gradient Echo (MPRAGE) sequence (repetition time TR = 13 ms; echo time TE = 10 ms; inversion time TI= 1100 ms; flip angle =15; 1 mm isotropic resolution; slice gap = 0 mm). Head motion was minimized with restraining foam pads provided by the manufacturer.

### 2.3 Cortical Variables Estimation

Cortical Thickness (CT) and Local Gyrification Index (LGI) estimation was performed with Freesurfer software, which is documented and freely available for download online (`http://surfer.nmr.mgh.harvard.edu/`). The technical details of these procedures were described in prior publications. Briefly, this processing included, removal of non-brain tissue using a hybrid watershed/surface deformation procedure[11], automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures [12,13], intensity normalization [14], tessellation of the gray matter white matter boundary, automated topology correction[15,16]. This method used both intensity and continuity information from the entire three dimensional MR volume in segmentation and deformation procedures to produce representations of cortical thickness, calculated as the closest distance from the gray/white boundary to the gray/CSF boundary at each vertex on the tessellated surface[17]. Local Gyrification Index was measured in each vertex as the ratio between areas of pial surface and an outer smoothed surface tightly wrapping the pial surface[18] using Matlab toolbox distributed with Freesurfer.

**Table 1.** Demographic and clinical characteristics of the study participants

| | HC | PDNC | PDMCI | PDD | Differences |
|---|---|---|---|---|---|
| No. Subjects | 19 | 21 | 26 | 17 | |
| Sex (M/F) | | 15/7 | 15/13 | | |
| Age | 68.0/3.1 | 67.0/7.0 | 71.6/3.8 | 73.2/7.3 | 0.001a |
| Education | 2.55/1.0 | 3.30/1.5 | 2.52/1.0 | 2.31/0.8 | N.S |
| Evolution (yr) | N.A | 12.4/3.6 | 14.5/6.1 | 13.9/4.7 | N.S |
| UPDRS III | N.A | 32.3/8.5 | 35.0/12.2 | 50.0/10.0 | P < 0.0a |
| HY | N.A | 2.54/0.6 | 2.89/0.7 | 3.78/0.7 | P < 0.001a |
| MMSE | 29.2/1.1 | 29.0/1.4 | 26.4/2.6 | 18.3/3.8 | P < 0.001a |

N.S: no significant; UPDRS : Unified Parkinson's Disease Rating Scale; H&Y: Hoehn and Yahr stage. (a) One way analysis of variance with Fisher LSD post-hoc comparisons

### 2.4 Features Extraction and Classification

Using general linear model (GLM) with Age and Gender as covariate nuisance with Freesurfer module MRI_GLMFIT and MATLAB scripts was investigated the regional difference patterns of CT and LGI between the different groups in pairs-wise analysis. Changes were examined with a threshold of $p<0.001$ (uncorrected) on the vertex level and $p<0.05$ (corrected for multiple comparison using Montecarlo simulation with 10,000 iterations) on the cluster level. Each one of the identified clusters expands to several cortical regions, using t-test ($p < 0.001$) we had determined cortical regions with significant difference in average value of each variable using Destrieux Atlas (a 148 regions atlas)[19], once eliminated, age and gender, confounding. CT was smoothed using a Gaussian kernel of 15 mm FWHM.

All contrast was evaluated to select those that gave us more information to differentiate between PD and PDMCI in both directions; first, changes at topographic extension and second on the intensity of variation. Average value of CT and LGI for each significant region integrated the feature-vector. Principal Component Analysis (PCA) was used to identify a set of orthogonal modes that capture the greatest amount of variance expressed spatially by the two feature-vectors. We proceeded on selecting a number of modes that accounted to per-model variance of 80%. Lineal Discriminant Analysis (LDA) assuming a common diagonal covariance matrix (or Naive-Bayes classifier) with same prior probability to all group and cross-validation was performed, using the leave-one-subject-out scheme in all analysis. Accuracy, sensibility and specificity was reported to five different analysis: CT-only/selected-regions, LGI-only/selected-regions, CT & LGI/selected-regions and CT & LGI/all-cortical-regions and CT & LGI/selected-regions/random-assigning-group.

# 3 Results

## 3.1 Global Analysis

Whole-cortex average CT was 2.44/0.09, 2.35/0.19, 2.19/0.17, 2.0/0.22 in HC, PDNC, PDMCI and PDD group respectively. ANCOVA revealed a significant difference between all groups except PDNC vs PDMCI and correlation with Age ($p<0.05$, tukey-kramer to compensate for multiple comparisons), no difference was found in Gender. Whole-cortex average LGI was 2.81/0.11, 2.77/0.15, 2.69/0.11, 2,69/0.12 in HC, PDNC, PDMCI and PDD group respectively, significant difference between HC vs PD and HC vs PDD and significant difference between Gender was revealed (ANOVA, $p=0.05$, tukey-kramer multiple comparisons). To avoid any possible effects of Age and Sex, both variables were included as covariates in the further analysis. A significant correlation between CT and LGI was found using four groups data (Pearson r=0,33, $p=0.002$).

## 3.2 Regional Analysis

Table 2 shows the number of clusters and regions identified with significant difference to CT and LGI. CT revealed significant changes to every contrast. HC-relative contrasts showed a progressive thinning from PDNC to PDD con values of 8.29%, 9.11% and 11.95% respectively. Topographic extension included 4 *regions (G_ pariet_ inf-Supramar_ left, S_ postcentral_ left/right and S_ intrapariet_ and_ P_ trans_ right),* 54 and 124 respectively. PDNC-relative contrast (PDNC vs PDMCI and PDNC vs PDD) revealed relative changes of 15% and 10.98%, the first one revealed significant different in *G_ occipital_ superior_ left* and the last one a number of 30 regions. LGI showed only significant clusters for HC-relative contrasts. The principal difference on its is reflexed by topographic extension, 2 regions in HC vs PDNC *(G_ cuneus_ left_ and S_ parieto_ occipital_ left)* compared to 14 and 7 in HC vs PDMCI and HC vs PDD respectively. In according to above results, we selected 31 regions (11 left and 20 right) provided by PDNC vs PDMCI and PDNC vs PDD contrasts to form a feature-vector to CT variable . In a similar way a feature-vector to LGI variable was compound for average value of LGI in 16 regions (11 left and 5 right) provided by HC vs PDNC and HC vs PDMCI contrasts. Figure 1 shows statistical parametric maps highlighting significant clusters that contains the selected regions (more details in supplementary material).

**Table 2.** Number of clusters and regions with significant difference in pairs-wise analysis for CT and LGI variables

| | **CT** | | | **LGI** | | |
|---|---|---|---|---|---|---|
| Groups | NoC | NoR | % | NoC | NoR | % |
| *HC vs PDNC* | 7 | 4 | 8.29 | 2 | 2 | 6.72 |
| *HCvsPDMCI* | 18 | 54 | 9.11 | 10 | 14 | 6.30 |
| *HC vs PDD* | 6 | 124 | 11.95 | 6 | 7 | 6.41 |
| *PDNC vs PDMCI* | 3 | 1 | 15.00 | 0 | 0 | |
| *PDNCvsPDD* | 16 | 30 | 10.98 | 0 | 0 | |
| *PDMCI vs PDD* | 10 | 11 | 9.49 | 0 | 0 | |

NoC: Number of significant cluster (p=0.05 cluster-wise, p=0.001 to form cluster)

NoR: Number of regions in clusters (Destrieux Atlas 2009, 148 regions) with significant difference (p=0.001) according to average value.

%: Relative percent of variation between pairs of groups

### 3.3 Classification

The 80% of variance of feature-vector CT was explained by the first five principal components. MANOVA discriminated between groups ($p<0.05$, chisq=15.02, wilk's lamda=0.7) using this modes of variations. With feature-vector LGI was necessary the first four principal components, witch ones discriminate between groups too (MANOVA, $p<0.05$, chisq=28.87, wilk's lamda=0.5). Table 3 summarizes classification results of different analysis. Multivariate classification using combined modes of CT and LGI limited to selected regions revealed the best discriminant accuracy with 82,98% compared to remainder analysis. Using all cortical regions was obtained a accuracy of 72.34%, using CT variable only 65.96% and using LGI variable only the result was of 78.72%. Similar results showed the sensitivity (85,71%) and specificity (80.77%) values. Using 10 trials of random assigning to all subjects of the two groups accuracy result was 38.30%.

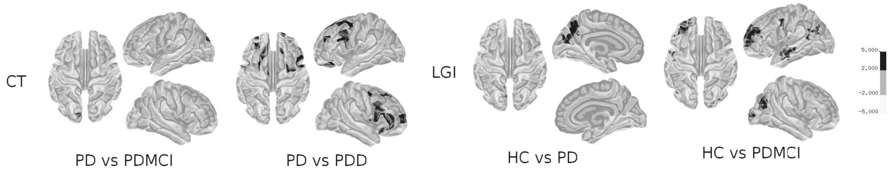


**Fig. 1.** Statistical parametric maps showing significant clusters on the four main contrasts selected to classification between PDNC and PDMCI *(Freesurfer MRI_ GLMFIT module and GLM, CT smoothed 15 mm FWHM, p=0.001 uncorrected and p=0.05 FWE cluster-wise*, corrected for multiple comparison using Montecarlo simulation with 10,000 iterations.

### 3.4 Discussion

We assessed cortical thickness combined with cortical folding accuracy for differentiate MCI in PD patients. In this first exploratory stage we have used LDA assuming diagonal covariance matrix or Naive-Bayes classifier on the basis of we have considered both CT and LGI variables normally distributed and independent each other within each group in accordance with the results of previous study[20], where no significant correlation were found between that variables in the control's group; In addition, feature-vector for classification was formed by first orthogonal modes of variation or principal components. However a comparison to a discriminative classifier, such as Logistic Regression or Support Vector Machine would be advisable to confirm the results. The used approach for discrimination not only captures univariate relationships of a single region across all subjects, but also detect multivariate relationships between different structures in each cortical variable[6]. CT showed a progressive reduction consistent with preview studies ([7,8]) and discriminate PDMCI with 65.96% of accuracy. In contrast with a recently study [8] and according to a previous one [20]LGI revealed structural changes between PDNC, PDMCI and PDD relative to control subjects. Figure 1 illustrate a extension of differences to others regions that should be associated to cognitive decline, that subtle differences between PDNC and PDMCI are no detected by univariate analysis, however, multivariate approach revealed difference between PDNC vs PDMCI of LGI with an accuracy of 78.72%. All classification results exceeded the accuracy obtained by chance (38.30%). By combining CT and LGI and using proposed regions we obtained the better accuracy (82,98%), similar to reported accuracy to differentiate dementia[8] . This results endorse the using of selected regions for classification. However, as the groups used in this study were recruited from a single clinical center, the results might be less generalizable to other clinical data and a validation process with more heterogeneous data sets is necessary, other lack is that we modeled a apparent progression of cognitive impairment using information relate to different contrasts obtained from cross-sectional design, this fact influences the results, witch ones should be confirm with longitudinal study fallowing quality criteria as were recommended recently [3,4] .

**Table 3.** Accuracy, sensitivity and specificity values to differentiate PDNC and PDMCI groups. Five different classification analysis are reported.

| Variables/Regions | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| CT and LGI/selected-regions | 85.71% | 80.77% | 82.98% |
| CT-only/selected-regions | 66.67% | 65.38% | 65.96% |
| LGI-only/selected-regions | 80.95% | 76.92% | 78.72% |
| CT and LGI/all-cortical regions | 71.43% | 73.08% | 72.34% |
| CT and LGI/selected-regions/randomly-assigning-group* | 28.57% | 46.15% | 38.30% |

*Naive-Bayes classifier with cross-validation leave-one-subject-out squeme for all analysis.*

(*) Average value resultant of ten trials of random assignations.

## 4 Conclusions

Our study supply additional evidence about existent relations between cognitive impairment and structural changes in brain cortex and reveal the capacity of cortical thickness and cortical folding to discriminate MCI, specially, when both features are combined and we use specific cortical regions, PCA and a Naive-Bayes classifier. A validation process using independent and more heterogeneous data set and further longitudinal studies are necessary to confirm our results.

## References

1. Pedersen, K.F., Larsen, J.P., Tysnes, O.B., Alves, G.: Prognosis of Mild Cognitive Impairment in Early Parkinson Disease: The Norwegian ParkWest Study. JAMA Neurology 70(5), 580–586 (2013)
2. Litvan, I., Aarsland, D., Adler, C., et al.: MDS task force on mild cognitive impairment in Parkinson's disease: critical review of PD-MCI. Mov. Disord. 26, 1814–1824 (2011)
3. Duncan, G.W., Firbank, M.J., O'Brien, J.T., Burn, D.: Magnetic resonance imaging: a biomarker for cognitive impairment in Parkinson's disease? Movement Disorders: Official Journal of the Movement Disorder Society 28(4), 425–438 (2013)
4. McGhee, D.J.M., Royle, P.L., Thompson, P.A., Wright, D.E., Zajicek, J.P., Counsell, C.E.: A systematic review of biomarkers for disease progression in Parkinson's disease. BMC Neurology, 35 (2013)
5. Litvan, I., Goldman, J., Tröster, A., Schmand, B., et al.: Diagnostic criteria for mild cognitive impairment in Parkinson's disease: Movement Disorder Society Task Force guidelines. Mov. Disord. 27, 349–356 (2012)
6. McGill University Montreal Quebec (CA), assignee. Systems and Methods of Clinical State Prediction Utilizing Medial Images Data. US-7899225-B2 (2011)

7. Pagonabarraga, J., Corcuera-Solano, I., Vives-Gilabert, Y., Llebaria, G., García-Sánchez, C., Pascual-Sedano, B., et al.: Pattern of regional cortical thinning associated with cognitive deterioration in Parkinson's disease. PloS One 8(1), e54980 (2013)
8. Zarei, M., Ibarretxe-Bilbao, N., Compta, Y., Hough, M., Junque, C., Bargallo, N., et al.: Cortical thinning is associated with disease stages and dementia in Parkinson's disease. Journal of Neurology, Neurosurgery, and Psychiatry (March 2013)
9. Winblad, B., Kivipelto, M., Jelic, V.: Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J. Intern. Med. 256(3), 240–246 (2004)
10. Association AP. Diagnostic and Statistical Manual of Mental Disorders, 4th edn. American Psychiatric Association, Washington, DC (1994)
11. Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., et al.: A hybrid approach to the skull stripping problem in MRI. NeuroImage 22(3), 1060–1075 (2004)
12. Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 33, 341–355 (2002)
13. Salat, D., Buckner, R.L., Snyder, A.Z., Greve, D.N., Desikan, R.S., Busa, E., et al.: Thinning of the cerebral cortex in aging. Cerebral Cortex 14, 721–730 (2004)
14. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 17, 87–97 (1998)
15. Fischl, B., Liu, A., Dale, A.M.: Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. IEEE Medical Imaging 20(1), 70–80 (2001)
16. Segonne, F., Pacheco, J., Fischl, B.: Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. IEEE Trans. Med. Imag. 26, 518–529 (2007)
17. Fischl, B., Dale, A.M.: Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences of the United States of America 97(20), 11050–11055 (2000)
18. Mea, S.: A surface-based approach to quantify local cortical gyrification. IEEE Trans. Med. Imag. 27, 161–170 (2008)
19. Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., et al.: Automatically Parcellating the Human Cerebral Cortex. Cerebral Cortex 14(1), 11–22 (2004)
20. Pereira, J.B., Ibarretxe-Bilbao, N., Marti, M.J., Compta, Y., Junqué, C., Bargallo, N., et al.: Assessment of cortical degeneration in patients with parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. Human Brain Mapping (September 2011)

# Performance Profile of Online Training Assessment Based on Virtual Reality:

## Embedded System versus PC-only

José Taunaí Segundo, Elaine Soares, Liliane S. Machado, and Ronei M. Moraes[*]

Federal University of Paraíba, João Pessoa/PB, Brazil
{taunai2,elaineanita1}@gmail.com, liliane@di.ufpb.br,
ronei@de.ufpb.br

**Abstract.** Training systems based on virtual reality are used in several areas of human activities. In some kinds of training is important to know the trainee's skills. It can be done in those systems but requires high-end computers to achieve good performance. Recently, the use of embedded systems connected to the training system was proposed for training assessment, with the goal of decreasing requirements of the main system. However, some questions are still open and a deep study of this proposal was not performed. This paper provides answers for some of those questions.

**Keywords:** Embedded Systems, Virtual Reality, Training Assessment, Fuzzy Naive Bayes, Possibility and Necessity Measures.

## 1 Introduction

Continuous learning and improvement of skills for staffs are a demand of several areas to guarantee good offer of services. With this purpose, applications based on virtual reality (VR) have been developed in order to provide realistic training, particularly in the medical area [3]. In those systems, users are exposed to simulated problems in 3D environments to practice and get technological and psychological skills to perform them in a real condition. Also, VR systems demand integration and synchronizations of routines, hardware and techniques [1], what requires high processing rates to provide real time feedback.

One of the main advantages of training in VR simulators is the possibility of monitoring user actions to register their movements. Then, information as force, position and acceleration, among others performed with interaction devices (as haptics) must be acquired and processed during all the training session. This information is used to feed assessment routines that must provide feedback about users skills. Thus, an online feedback [3] will demand expressive time of CPU, which can compromise the other tasks of the simulator. Moraes and Machado [2] proposed an architecture for assessment based on embedded systems. In this architecture, an embedded system is connected to the VR simulator to enhance the execution of the

[*] Corresponding author.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 158–165, 2013.

assessment tasks and release the CPU for other tasks related to the simulation. However, no further studies to analyze the efficiency of the architecture, its limitations to provide online assessment and its performance if compared to a CPU based approach were identified in the literature. The paper has as goal to provide this analysis for two previously proposed online methods based on fuzzy sets: Fuzzy Naive Bayes (FNB) [5] and Possibility and Necessity Measures (PMN) [14]. Both were implemented in a CPU and also in an embedded system. Monte Carlo simulations were used to describe profiles for both methods with increasing size databases.

## 2 Virtual Reality and Training Assessment

VR for training of procedures allows simulating real problems in a realistic way, avoiding risks and ethical issues [3], as the acquisition of guinea pigs or cadaveric bodies. Advantages of the use of VR for training are related to the variability of cases that can be simulated, including rare and atypical occurrences, the possibility of repetition without degradation of materials and the absence of risks for the people involved.

The use of special devices in VR systems allows reaching high levels of immersion and interactivity, providing for users the feeling of presence by the manipulation of elements in the virtual environment [8]. Those devices usually explore users´ senses as the sight, hearing and touch. The VR system processes all interactions and the feedback to be provided. Also, the VR system is responsible by the synchronization of all tasks in order to guarantee the sequence and coherence among the several tasks. These tasks include the calculus of physical phenomena, lighting and collision detection, as examples. Because interaction devices can acquire data with rates that start on 30Hz, the amount of information processed in a simulation can be massive, depending on the type of interaction. Haptic devices, as example, can capture interaction in rates between 500 and 1000 Hz.

Since VR systems are computational simulations, interaction data of the procedure can be collected and used to assess trainees´ performance. This can occur in two different ways: offline and online. Offline assessment methods are those that cannot provide immediate feedback for users: post-analysis of recorded training [10] sessions or questionnaires answered by users [11]. However, users can forget their actions after some time, which gives to this type of assessment a lower didactic impact. Online assessment methods can provide real-time feedback and trainees can immediately repeat the training and try to correct their actions to improve their performance. It is important since it can provide a more effective learning process. Several methods for online assessment have been proposed for medical simulators based on VR [4,5,6,9,14].

The calibration of the evaluation system is necessary to acquire and label correct and incorrect ways to perform the procedure. It is provided by an expert of the simulator subject that executes several times the procedure in different ways in order to generate parameters for each execution. A previously defined number for classes of performance is used to label each execution. All interaction and environment parameters are acquired during this process to be used by the assessment method, which is normally based on a pattern recognition technique. As example, a M=3 number of classes can refers to 1: “good performance“, 2 - “regular performance“, 3 - “bad performance“ [14].

# 3 Assessment Methods and Decision Rule

## *3.1* Fuzzy Naive Bayes (FNB) Method

A fuzzy set A in Ω is defined for each element ω ϵ Ω as a mapping $\mu_A$, called membership function, which can associate each element from Ω to [0,1], and is interpreted as the degree of membership of ω in A [7]. Some fuzzy versions for the Naive Bayes classifier were proposed. In this work we follow the version proposed by Störr [17] and used by [5] as a kernel of an assessment system for training based on VR.

Formally, let be $\Omega=\{1,...,M\}$ the classes of performance in space of decision, where $M$ is the total number of classes of performance. Let be $w_i$, $i \in \Omega$ the class of performance for a trainee. It is possible to determine the class of performance most probable for this trainee given a data vector $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ and assuming that each $X_k$, $k=1,...,n$ is a fuzzy variable, with normalized membership functions $\mu_i(X_k)$, where $i=1,...,M$. The method is defined by [5]:

$$P(w_i \backslash \boldsymbol{X}) = (1/S)\, P(w_i) * \mathrm{P(W_i)} \quad \prod\nolimits^{n}_{k=1} \quad [P(X_k \backslash w_i)\, \mu_i(X)],\ \mathrm{i} \in \Omega \tag{1}$$

where $S$ is a scale factor which depends on $X_1, X_2, ..., X_n$.

The classification rule for Fuzzy Naive Bayes is: select performance class $w_i$ for the vector $\boldsymbol{X}$ if:

$$P(w_i \backslash X) > P(w_j \backslash X) \text{ for all } i \neq j \text{ and } i, j \in \Omega \tag{2}$$

## *3.2* Possibility and Necessity Measures (PMN) Method

Let A be a fuzzy subset of Ω, with its membership function $\mu_{A,}$ and let X be a variable which assumes values ω in Ω. Then, the possibility distribution π is a function associated to X and is defined as:

$$\pi_X(\omega) = \mu_A(\omega) \tag{3}$$

The possibility measure Π and the necessity measure are defined respectively by:

$$\Pi\,(A) = \sup\,\{\pi\,(u) \mid x \in A\} \quad \text{and} \quad N\,(A) = \inf\,\{1 - \pi\,(u) \mid x \notin A\}.$$

Other relations between them were provided in [16]:

$$\Pi\,(\varnothing) = N\,(\varnothing) = 0; \quad \Pi\,(A) = N\,(A) = 1;$$

$$\max(\Pi(A)\,, \Pi(\bar{A})) = 1; \min(N(A), N(\bar{A})) = 0;$$

The possibility and the necessity measures are dual:

$$\Pi(A) = 1 - N(\bar{A}) \quad \text{and} \quad N(A) = 1 - \Pi(\bar{A}).$$

Some relations between them can be provided [16]:

$$\Pi(A) \geq N(A); \quad N(A) > 0 \Rightarrow \Pi(A) = 1; \quad \Pi(A) < 1 \Rightarrow N(A) = 0$$

Let A and B be fuzzy subsets of $\Omega$, with membership functions $\mu_A$ and $\mu_B$, respectively. Let X be a variable which assumes values $\omega \in \Omega$. The conditional possibility and necessity measures are given by [16]:

$$\Pi (A \mid B) = \max_{u \in X} \min (\mu_A (u), \pi_B (u)) \text{ and}$$

$$N (A \mid B) = \min_{u \in X} \max (\mu_A (u), 1- \pi_B (u)). \tag{4}$$

From equation (4) is possible to construct an interval for the real value of the class of performance $w_i$, given each feature $X_k$ , with $k=\{1,...,n\}$, from the training data $\boldsymbol{X} = \{X_1, X_2, ..., X_n\}$ from a user [14]:

$$\mu_{\omega i} (X_j) \in [N(\omega_i \mid X_k); \Pi(\omega_i \mid X_k)].$$

The domain of membership function for the class of performance $\omega_i$ is an interval where the minimum value is the minimum compatibility and the maximum value is the maximum compatibility:

$$\mu\omega_i (\mathbf{X}) \in [\text{compat}_{\min}; \text{compat}_{\max}].$$

As the class of performance $w_i$ is expressed by a conjunction of features $X_j$ , then this aggregation is performed by a t-norm. In this case, the “min” operator preserves the semantics of possibility and necessity measures [14]:

$$\text{compat}_{\min} = \min_k (N (\omega_i \mid X_k)) \quad \text{and} \quad \text{compat}_{\max} = \min_k (\Pi (\omega_i \mid X_k)).$$

The defuzzification process can be done using, for instance, the centroid method, where C $[\mu_{wi} (\boldsymbol{X})]$ is the centroid between $\text{compat}_{\min}$ and $\text{compat}_{\max}$ for the pertinence function of class $w_i$, according to $\boldsymbol{X}$. Then, the decision rule is: select performance class $w_i$ for the vector $\boldsymbol{X}$ if [14]:

$$w_i = \arg\max_{1 \leq i \leq M} C [\mu_{wi} (\boldsymbol{X})].$$

## 4 Embedded System

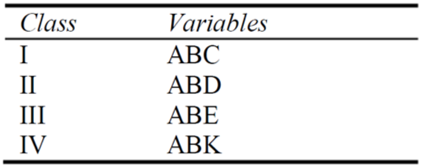
An embedded system is a combination of hardware and software with additional components (mechanical and/or electronic) performing a dedicated function [8]. Its hardware is specifically designed to fulfill requirements of a system making it cheaper. This kind of device is also characterized by having higher quality, higher reliability and lower cost components [8] than other computer systems. Its architecture is generally similar to that of a computer system and may be composed by main memory, secondary memory, processor and buses input and output, such as: USB port, VGA port, network adapter and others, according to the task to be performed.

As mentioned in Section 2, a VR system can demand processing of massive data, which can overload the main system. To relief it, we take the advantage of using an embedded system to execute assessment tasks to meet timing requirements [3].

## 5 Methodology

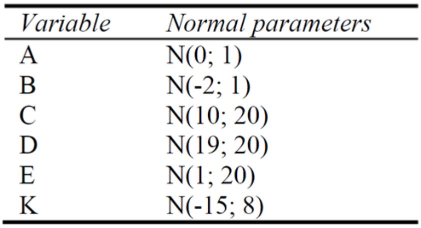
The objective of this paper is analyze the efficiency of the architecture proposed by [2] in producing good results, investigating the limits of this architecture in providing online assessments and what are performance relationships between the assessments provide by a PC-only system and the architecture proposed by [2]. Initially, the authors developed computational programs for both assessment methods, a PC-only system and an embedded system, which configurations were, respectively: Athlon 64 X2 AMD processor, 2,4GHz, 2GB of DDR2 RAM, under Fedora Linux and Geode AMD processor, 500MHz, 256MB of DDR RAM, under Gentoo Linux. After, were used Monte Carlo simulations [15] with 2000 databases, where the smallest database contained three dimensional vectors with 1000 positions each, with four classes of performance, as showing in the Table 1. Databases with increasing sizes were randomly generated, starting from the smaller one, and the performance profiles were outlined for both methods. Performance classes were defined from the combination of random variables and those from Gaussian distributions generated with the predefined parameters, with at least 30% of intersection. Variables and parameters used are present in Table 2.

**Table 1.** Classes setup



| *Class* | *Variables* |
|---|---|
| I | ABC |
| II | ABD |
| III | ABE |
| IV | ABK |

**Table 2.** Variables and parameters



| *Variable* | *Normal parameters* |
|---|---|
| A | N(0; 1) |
| B | N(-2; 1) |
| C | N(10; 20) |
| D | N(19; 20) |
| E | N(1; 20) |
| K | N(-15; 8) |

The calibration phase used 10 databases with 1000 positions each one. To check the performance of the methods were used in the calibration phase between 100 and 2000 databases. Each database had the same number of positions, but they were used several times to increase the final database. So, the final database had 20000 positions to assess. In the case of the trainee assessment performed by PC-only, the method was executed locally. However, for the embedded system was necessary defining a framework for communication via Ethernet network (client-server via TCP/IP). The server application ran on the PC. It simulated the generation of data from the VR system, received and displayed the assessment report. The client application ran on embedded system and received data from the server, conducted the trainee´s assessment, calculated its own average time of execution and sent the final report to the server. The average accuracy of each method was verified using the Kappa coefficient proposed by Cohen [12] and recommended by the literature of pattern recognition [13].

## 6 Results and Discussions

From the Monte Carlo simulation data, as described above, were obtained confusion matrices and the average time of execution for both methods, which are presented in Tables 3 for FNB method and in Table 4 for PNM method. The main diagonal of those matrices represents the correct classifications for each method. As was expected, the results obtained from the assessment system-only running on the PC and embedded system were numerically equal.

Using the Table 3 was computed the Kappa coefficient for FNB, which resulted in 0.653 with standard deviation 0.006. The percentage of correct classification was 73.95%. It were noticed 2084 misclassifications and those errors were concentrated in the classes I and IV. The percentage of correct classification for the class I was poor: only 1.55%. However, FNB achieved good results for the classes II and III (100.00%). The percentage was 94.25% for the class IV. The average time of execution for the FNB method was 0.2098 seconds for the embedded system and 0.0139 seconds for the PC-only.

For the PNM method, the time measured was 0.2104 and 0.0140 seconds for the embedded system and for the PC-only, respectively. From Table 4 data, the Kappa coefficient obtained was 0.693 with standard deviation 0.006. The percentage of correct classification was 77.00%. The number of misclassifications was 1840, distributed in the classes I, II and III. The percentages for correct classifications were 94.10%, 20,10% and 96,80% for the classes I, II and III, respectively. IN this case, PNM achieved good results for the class IV (100.00%). These results were not so good as those presented by [2], specially for the class II.

**Table 3.** Fuzzy Naive Bayes Error Matrix

| *Class* | *I* | *II* | *III* | *IV* |
|---|---|---|---|---|
| I | 31 | 633 | 1336 | 0 |
| II | 0 | 2000 | 0 | 0 |
| III | 0 | 0 | 2000 | 0 |
| IV | 0 | 0 | 115 | 1885 |

**Table 4.** Possibility and Necessity Error Matrix

| *Class* | *I* | *II* | *III* | *IV* |
|---|---|---|---|---|
| I | 1822 | 0 | 178 | 0 |
| II | 1589 | 402 | 9 | 0 |
| III | 64 | 0 | 1936 | 0 |
| IV | 0 | 0 | 0 | 2000 |

Both methods satisfied the requirement of processing time lower than one second either running in the embedded system as in the PC-only. However, the ratio between processing times when running in the PC-only and when running in the embedded system, with the configurations described above, was around 15 times. It was expected, and this is clearly due to the differences between their configurations.

It can be seen in the Figure 1 (left) that both curves are linear. The same can be observed for the PC-only (Figure 1 - right) when the databases are lower than 1500. After that, the behavior for PC-only simulations is nonlinear. This change of behavior can be due to the scheduling of the operational system. The time to run 1500 databases in the PC-only is approximately the same to run 100 databases in the embedded system. The graphic on left has a 400 seconds scale with 420s. The graphic on right has a 40 seconds scale and the limit of curves are below this value.

An embedded system was designed for this specific task but not for the PC-only. In this last one, its operational system spends time to perform other tasks not related to the assessment task or to VR-based simulation. It should be noted also that, although the total time for carrying out the evaluation for all databases was higher than one second (see Figure 1), the ratio for the amount of total time positions of each vector remained below one second to any number of databases and for both methods.

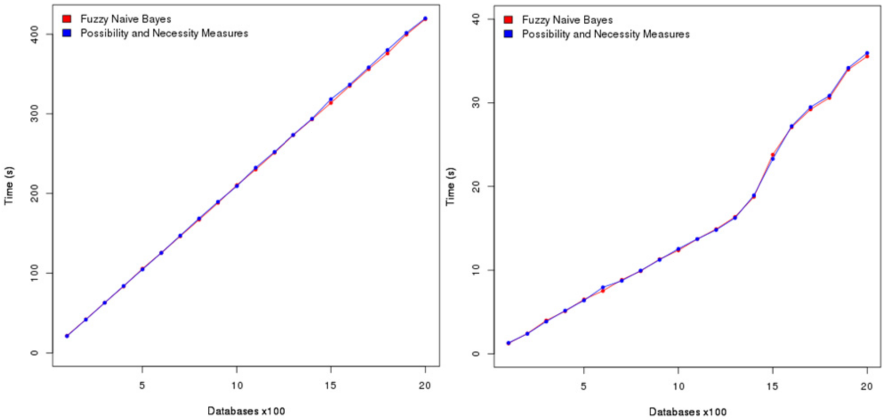


**Fig. 1.** Processing times for embedded system (left) and for the PC-only (right)

## 7 Conclusions

An architecture for assessment based on embedded systems was proposed by [2] some years ago. In the present paper were presented answers for some questions open, as the real efficiency of this architecture, its limitations to provide online assessment and its performance if compared to a CPU based approach. The results obtained shown this assessment architecture can provide the same results as methods implemented for PC-only. When comparing the processing times to perform the assessment task using the embedded system, the PC-only was around 10.5 times faster, but the embedded system also achieved the requirement of assessment time lower than 1 second.

By the results obtained was possible to confirm the proposal [2] of transferring assessment tasks to an embedded system and leave the main system available to process the tasks related to the VR simulation. It was important to observe that even with a less powerful configuration, the embedded system was able to provide fast answers. Therefore, the use of assessment methodologies, even if computationally expensive, can be considered to be implemented in embedded systems without compromising the simulation performance.

**Acknowledgement.** This work is supported by the National Council for Scientific and Technological Development (CNPq): processes 310470/2012-9, 310561/2012-4 , PIBIC//UFPB and 181813/2010-6 (INCT-MACC).

## References

1. Machado, L.S., Mello, A.N., Lopes, R.D., Odone Fo, V., Zuffo, M.K.: A Virtual Reality Simulator for Bone Marrow Harvest for Pediatric Transplant. Studies In Health Technology and Informatics 81, 293–297 (2001)
2. Moraes, R.M., Machado, L.S.: Using Embedded Systems to Improve Performance of Assessment in Virtual Reality Training Environments. In: Int. Conf. Engineering and Technology Education (Intertech 2008), Santos, Brazil, pp. 140–144 (2008)
3. Moraes, R., Machado, L.: Development of a Medical Training System with Integration of Users' Assessment. In: Kim, J.-J. (ed.) Virtual Reality, ch. 15. Intech (2011)
4. Moraes, R.M., Machado, L.S.: Assessment Based on Naive Bayes for Training Based on Virtual Reality. In: Int. Conf. Engineering and Computer Education (ICECE 2007), Santos, Brasil, pp. 269–273 (2007)
5. Moraes, R.M., Machado, L.S.: Another Approach for Fuzzy Naive Bayes Applied on Online Training Assessment in Virtual Reality Simulators. In: Safety, Health and Environmental World Congress (SHEWC 2009), Mongaguá, Brazil, pp. 62–66 (2009)
6. Moraes, R.M., Machado, L.S.: Online Assessment in Medical Simulators Based on Virtual Reality Using Fuzzy Gaussian Naive Bayes. Journal of Multiple-Valued Logic and Soft Computing 18(5), 479–492 (2012)
7. Zadeh, L.A.: Fuzzy Sets. Information and Control 8, 338–353 (1965)
8. Barr, M.: Programming Embedded Systems in C and C++. O'Reilly (1999)
9. Moraes, R.M., Machado, L.S.: Gaussian Naive Bayes for Online Training Assessment in Virtual Reality-Based Simulator. Mathware & Soft Computing 16, 123–132 (2009)
10. McBeth, P.B., et al.: Quantitative Methodology of Evaluating Surgeon Performance in Laparoscopic Surgery". Studies in Health Technology and Informatics 85, 280–286 (2002)
11. Dinsmore, M., Lagrana, N., Burdea, G., Ladeji, J.: Virtual Reality Training Simulation for Palpation of Subsurface Tumors. In: IEEE VRAIS 1997, pp. 54–60. IEEE Press (1997)
12. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37–46 (1960)
13. Webb, A.: Statistical Pattern Recognition, 2nd edn. John Wiley, Chichester (2005)
14. Machado, L.S., Moraes, R.M.: Medical Skills Assessment in Training Based on Virtual Reality Using a Possibilistic Approach. In: 10th Int. FLINS Conf (FLINS 2012), Istanbul, Turkish, pp. 339–345 (2012)
15. Gentle, J.E.: Random Number Generation and Monte Carlo Methods, 2nd edn. Springer Science (2005)
16. Dubois, D., Prade, H.: Possibility theory. Plenum Press, New-York (1988)
17. Störr, H.-P.: A compact fuzzy extension of the naive Bayesian classification algorithm. In: 3rd International Conference on Intelligent Technologies and Vietnam-Japan Symposium on Fuzzy Systems and Applications, Hanoi, Vietnam, pp. 172–177 (2002)

# Improving the Efficiency of MECoMaP: A Protein Residue-Residue Contact Predictor

Alfonso E. Márquez-Chamorro[1], Federico Divina[1], Jesús S. Aguilar-Ruiz[1], and Cosme E. Santiesteban-Toca[2]

[1] School of Engineering, Pablo de Olavide University of Sevilla, Spain
{amarcha,fdivina,aguilar}@upo.es
[2] Centro de Bioplantas, University of Ciego de Avila, Cuba
cosme@bioplantas.cu

**Abstract.** This work proposes an improvement of the multi-objective evolutionary method for the protein residue-residue contact prediction called MECoMaP. This method bases its prediction on physico-chemical properties of amino acids, structural features and evolutionary information of the proteins. The evolutionary algorithm produces a set of decision rules that identifies contacts between amino acids. These decision rules generated by the algorithm represent a set of conditions to predict residue-residue contacts. A new encoding used, a fast evaluation of the examples from the training data set and a treatment of unbalanced classes of data were considered to improve the the efficiency of the algorithm.

**Keywords:** protein structure prediction, residue-residue contact, multi-objective optimization, evolutionary computation.

## 1 Introduction

One of the central goals of bioinformatics is the prediction of protein function and tertiary structure from the linear sequence of amino acids (primary structure). Determining the three dimensional structure of proteins is necessary to understand the functions of molecular protein level. On the other hand, misfolding proteins can be the principal cause of some diseases. Since protein function is determined by its structure, a misfold implies that a protein can not fulfill its function correctly. Alzheimer's disease, cystic fibrosis, bovine spongiform encephalopathy (mad cow disease) and its human variant are now all attributed to protein misfolding. The knowledge of the misfolding factors and understanding the protein folding process, would help in developing cures for these diseases.

The primary structure, or amino acid sequence, of a protein is much easier to determine than its tertiary structure. Moreover, the gap between the number of proteins with known sequence and the number of proteins with known tertiary structure is rapidly increasing. In order to reduce this gap, there have been many researches focused on determining the tertiary structure of a protein

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 166–173, 2013.

from its sequence [1,2]. The high number of protein sequences whose three-dimensional structures must be determined, make computational methods for protein structure prediction (PSP) an essential tool. We believe that EAs well suited for solving the PSP problem, since PSP can be seen as a search problem through the space determined by all the possible protein foldings. Moreover, PSP problem can be considered as a optimization problem with several objectives [3]. The task of finding one or more suboptimal solutions is called Multi-objective optimization. Our algorithm is based on these approaches.

An useful, and commonly used, representation for protein 3D structure is the protein contact map, which represents binary proximities (contact or non-contact) between each pair of amino acids of a protein. Our approach is included in this category.

The aim of this work consists of improving our proposal MECoMaP (Multi-objective Evolutionary Contact Map Predictor) [4] in order to increase the efficiency of the protein contact map prediction. The prediction is based on three physico-chemical properties: hydrophobicity (H), polarity (P) and charge (C), structural features: solvent accessibility (SA) and secondary structure (SS) and evolutionary information in form of Position Specific Scoring Matrix (PSSM). It is known that amino acid properties play an important role in the PSP problem [5]. Several PSP methods rely on amino acids properties, *e.g.*, HP models. On the other hand, a vast majority of PSP algorithms used SS, SA and PSSM as predictive features.

The remainder of this paper is organized as follows. Our multi-objective evolutionary approach is described in section 2. Section 3 presents the experimentation and obtained results. Finally, section 4, includes some conclusions and possible future works.

## 2 Methodology

MECoMaP is based on the Strength Pareto Evolutionary Algorithm (SPEA). Each individual of the population represents a decision rule. In particular, rules are based on the previously mentioned amino acid properties. Basically rules specify a set of conditions on each property, that, if satisfied, predict a contact between two amino acids.

In the following the preparation of data, attribute selection, the encoding, the fitness function and the genetic operators used by the EA will be presented.

### 2.1 Preparation of Data

We selected from PDB a protein data set (DS1) that consists of 173 non-redundant proteins with sequence identity less than 25%, and was obtained from [6]. The minimum and maximum lengths of proteins are 31 and 753 amino acids, respectively. DS1 contains 240501 positive examples (contacts) and 5034050 negative examples (non-contacts).

The second data set (DS2), with 53 non-redundant and non-homologous globulin proteins, is detailed in [7]. The sequence identity of DS2 dataset is also lower than 25%. DS2 is formed by a total of 30546 contacts and 356528 non-contacts.

As we can see, the positive and negative classes (contact and non-contacts) are notably unbalanced. We have performed a resampling of data using 1:1 and 2:1 contact/non-contacts ratios. Using 1:1 ratio we obtain a higher rate of predicted contacts, however the rate of false positives of the predictor is increased. Specifically, the accuracy results for both ratios on DS1 and DS2 are shown in Table 1. As seen in the table, the 2:1 ratio presented better performance. This is also the case for DS2 data set. The optimization of this parameter also implies a lower computational cost for the algorithm. Based on the results of the table, we decided to perform a re-sampling using the 2:1 ratio.

**Table 1.** Average accuracy results obtained for different contact/non-contacts ratios for the DS1 and DS2 protein data set

| Ratio | Data Set | $Accuracy_\mu$ |
|---|---|---|
| 1:1 | DS1 | $0.21_{\pm 0.10}$ |
| 2:1 | DS1 | $0.23_{\pm 0.08}$ |
| 1:1 | DS2 | $0.16_{\pm 0.13}$ |
| 2:1 | DS2 | $0.20_{\pm 0.11}$ |

### 2.2 Feature Selection

As stated before, the prediction is based on a set of amino acid properties which are very important in the folding process. The reason for basing the prediction on such properties, is that it has been shown that amino acids that are in contact, are characterized by similar properties [8]. We selected Kyte-Doolittle hydropathy profile [9], the Grantham profile [10] for polarity and the Klein scale for net charge [11]. Hydrophobic amino acids are generally found in the inner of proteins protected from direct contact with water. Inversely, the hydrophilic amino acids are generally found on the outside of proteins as well as in the active centers of enzymatically active proteins. The net charge takes into account the charged groups present in any amino acid, peptide or protein nd the pH of its environment. In addition to these properties, we also use two structural features of proteins (SS and SA) and evolutionary information, in form of PSSM.

Secondary structure prediction consists of predicting the location of $\alpha$-helices, $\beta$-sheets and turns from a sequence of amino acids. The location of these motifs could be used by approximation algorithms to obtain the tertiary structure of the protein. We obtain SS predictions using PSIPRED. SA refers to the degree to which a residue interacts with the solvent molecules. The prediction of SA value is performed using ICOS Server for the prediction of structural aspects of protein residues *http://cruncher.cs.nott.ac.uk/psp/prediction.*

A PSSM determines the substitution scores between amino acids according to their positions in the alignment. Each cell of the matrix represents the observed substitution frequency at a given position divided by the expected substitution frequency at that position. PSSM is obtained using PSI-BLAST.

H, P and PSSM values were normalized between -1 and 1. C values are represented with -1, 0 and 1 for negative, neutral and positive charges. SS values are identified with 1, 2 and 0 for alpha-helices, beta-sheets and random coils, respectively. SA values are ranging from 0 to 4 according to the exposure level.

The procedure scheme of preproccessing of the data is represented in Figure 1. We have obtained five different files with the information of the properties. They constitute the training data of the algorithm.

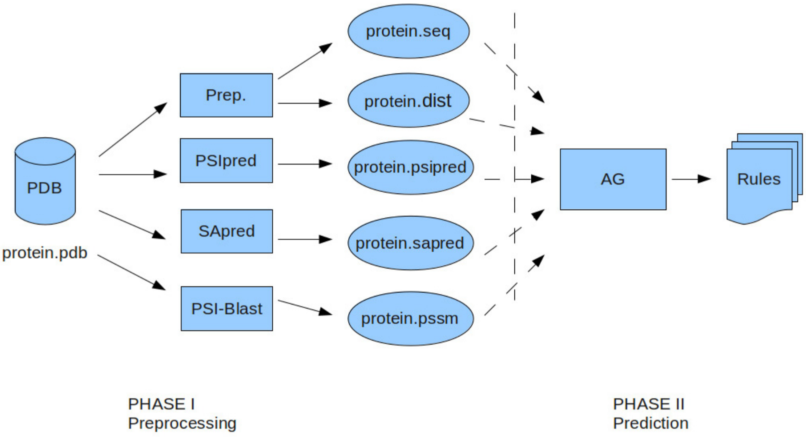


**Fig. 1.** Preprocessing procedure scheme

### 2.3 Encoding

An individual is constituted by six blocks which represent the different properties of amino acids. Each block indicates the values of a respective property in all the positions of the residues in the window. We use two windows of ±3 residues centered around the two target amino acids $i$ and $j$. Therefore, one window is relative to amino acids $i-3, i-2, i-1, i, i+1, i+2, i+3$ and the other one is relative to amino acids $j-3, j-2, j-1, j, j+1, j+2, j+3$.

We define each individual as a decision rule $R_{i,j}$ for amino acids $i$ and $j$:

$$R_{ij} = \{\{H_{min}, H_{max}\}^{1..n}, \{P_{min}, P_{max}\}^{1..n},$$
$$C^{1..n}, SS^{1..n}, SA^{1..n}, \{PSSM_{min_{ij}}, PSSM_{max_{ij}}\}^{1..20}\} \qquad (1)$$

where $n$ indicates the total number of amino acids (in this case $n = 14$). Each element of $R_{ij}$ must fulfill the following requirements:

$$\begin{gathered}
-1 \leq H_{min} < H_{max} \leq 1 \\
-1 \leq P_{min} < P_{max} \leq 1 \\
C \in \{-1, 0, 1\} \\
SS \in \{-1, 0, 1, 2\} \\
SA \in \{-1, 0, 1, 2, 3, 4\} \\
-1 \leq PSSM_{min}^{1..20} < PSSM_{max}^{1..20} \leq 1
\end{gathered} \tag{2}$$

This decision rule determines whether two amino acids $i$ and $j$ are in contact, where $1 \leq i < j \leq L$, being $L$ the sequence length. Our representation consists in $14 \times 2$ attributes for H, $14 \times 2$ for P, 14 for C, 14 for SS, 14 for SS and $2 \times 2 \times 20$ for PSSM, 178 attributes in total.

### 2.4 Fitness Function

As stated in [4], we consider two objectives to be optimized: coverage and accuracy. Coverage represents the number of predicted contacts and accuracy evaluates the real predicted contacts rate. Therefore, $Coverage = C/C_t$ and $Accuracy = C/C_p$, where $C$ is the number of correctly predicted contacts of a protein, $C_t$ is the total number of contacts of the protein and $C_p$ is the number of predicted contacts. We aim at finding the best compromise between these two measures. The fitness of an individual $x$ is given by the number of individuals that $x$ dominates.

### 2.5 Genetic Operators

A 2-point crossover operation was employed with a binary tournament selection and a 0.5 probability. In each tournament, we select the individual which is located in the better Pareto front.

A first mutation operator follows a Gaussian distribution for a randomly selected individual. This operator increases or decreases a gene value with a probability of 0.5 randomly interval. A second mutation operator randomly selects a gene that is related to a given property, with a 0.1 probability, and moves the bounds to the maximum or minimum of the domain, making the property irrelevant in this rule. For example, if the property is the polarity, we change the range to -1, 1 so the rule does not take into account this property in this case. After the mutation, we test if the obtained values are in the adequate ranges for the corresponding property.

The population size is set to 100, and the initial population is randomly initialized with a 0.6 probability. The maximum number of generations that can be performed is set to 100. However, if the fitness of the best individual does not increase over twenty generations, the algorithm is stopped and a solution is provided. At the end of the execution, repeated or redundant rules are discarded from the solution set.

### 2.6 Efficient Evaluation Structure

In order to reduce the computational time of our method, we have implemented an AVL tree [12] to order and classify the training examples according to their property values. This tree organizes the information in such a way that it is not necessary to process all the examples to evaluate individuals (candidate decision rules) from the genetic population generated by MECoMaP. The time of the operations on an AVL tree is O(log n) average, where n is the number of elements. Each node determines a condition of a property and each leaf represents a list with the training examples that fulfills all the conditions impose in the predecesor nodes. Each level of the tree represents a determined property of a determined position of an amino acid. We consider a tree example in figure 2. Level 1 represents the hydrophobicity of amino acid $i$ and level 2 indicates the polarity of amino acid $j$. As example, leaf node $N1$ stores all the training examples whose amino acid in position $i$ has a hydrophobicity value lower than 0 and a polarity value in position $j$ is also lower than 0. We achieve a reduction of the computational cost about 50% by means of a fast evaluation of examples from the dataset.

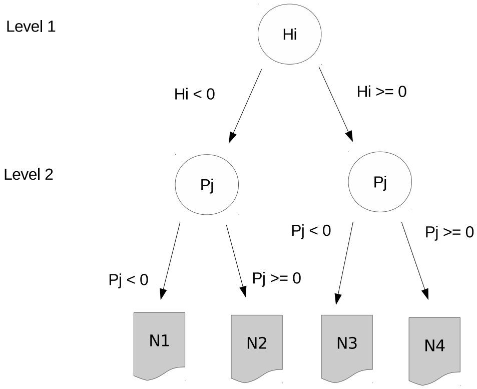


**Fig. 2.** Example of efficient evaluation structure (AVL tree)

## 3 Experiments and Results

We have built a file in arff format with all the training data information. This file is constituted by all the protein subsequences of two windows of seven amino acids encoded with the values of the cited attributes. The positive class (contact) is represented with 1 and the negative class (non-contact) is represented with 0. The ratio between the positive and negative classes was set to 2:1 for DS1 and DS2 data sets. The training data used contained all the possible subsequences with a minimum separation between contact residues of 7 amino acids for DS1 and a separation 6 amino acids for DS2. We have performed several experiments with three Weka classifiers [13]: Näive Bayes (NB), C4.5 classifier tree (J48) and Nearest Neighbor approach with $k = 1$ (IB1). The obtained results can be seen in Table 2 for a 3-fold cross-validation. We appreciate low coverage and accuracy values in all the cases. This experiment was performed with the aim of validating our representation and confirms that the new encoding provides enough information for a good performance of a learning classifier. Moreover, we can also notice that MECoMaP achieved the best results for this experiment and improve the results for DS1 and DS2 data set shown in [4].

**Table 2.** Average results obtained for MECoMaP and different classification Weka algorithms for the DS1 and DS2 protein data set

| Algorithm | Data Set | $Coverage_{\mu\pm\sigma}$ | $Accuracy_{\mu\pm\sigma}$ |
|---|---|---|---|
| J48 | DS1 | $0.04_{\pm 0.07}$ | $0.19_{\pm 0.08}$ |
| IB1 | DS1 | $0.08_{\pm 0.05}$ | $0.07_{\pm 0.05}$ |
| NB | DS1 | $0.15_{\pm 0.03}$ | $0.08_{\pm 0.02}$ |
| MECoMaP | DS1 | $0.18_{\pm 0.13}$ | $0.26_{\pm 0.32}$ |
| MECoMaP 2.0 | DS1 | $0.20_{\pm 0.15}$ | $0.29_{\pm 0.11}$ |
| J48 | DS2 | $0.10_{\pm 0.02}$ | $0.10_{\pm 0.05}$ |
| IB1 | DS2 | $0.07_{\pm 0.10}$ | $0.07_{\pm 0.05}$ |
| NB | DS2 | $0.10_{\pm 0.10}$ | $0.18_{\pm 0.10}$ |
| MECoMaP | DS2 | $0.12_{\pm 0.01}$ | $0.38_{\pm 0.09}$ |
| MECoMaP 2.0 | DS2 | $0.18_{\pm 0.08}$ | $0.39_{\pm 0.07}$ |

## 4 Conclusions and Future Work

In this work, we presented some improvements to a multi-objective optimization algorithm for the residue-residue contact prediction. Two of these improvements enhance the efficiency of the algorithm: the introduction of new features based on evolutionary information (PSSM) for the encoding and a treatment for the unbalanced classes. An efficient evaluation structure for a fast evaluation of the training data is also included to reduce the time complexity of the EA. This algorithm generates rules that predict the necessary conditions for the contact between two amino acids based on their physico-chemical properties. The algorithm was tested on two sets of proteins that had been previously used in the literature and achieved better coverage and accuracy rates than the predecessor version of the algorithm. As future work, the incorporation of new evolutionary information such as correlated mutations must be taken into account. Furthermore, our algorithm must be validated with a higher number of proteins data set.

## References

1. Tegge, A., Wang, Z., Eickholt, J., Cheng, J.: Nncon: Improved protein contact map prediction using 2d-recursive neural networks. Nucleic Acids Research 37(2), 515–518 (2009)

2. Jones, D.T., Buchan, D.W., Cozzetto, D., Pontil, M.: Psicov: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2), 184–190 (2012)
3. Calvo, J.C., Ortega, J.: Parallel protein structure prediction by multiobjective optimization. Parallel, Distributed and Network-based Processing 12(4), 407–413 (2009)
4. Marquez-Chamorro, A.E., Asencio, G., Divina, F., Aguilar-Ruiz, J.S.: Evolutionary decision rules for predicting protein contact maps. Pattern Analysis and Applications, PAAA (September 1-13, 2012)
5. Russell, R.B., Betts, M.J., Barnes, M.R.: Amino acid properties and consequences of subsitutions. In: Bioinformatics for Geneticists. Wiley (2003)
6. Fariselli, P., Olmea, O., Valencia, A., Casadio, R.: Prediction of contact map with neural networks and correlated mutations. Protein Engineering 14, 133–154 (2001)
7. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. Bioinformatics 8, 113 (2007)
8. Gupta, N., Mangal, N., Biswas, S.: Evolution and similarity evaluation of protein structures in contact map space. Proteins: Structure, Function, and Bioinformatics 59, 196–204 (2005)
9. Kyte, J., Doolittle, R.F.: A simple method for displaying the hydropathic character of a protein. J. J. Mol. Bio. 157, 105–132 (1982)
10. Grantham, R.: Amino acid difference formula to help explain protein evolution. J. J. Mol. Bio. 185, 862–864 (1974)
11. Klein, P., Kanehisa, M., DeLisi, C.: Prediction of protein function from sequence properties: Discriminant analysis of a data base. Bioch. Bioph. 787, 221–226 (1984)
12. Adelson-Velskii, G., Landis, E.M.: An algorithm for the organization of information. Proceedings of the USSR Academy of Sciences; Soviet Math. 3, 1259–1263
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: An update. SIGKDD Explorations 11 (2009)

# Identifying Loose Connective and Muscle Tissues on Histology Images

Claudia Mazo[1], Maria Trujillo[1], and Liliana Salazar[2]

[1] School of Computer and Systems Engineering
[2] Department of Morphology
Universidad del Valle

**Abstract.** Histology images are used to identify biological structures present in living organisms — cells, tissues, and organs — correctly. The structure of tissues varies according to the type and purpose of the tissue. Automatic identification of tissues is an open problem in image processing. In this paper, the identification of loose connective and muscle tissues based on morphological tissue information is presented.

Image identification is commonly evaluated in isolation. This is done either by eye or via some other quality measure. Expert criteria — by eye — are used to evaluate the identification results. Experimental results show that the proposed approach yields results close to the real results, according to expert opinion.

## 1 Introduction

The development of digital technologies has made available, to physicians and biologists, digital cameras connected to microscopes for image capture in order to preserve observed samples. Thus, large repositories of images are gathered of histological samples and allow automatic identification of tissues.

Connective and muscle are two of the four basic body tissues. The connective tissue is divided into: loose and dense. The dense connective is classified into: regular and irregular. The muscle tissue is divided into: striated or skeletal, smooth, striatal heart [6]. In this paper, we focus on identify loose connective and muscle tissues.

The identification of loose connective and muscle tissues is an open problem because of the close relation between them and the difficulty to demarcate the boundaries of each one. This process starts with a segmentation of each of the tissues and their refinement to eliminate additional information.

Automatic segmentation of the loose connective and muscle tissues involves several problems: the hard boundary between loose connective and muscle tissues in areas where they interrelate, the presence of red blood cells in some samples, the similarity between smooth muscle and dense regular connective tissue, among others.

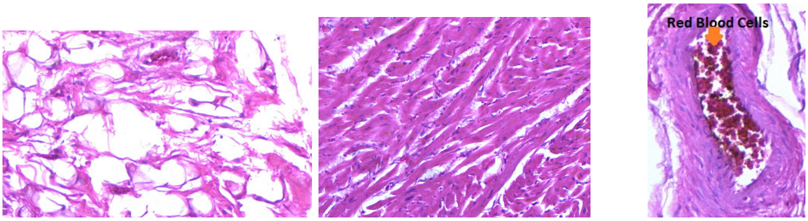
In this paper, an automatic segmentation of loose connective and muscle tissue approach based on morphological information is presented. Obtained results of the largest eigenvalue of structure tensor along with the red and the green

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 174–180, 2013.

color channels are combined in the K-means clustering for identifying loose connective tissue and obtaining a first approximation of the identification of muscle tissue. Finally morphological operations, thresholding, subtraction of images and thresholds are used to select areas of muscle tissue and refine results of muscle identification. Experimental results show that the proposed approach identifies correctly loose connective and muscle tissue in histological images. The rest of the paper is organised as follows. The proposed segmentation approach is presented in Section II. Experimental validation is included in Section III. Finally, conclusions are in Section IV.

## 2 Proposed Approach

The loose connective tissue is characterised by abundant fluid and tissue basic substance, dispersed structure and more numerous cells than the fibers. The smooth muscle is composed of spindle-shaped cells with a central nuclei which lack transverse striations while exhibiting weak longitudinal grooves. The heart muscle central nuclei and ramifications and interconnections between the fibers. The overall muscle tissue is more dense and compact. The example images of the tissues are illustrated in Fig. 1.



*Connective Tissue* *Heart Muscle Tissue* *Smooth Muscle Tissue*

**Fig. 1.** Example of loose connective and muscle tissue

Obtained borders with the largest eigenvalue of structure tensor algorithm along with the red and the green color channels are used as input into the K-means algorithm in order to obtain loose connective tissues and muscle tissues segmentation. Morphological operations, thresholding and region thresholds are used to refine areas of muscle tissue obtained with the K-means.

### 2.1 Identification of Loose Connective Tissue

Initially, the Structure tensor using the maximum eigenvalues [1], [2] is calculated. This algorithm allows us to obtain the nuclei of the cells [7]. Finally, the identification of loose connective is performed using the K-means algorithm.

**Largest Eigenvalue of Structure Tensor.** The largest eigenvalue of structure tensor [5] is described as follows. Given the red color channel of an image, the red channel is selected since it has the greatest information contains cell nuclei, the structure tensor $J_0$ is defined as the outer product of the gradient vector $\bigtriangledown I$:

$$J_0 = \bigtriangledown I \bigtriangledown I^T = \begin{pmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{pmatrix}, \qquad (1)$$

where $I^T$ symbolised the transpose of I. $J_0$ is extended to the linear structure tensor by a convolution of the components of $J_0$ with a Gaussian kernel $K_p$ (Gaussian smoothing)in order to consider neighbouring information:

$$J_\rho = J_0 * K_\rho = \begin{pmatrix} j_{11} & j_{12} \\ j_{12} & j_{22} \end{pmatrix}. \qquad (2)$$

The matrix $J_p$ has orthonormal eigenvectors $v_1$ and $v_2$ with $v_1$ parallel to

$$\left( \frac{2j_{11}}{j_{11} + j_{22} - \sqrt{j_{11} - j_{22}{}^2 + 4j_{12}^2}} \right). \qquad (3)$$

The eigenvalues are given by

$$\mu_1 = \frac{1}{2}\left[ j_{11} + j_{22} + \sqrt{j_{11} - j_{22}{}^2 + 4j_{12}^2} \right], \qquad (4)$$

and

$$\mu_2 = \frac{1}{2}\left[ j_{11} + j_{22} - \sqrt{j_{11} - j_{22}{}^2 + 4j_{12}^2} \right]. \qquad (5)$$

The eigenvalues describe the average contrast in the eigen-directions within a neighbourhood of size ($\rho$). The vector $v_1$ indicates the orientation with the highest red value fluctuations, while $v_2$ gives the preferred local orientation, the coherence direction. Furthermore, $\mu_1$ and $\mu_2$ serve as descriptors of local structure. Isotropic areas are characterised by $\mu_1 \cong \mu_2$, straight edges gives $\mu_1 \gg \mu_1 = 0$, corner by $\mu_1 \geq \mu_2 \gg 0$ [5].

The structure tensor of an image is a method of analysing the edge structure in an image. Eigenvectors point in the direction orthogonal and across the local edge, with the Eigenvalues indicating the strength of the directional intensity change. The larger eigenvalue shows the strength of the local image edges, the corresponding eigenvector points across the edge (in gradient direction). In this way we get a grayscale image as a result.

**The K-means Algorithm.** To perform the segmentation of images identifying: loose connective tissue, muscle tissue and light regions, the K-means algorithm is used [3]. A psudocode of the K-means algorithm is sketched as follows:

```
Begin
  Determine k centroids randomly
  Calculate the distance between each data and the
  centroids
  Assign each data to the group represented by the
  nearest centroid
  Recalculate centroids
  While centroids do not change
    Calculate the distance between each data and
    the centroids
    Assign each data to the group represented by
    the nearest centroid
    Recalculate centroids
  End While
End
```

### 2.2 Identification of Muscle Tissue

A first segmentation of muscle was performed in the previous step for the K-means. However, it is necessary to refine the results since it included the loose connective tissue in some areas and the red blood cells. To improve the segmentation result will apply three steps: first, erosion [8] is performed on the muscle segmented image. Second, the result of the erosion is subjected to a thresholding, regions under threshold1 are removed. The size of the regions is controlled by the Flood-fill algorithm [4]. Finally the red blood cells are removed with a process of thresholding and morphological operation.

**Erosion.** A problem of the segmented images is the presence of irrelevant details — from the point of view of size. To eliminate these small islands and bumps the segmented image is erode [8]. The erosion and morphological operation is defined as:

$$I \ominus C = \{x \in E | C_x \subseteq I\}, \tag{6}$$

$$C_x = \{c + x | c \in C\}, \forall_x \in E, \tag{7}$$

let $E$ be a Euclidean space $\mathbb{R}^d$ or an integer grid $\mathbb{Z}^d$. $I$ is a binary image in E. $C_z$ is the translation of $B$ by the vector $z$. For erosion to be satisfied that the set of all points $x$, such that $C$ forward $x$, are contained in $I$.

**Threshold.** After performing the erosion an image without irrelevant size areas is obteined, however there are irrelevant areas that are not eliminated since have a larger size, but are reduced. To eliminate these areas a thresholding process is performed, then regions under a threshold are removed. The size of the regions is controlled by the Flood-fill algorithm [4]. A psudocode of the Flood-fill algorithm is sketched as follows:

```
Flood-fill (node, target-color, replacement-color)
  Set Q to the empty queue
  Add node to the end of Q
  While Q is not empty
    Set n equal to the last element of Q
    Remove last element from Q
    If the color of n is equal to target-color
      Set the color of n to replacement-color
      Add the neighbors of current positon (east,
      west, north, south) to then end of Q
    End If
  End While
End
```

**Removing Red Blood Cells.** To remove red blood cells of muscle tissue a segmentation process is performed similar to the removal of connective tissue. Initially, a thresholding on the red channel is performed, obtaining a segmentation of the regions belonging to the red blood cells. About thresholding the result is applied to the morphological operation, erosion, in order not to be part of muscle tissue. After, regions under a threshold2 are removed to avoid segmented muscle or connective cells nuclei. Finally, the muscle tissue image is segmented to subtract the red blood cells from the image, the result of subtracting the second from the first is:

$$I - (I \cap B), \tag{8}$$

let $I$ be the muscle image segmentation. Let B be the red blood cells image segmentation. The regions under a threshold3 are removed to eliminate irrelevant regions resulting from the subtraction.

## 3 Experiments and Analysis of Results

In order to assess the proposed approach, loose connective and muscle tissues samples were processed with hematoxylin-eosin staining. The K-means algorithm for this particular case will use a K=3, each of these clusters will recognize: loose connective tissue, muscle tissue and light regions. The input for the K-means are: the red and the green color channels along with the tensor values obtained. Identification is evaluated by eye, evalution of experts. Automatic segmentation results can be observed in selected images in Fig. 2.

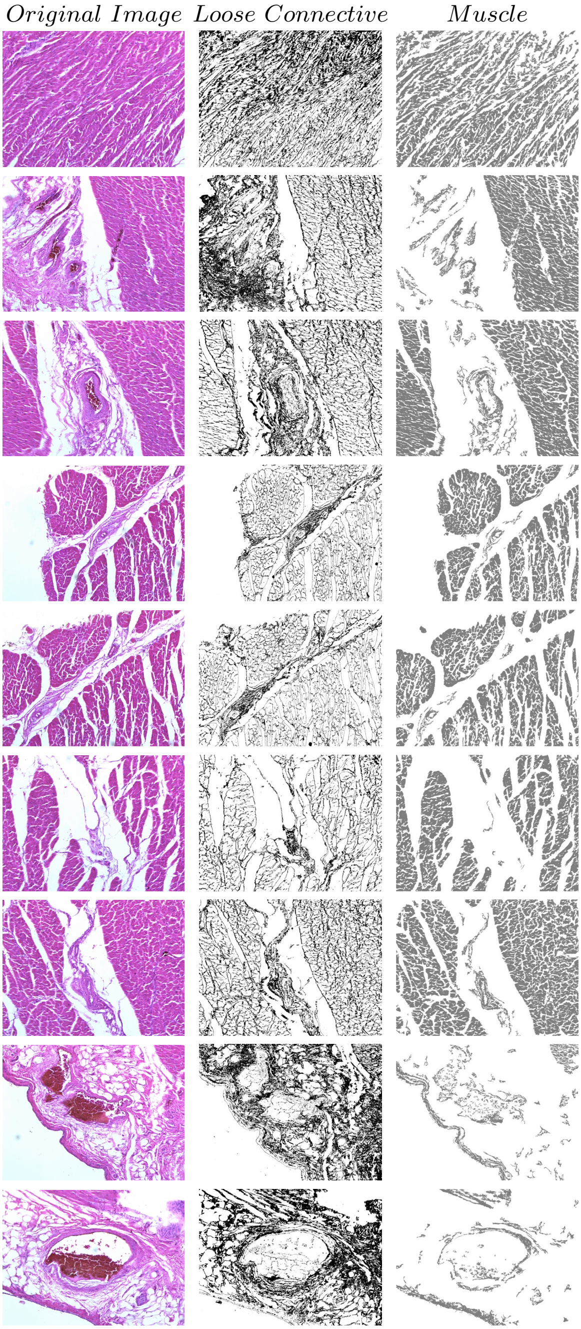


**Fig. 2.** Results obtained using automatic technique for identify loose connective and muscle tissues

## 4 Conclusions

During the research process different approaches were evaluated, such as: thresholding, edge detection algorithms, combining information as input to the k-means algorithm and segmentation algorithms for identifying loose connective and muscle tissue. However, the presented approach provides the closest identification to the real identification — by eye — according to expert opinion.

The difficulty for identifing the loose connective tissue when is immersed in muscle tissue — even visually — is solved with the proposed method.

The experimental evaluation shows that the obtained segmentation is very close to the real one. Additional constrains are required in order to reduce false positive.

## References

1. Rao, A.R., Schunck, B.G.: Computing oriented texture fields. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1989, pp. 61–68 (1989)
2. Weickert, J.: A Scheme for Coherence-Enhancing Diffusion Filtering with Optimized Rotation Invariance. Journal of Visual Communication and Image Representation 13, 103–118 (2002)
3. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 881–892 (2002)
4. Nosal, E.-M.: Flood-fill algorithms used for passive acoustic detection and tracking. In: New Trends for Environmental Monitoring Using Passive Systems, pp. 1–5 (2008)
5. Lu, B., Miao, C., Wang, H.: Pixel level image fusion based on linear structure tensor. In: 2010 IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT), pp. 303–306 (2010)
6. Vegue, J.B.: Atlas de Histología y Organografía Microscópica. Editorial Medica Panamericana, S.A. Madrid, España (2011)
7. Mazo, C., Trujillo, M., Salazar, L.: An Automatic Segmentation Approach of Epithelial Cells Nuclei. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 567–574. Springer, Heidelberg (2012)
8. Haralick, R., Shapiro, L.: Computer and Robot Vision, vol. 1, ch. 5. Addison-Wesley Publishing Company (1992)

# Polyps Flagging in Virtual Colonoscopy

Marcelo Fiori[1], Pablo Musé[1], and Guillermo Sapiro[2]

[1] Universidad de la República, Uruguay
[2] Duke University, Durham, NC 27708, USA

**Abstract.** Computer tomographic colonography, combined with computer-aided detection, is a promising emerging technique for colonic polyp analysis. We present a complete pipeline for polyp detection, starting with a simple colon segmentation technique that enhances polyps, followed by an adaptive-scale candidate polyp delineation and classification based on new texture and geometric features that consider both the information in the candidate polyp and its immediate surrounding area. The proposed system is tested with ground truth data, including challenging flat and small polyps. For polyps larger than $6mm$ in size we achieve 100% sensitivity with just 0.9 false positives per case, and for polyps larger than $3mm$ in size we achieve 93% sensitivity with 2.8 false positives per case.

## 1 Introduction

Colorectal cancer is nowadays the third leading cause of cancer-related deaths worldwide. The early detection of polyps is fundamental, allowing to reduce mortality rates up to 90%. Nowadays, optical colonoscopy (OC) is the most used detection method due in part to its high performance. However, this technique is invasive and expensive, making it hard to use in large screening campaigns.

Virtual Colonoscopy (VC) is a promising alternative technique that emerged in the 90's, which uses volumetric Computed Tomographic data of the cleansed and air-distended colon. It is less invasive than optical colonoscopy, and much more suitable for screening campaigns once its performance is demonstrated.

However, it takes more than 15 minutes for a trained radiologist to complete a VC study, and the overall performance of OC is still considered better. In this regard, Computer-Aided Detection (CAD) algorithms can play a key role, assisting the expert to both reduce the procedure time and improve its accuracy.

Flat polyps (those having $< 3mm$ of elevation above the mucosa) and "small" polyps are of special interest because these are an important source of false negatives in VC, and many authors claim that flat polyps are around 10 times more likely to contain high-grade epithelial dysplasia [1]. The goal of this work is to exploit VC to automatically flag colon regions with high probability of being polyps, with special attention to challenging small and flat polyps.

Automatic polyp detection is a very challenging problem, not only because the polyps can have different shapes and sizes, but also because they can be located in very different surroundings. Most of the previous work on CAD of colonic

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 181–189, 2013.

polyps consists in a segmentation step followed by a classification stage based on geometric features, some using additional texture information, but none of them takes into account the information of the tissues *surrounding* the polyp. On the other hand, for the segmentation step, not much work has been done in comparing the smoothing techniques to see which one is more adapted to polyp detection. To the best of our knowledge, no algorithm reported in the literature can detect small polyps properly, and for polyps larger than $6mm$ in size, no algorithm can achieve 100% sensitivity with less than one false positive per case.

The proposed system is illustrated in Figure 1, and consists of the following steps: colon segmentation, an adaptive-scale search of candidates in order to capture the appropriate size, computation of geometrical and textural features, and a machine learning algorithm to classify patches as polyps or normal tissue.

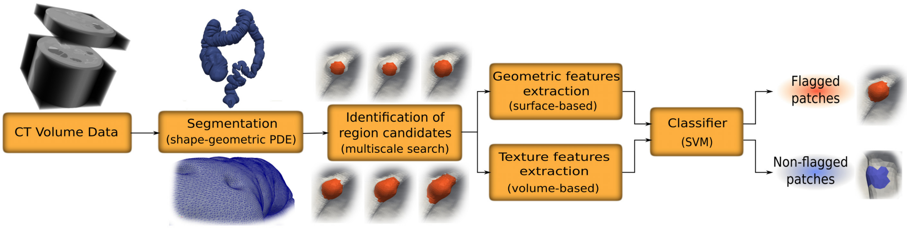

CT Volume Data
Segmentation (shape-geometric PDE)
Identification of region candidates (multiscale search)
Geometric features extraction (surface-based)
Texture features extraction (volume-based)
Classifier (SVM)
Flagged patches
Non-flagged patches

**Fig. 1.** Basic pipeline of the proposed polyp flagging system

## 2 Summary of the Colon Segmentation Method

The segmentation of the colon surface, which is critical in particular to compute geometric features, is divided into two parts: a pre-processing stage for dealing with the air-fluid composition of the colon volume, and a second stage that consists on smoothing the pre-processed image and obtaining the final colon surface by thresholding the smoothed volume. More details are available in [2].

### Classifying CT Regions

All the database cases have the same preparation, which includes solid-stool tagging and opacification of luminal fluid. Figure 2 shows a CT slice and its pixel values over the highlighted vertical profile. There are 3 clearly distinguishable classes: lowest gray levels correspond to air, highest levels to fluid, and middle gray values to tissue. However, there are around 6 interface voxels between air and fluid whose gray values lie within the normal tissue range. Therefore, a naïve approach is not suitable for tissue classification. We propose to compute a volume $u_0$ intended to have homogeneous values in the colon interior and exterior, and a smooth transition between them. To do that, we assign to each voxel the likelihood of being air, fluid, or air-fluid interface. Air and fluid distributions are estimated using standard kernel density estimation methods; these functions are then used to assign air and fluid likelihood values to the voxels.

Note that this assignment fails on the air-fluid and air-fluid-tissue interfaces. For assigning a value to these voxels, we take advantage of the physics of the

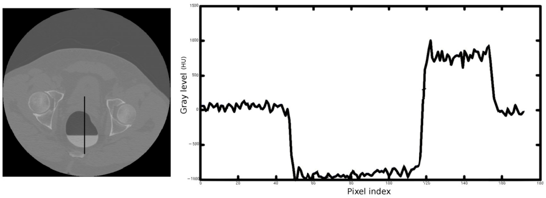


**Fig. 2.** CT slice and its different gray values for air, fluid and normal tissue, along the vertical profile

problem: the subject is laid horizontally so the interface between the fluid and the air is a plane parallel to the floor. The voxels situated on the interface then have a large gradient in the vertical direction. The implementation of these criteria is as follows. A cubic neighborhood around each voxel $\mathbf{x}$ is considered, and for each one of the "columns" that result of fixing the $x$ and $y$ coordinates, the air-likelihoods of the upper voxels and the fluid-likelihoods of the lower voxels are accumulated. The value $IC(\mathbf{x})$ that represents the confidence level of $\mathbf{x}$ being an interface voxel is then an increasing function of this accumulated measures.

We then assign to the initial segmentation $u_0$ the maximum of these three values, namely, the air and fluid likelihoods and the interface confidence level.

It is not rare that segmentation algorithms result in "gutter-like" shapes along the air-fluid-tissue interface. This is a critical point because of the potential of yielding several FPs in the detection step. If small oscillations occur along the "gutter" (which is expectable), artifacts with polyp-like shape are produced, thus degrading the overall performance. We paid special attention to this issue: the $IC$ computation allows to avoid these artifacts. Figure 2 shows the comparison of our segmentation with a version of the method without the $IC$ computation.

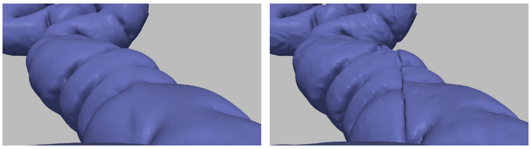


**Fig. 3.** Comparison of reduced artifacts in our segmentation (left) with a previously tested more standard version (right)

### Smoothing and Colon Surface Computation

In order to eliminate noise and to obtain a smoother colon surface after the segmentation stage, we proceed to smooth the initial segmentation $u_0$. We derive a PDE-driven smoothing technique that preserves the shape of the polyps, while obtaining a smooth enough surface to reliably compute local geometric features.

We concentrate on a family of smoothing PDEs of the form

$$\frac{\partial u(\mathbf{x},t)}{\partial t} = \beta|\nabla u| \quad , \quad u(\mathbf{x},0) = u_0(\mathbf{x}) \ , \tag{1}$$

where the initial volume $u_0$ results form the preprocessing described in the previous section. After a few iterations of this evolution, the inner colonic wall will be extracted as a suitable iso-level surface of the resulting $3D$ image $u(\mathbf{x},T)$.

We recall that the Level Set Method [3] states that if $u(\mathbf{x},t)$ evolves according to (1), then its iso-levels (level sets) satisfy $\frac{\partial \mathcal{S}}{\partial t} = \beta\boldsymbol{\mathcal{N}}$, where $\mathcal{S}$ is any iso-level surface and $\mathcal{N}$ its unit normal. This geometric view enables to design $\beta$ to fulfill a set of requirements we will impose to the surface evolution. In particular, we are interested in motions driven by the principal curvatures $\kappa_{max}$ and $\kappa_{min}$.

With the mean curvature motion ($\beta = \mathcal{H}$), and the affine motion ($(K^{+})^{1/4}$), the polyps are flattened too fast [2]. As an alternative, a motion that seems to be well suited for our problem is the motion by minimal curvature. Indeed, polyps have a curve of inflection points all around it, separating its upper and lower sections. Along this curve, the minimal curvature is $\kappa_{min} = 0$, and therefore this part of the polyp does not move (or moves very slowly), so intuitively under this motion the polyps should persist longer. This PDE already yields very good results in terms of both surface smoothing and polyp enhancement.

We further derive two modifications that lead us to the proposed PDE. The first one is inspired by the exponent 1/4 of the affine motions in dimension 3: $\frac{\partial \mathcal{S}}{\partial t} = \kappa_{min}^{1/4}\mathcal{N}$. Figure 4 shows the result after a few iterations, and Figure 5 evidences the difference between the motions by $\kappa_{min}$ and $\kappa_{min}^{1/4}$ (gray and orange respectively) with a comparative image. On the polyp protrusion, the orange surface is above the gray one, while the opposite is observed in the surrounding area, showing that the evolution by $\kappa_{min}^{1/4}$ leads to better polyp enhancement.

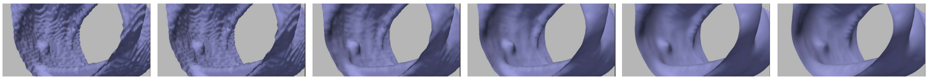
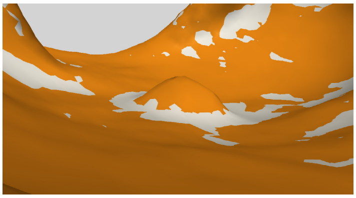


**Fig. 4.** Evolution by $\kappa_{min}^{1/4}$: original surface and result after 2, 8, 15, 30 and 50 iterations

**Fig. 5.** Comparison between evolutions. Motion by $k_{min}$ in light gray vs. motion by $k_{min}^{1/4}$ in dark gray. Both surfaces are overlaid, so sections that are not visible are hidden below the other surface.



The second modification is based on the idea of preserving the polyps qualities that we later use to identify them. A measure of the local shape of a surface is the so-called *shape index SI*, and the complementary *curvedness C* [4]:

$$SI := -\frac{2}{\pi}\arctan\left(\frac{\kappa_{max}+\kappa_{min}}{\kappa_{max}-\kappa_{min}}\right) \quad , \quad C := \frac{2}{\pi}\ln\sqrt{\frac{\kappa_{max}^2+\kappa_{min}^2}{2}} \ .$$

While the value of $SI$ is scale-invariant and measures the local shape of the surface, the value of $C$ indicates how pronounced it is. We now include this information in order to make potential polyps evolve differently than the rest of the colon surface. We define a function of the shape index that acts as a multiplying factor to the term $\kappa_{min}^{1/4}$, making the surface evolve slower at the interest points. These function should assign low values to shape index near $-1$, and values close to unity to other points. A smooth function $g(SI)$ verifying these constraints is $g(SI) = \frac{1}{\pi}\arctan\left((SI - 0.75)\cdot 10\right) + \frac{1}{2}$.

The final evolution keeps all the advantages of the motion by $\kappa_{min}^{1/4}$ and in addition, polyps are flattened slower:

$$\frac{\partial \mathcal{S}}{\partial t} = g(SI)\,\kappa_{min}^{1/4}\boldsymbol{\mathcal{N}}\ . \tag{2}$$

The number of iterations can be set by experimentally choosing the value that maximizes the overall performance of the system, measured in terms of the free-response ROC curve (FROC). Alternatively, we can consider a sphere of the size of the CT resolution and compute analytically the number of iterations needed to make it vanish (see [2]). These two approaches led to the same result, namely 15 iterations, and therefore this is the chosen value for the experiments.

At this point we have a smoothed volume $u(\mathbf{x}, T)$ indicating the volume inside of the colon. We then extract the surface of the colon as the iso-value surface of level $\alpha \in [0, 1]$. The choice of the value $\alpha$ can be made by maximizing some criteria, in order to obtain the most contrasted surface in a given sense.

## 3 Polyp Delineation, Feature Extraction and Classification

All the polyp detection methods reported try to classify polyps from properties defined only within the candidate region. However, it is important to analyze the spatial context in which the candidate patch is located, not only because different sections of the colon present different characteristics, but also because polyps can be situated over different structures such as folds or plain colonic wall. In this regard, most of the features here described take into account the information of the area surrounding the candidate patch. This makes the features more robust to the local phenomena. The normal tissue of different cases may vary, so absolute thresholds lack meaning; while texture patterns differ from study to study, what does not vary is the fact that polyps have different properties than normal tissue.

**Candidate Detection and Geometrical Features**

Consider the shape index as a function $SI : \mathcal{S} \rightarrow [-1, 1]$, and recall that the polyps' $SI$ are close to $-1$. Therefore, a region of the surface corresponding to a polyp has at least one local minimum of $SI$. Detection of candidate patches follows an adaptive-scale search: for each local minimum $x_0 \in \mathcal{S}$ of the function $SI$, several level sets of $SI$ $(\mathcal{P}_1 \ldots \mathcal{P}_n)$ around $x_0$ are tested, and the level set $\mathcal{P}_i$ that maximizes the distances between the histograms described below, is the considered candidate patch, denoted by $\mathcal{P}$ (Fig. 6). A total of $n = 7$ level sets are tested, corresponding to $SI$ values from $-0.8$ to $-0.5$ with a 0.05 step. The following description is given for the final chosen patch $\mathcal{P}$, but the computations are made for all the level sets $\mathcal{P}_i$ in order to select the most appropriate one.

Given a candidate patch $\mathcal{P}$, a ring $\mathcal{R}$ around $\mathcal{P}$ is computed, in order to consider geometrical measurements with respect to the area surrounding the patch. The ring is calculated by dilating the patch $\mathcal{P}$ a certain geodesic distance, such that the areas of $\mathcal{P}$ and $\mathcal{R}$ are equal, see Figure 7.

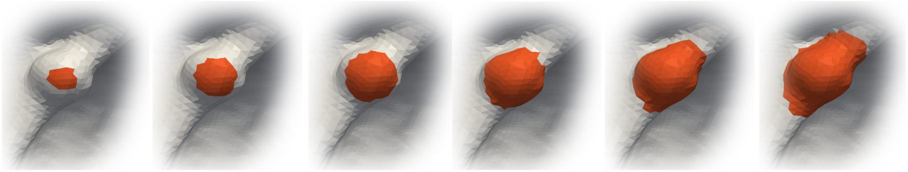
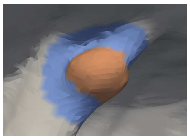


**Fig. 6.** $\mathcal{P}_1 \ldots \mathcal{P}_n$: different sizes are tested in order to select the most appropriate patch

**Fig. 7.** Ring, in blue, surrounding a candidate polyp (actually a true polyp), in orange



Histograms of the shape index values are then computed for the patch $\mathcal{P}$ and the ring $\mathcal{R}$, and two different distances between them are computed: the $L_1$ distance and the symmetric Kullback-Leibler divergence. If the patch corresponds to a polyp-like shape then the values of the $\mathcal{P}$ histogram will be concentrated around $-1$. The histogram of $\mathcal{R}$ will be concentrated near 1 in case of a polyp on a normal colon wall (concave), or around $-0.5$ if the polyp is on a fold. These two features give a measure of the geometric local variation of the candidate patch $\mathcal{P}$. Although these two distances are the most discriminative features, we also consider the following ones since they help discriminating typical false positives:

- The mean value of the shape index over the patch $\mathcal{P}$.
- The area of the patch.
- The growth rate at the adaptive-size stage, meaning the ratio between the area of the chosen patch $\mathcal{P} = \mathcal{P}_i$ and the area of the immediately smaller patch $\mathcal{P}_{i-1}$; this feature measures how fast the shape of the patch is changing.
- And finally the *shape factor* $SF = \frac{4\pi \cdot Area}{Perimeter^2}$, which measures how efficiently the perimeter is used in order to gain area. It favors circle-like patches (like the polyp in Fig. 7), avoiding elongated patches (like false positives in folds).

We then end-up with a total of 6 geometric features.

### Texture Features

There is evidence that the gray-level of the CT image and its texture can be very helpful for detecting polyps. This is particularly useful for flat or small polyps, where geometric information is limited [5]. Some work has been done on the inclusion of texture features (inside the candidate polyps only), in order to reduce false positives [6]. We propose both the use of new texture features and the inclusion of the information on the candidate's surrounding area.

First, for each polyp candidate $\mathcal{P}$, a volume $V_1$ is computed, containing the patch $\mathcal{P}$ and a portion of the inner tissue bounded by the patch. Volume $V_1$ is obtained by dilating $\mathcal{P}$ (in 3D) towards the inner colon tissue. A second volume $V_2$ surrounding $V_1$ is computed dilating $V_1$. The tissue in $V_2$ is intended to be normal, to be compared with the polyp candidate tissue. The dilation is chosen as before: several distances are tested, keeping the most discriminative one.

The chosen texture features are a subset of the classical Haralick features, namely, entropy, energy, contrast, sumMean, and homogeneity. Seven cooccurrence matrices are computed with the voxels of $V_1$, and the five features are averaged over the seven directions. The analogous computation is made for $V_2$, and the differences between the two volumes, for each texture feature, are considered. Additionally, the mean gray levels in both volumes is computed, and their difference is considered as a feature. In this way, six texture features are considered.

**Classification**

Once the the candidates detection has been performed, the number of true polyps was much lower than the number of non-polyps patches, a relation on the order of 500:1, which is a significant problem for the learning stage of the classifier, since most classifiers are designed to maximize the accuracy, which is not adequate for imbalanced problems [7]. For instance, if we classify all candidates as "non-polyps," we would get an accuracy of 99.8% but without detecting any polyps. Three techniques were considered to overcome this problem: MetaCost, Cost Sensitive Learning (CSL), and Synthetic Minority Over-sampling TEchnique (SMOTE). The best results were obtained with CSL+SVM.

## 4 Results

A total of 150 patients of the Walter Reed Army Medical Center database [8] were used to test our CAD algorithm. The database contains 134 polyps detected by OC, including 12 flat polyps. Among these 134 polyps, 86 are larger than $6mm$, and 48 range from $3mm$ and $6mm$. The evaluation was carried out by splitting the dataset into halves, training and testing. Under this setting, classification with CSL+SVM yields the FROCs in Fig. 8, which shows the performance for different polyps sizes. These values are comparable with state-of-the-art results [6,9], however our study includes very small polyps. A more precise comparison is not necessarily meaningful, since in general each work considers its own database.

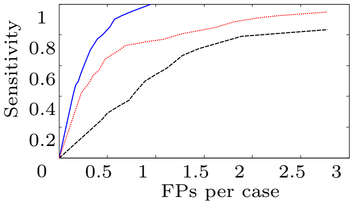


**Fig. 8.** FROC of our method for different polyps sizes: larger than $6mm$ (solid), smaller than $6mm$ (dashed), and all polyps (dotted)

The FROCs in Figure 9 compare the performance of our system when using different smoothing schemes (Section 2). The chosen one yields the best results.

The FROCs in Fig. 10 compare the influence of absolute and differential texture features. The classification was performed using all the geometric features, and either absolute (computed just for $V_1$) or differential texture features.

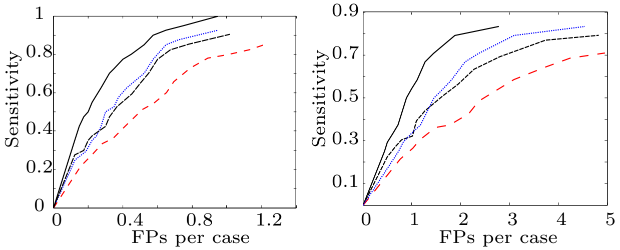


**Fig. 9.** FROCs comparing different smoothing methods, classifying large (left) and small polyps (right). The curve for the proposed evolution is shown in solid line, the results for the evolution by $\mathcal{H}$ and $\kappa_{min}$ are shown in dotted and dashed lines respectively, and the lower curve is the result when no smoothing is performed.

The results show that differential texture features are more discriminative than the absolute ones. Finally the FROCs in Fig. 11 compare the results of different classification approaches. CSL, SMOTE, and MetaCost were used as a pre-processing stage for SVM, and C4.5 trees stabilized with AdaBoost. Parameters in all classifiers were optimized via cross validation.

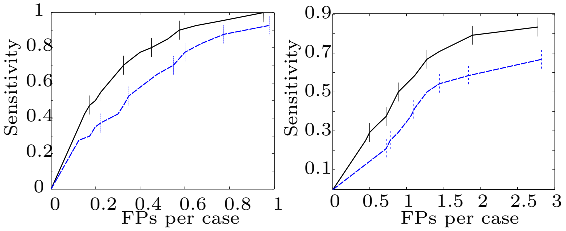


**Fig. 10.** FROCs (95% confidence intervals), comparing the performance with differential (solid) and absolute (dashed) texture features, for polyps larger (left) and smaller (right) than $6mm$

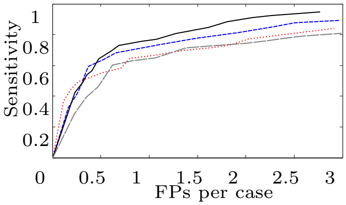


**Fig. 11.** FROCs comparing the performances of different classification approaches. SVM+CSL (solid), SVM+SMOTE (dashed), C4.5+AdaBoost (dotted) and plain SVM (long-dashed).

## 5 Conclusion

We introduced a complete pipeline for a Computer Aided Detection algorithm that flags candidate polyp regions. The segmentation stage is very simple and fast, and its main novelty is the smoothing PDE which enhances the polyps, enabling better detection rates. In addition to the incorporation of the Haralick texture features, the main yet simple novelties of the proposed features and classification stages are twofold. First, the surrounding area of candidate polyps are explicitly taken into account. Indeed, the proposed (so-called differential) features are computed by comparing properties in the central and surrounding regions of the polyps. We show that differential features are more discriminative than the absolute ones, as they emphasize local deviations of geometry and texture over the colon. The other novelty is an adaptive-scale strategy that test regions of different sizes and automatically selects the region that best delineates each candidate polyp. The obtained quantitative results are very promising.

## References

1. Fidler, J., Johnson, C.: Flat polyps of the colon: accuracy of detection by CT colonography and histologic significance. Abdom Imaging 34(2), 157–171 (2009)
2. Fiori, M., Musé, P., Sapiro, G.: A complete system for candidate polyps detection in virtual colonoscopy (2012), `http://arxiv.org/abs/1209.6525`
3. Osher, S., Sethian, J.: Fronts propagating with curvature- dependent speed: Algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. 79, 12–49 (1988)
4. Koenderink, J.J.: Solid Shape. MIT Press, Cambridge (1990)
5. Fiori, M., Musé, P., Aguirre, S., Sapiro, G.: Automatic colon polyp flagging via geometric and texture features. In: IEEE EMBS, pp. 3170–3173 (2010)
6. Wang, Z., Liang, Z., Li, L., Li, X., Li, B., Anderson, J., Harrington, D.: Reduction of false positives by internal features for polyp detection in CT-based virtual colonoscopy. Medical Physics 32(12), 3602–3616 (2005)
7. Martino, M.D., Hernández, G., Fiori, M., Fernández, A.: A new framework for optimal classifier design. Pattern Recognition 46(8), 2249–2255 (2013)
8. Pickhardt, P., Choi, J., Hwang, I., Butler, J., Puckett, M., Hildebrandt, H., Wong, R., Nugent, P., Mysliwiec, P., Schindler, W.: Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. New England Journal of Medicine 349(23), 2191–2200 (2003)
9. Suzuki, K., Yoshida, H., Näppi, J., Armato, S.G., Dachman, A.H.: Mixture of expert 3D massive-training ANNs for reduction of multiple types of false positives in CAD for detection of polyps in CT colonography. Medical Physics 35(2), 694 (2008)

# Predicting HIV-1 Protease and Reverse Transcriptase Drug Resistance Using Fuzzy Cognitive Maps

Isel Grau, Gonzalo Nápoles, and María M. García

Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
{igrau,gnapoles,mmgarcia}@uclv.edu.cu

**Abstract.** Several antiviral drugs have been approved for treating HIV infected patients. These drugs inhibit the function of proteins which are essential in the virus life cycle, thus preventing the virus reproduction. However, due to its high mutation rate the HIV is capable to develop resistance to administered therapy. For this reason, it is important to study the resistance mechanisms of the HIV proteins in order to make a better use of existing drugs and design new ones. In the last ten years, numerous statistical and machine learning approaches were applied for predicting drug resistance from protein genome information. In this paper we first review the most relevant techniques reported for addressing this problem. Afterward, we describe a Fuzzy Cognitive Map based modeling which allows representing the causal interactions among the protein positions and their influence on the resistance. Finally, an extended comparison experimentation is carried out, which reveals that this model is competitive with well-known approaches and notably outperforms other techniques from literature.

**Keywords:** HIV proteins, Drug resistance, Prediction, Fuzzy Cognitive Maps.

## 1 Introduction

In the last two decades several antiretroviral (ARV) drugs have been designed for treating Human Immunodeficiency Virus (HIV). The main goal of ARV therapies is to inhibit the function of essential proteins for the virus life cycle such as *protease*, *reverse transcriptase* or *integrase*. For instance, the *reverse transcriptase* protein catalyzes the reverse transcription process, which transforms RNA into DNA and incorporates the resulting DNA into the host cell. As a result, the infected cell produces viral particles which are maturated by *protease* protein, cleaving precursor proteins and therefore completing the virus replication process. Evidently, inhibiting these processes could help on preventing the virus reproduction. Nevertheless, due to its high mutation rate, the HIV is capable to develop resistance to administered drugs, evading the immune system and causing the therapy failure.

Consequently, understanding resistance mechanisms is critical for designing more effective treatment strategies or even developing new ARV drugs [1]. In general, the resistance testing of an observed mutation can be performed by two main approaches: the genotypic and the phenotypic tests. The genotypic testing is based on identifying drug-resistance mutations (and the combination of them) which have been associated to decreased susceptibility of a target drug; while phenotypic testing measures the

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 190–197, 2013.

viral replication in presence of different ARV concentrations. In clinical practice, genotype assays are more frequently used than phenotype ones since they are less expensive in time and effort; however, they only provide indirect evidence of resistance. On the other hand, phenotypic testing is more useful for determining the susceptibility of new approved ARV drugs, where patterns on resistance have not yet been well described. Phenotype assays are clinically suitable for viruses with complex mutational patterns where genotype interpretation becomes really difficult [1].

The paired results of such tests, performed for several protein mutations, constitute a valuable historical data in order to understand the HIV behavior. Based on this knowledge, in the last ten years several statistical and machine learning methods have been proposed for predicting the phenotype resistance to a target drug from the genotype information (known as virtual phenotype), that is, the resistance degree of a mutation (target attribute) given its amino acid sequence (predictive features). In some cases these data-driven approaches lead to parsimonious models, but in general they are harder to interpret [2].

In a recent attempt to use more interpretable techniques, in previous works [3, 4] the authors proposed a model based on Fuzzy Cognitive Maps (FCM) [5] with the goal of discovering knowledge on the causal patterns among the sequence positions and the phenotype resistance. Although this research was mainly focused on the causality interpretation of learned maps, we observed that the prediction accuracies notably outperformed several well-known classifiers. Inspired on this result, we propose an extended comparison experiment for measuring the accuracy of this model using historical data from several *protease* and *reverse transcriptase* inhibitors. Before, we first review the most relevant techniques reported in the last ten years for addressing this classification problem. In addition, the FCM model is described in Section 3, but now it is investigated from the prediction point of view.

## 2 Computational Approaches for Drug Resistance Analysis

Since not only the number of approved ARV drugs is increasing, but also the resistant mutations to these treatments, the use of intelligent systems have progressively become more important for understanding the resistance phenomena in HIV proteins. Actually, in the last years the use of computational methods for prediction or interpretation of HIV drug resistance has been growing. In general terms, such models constitute a very useful tool for guiding physicians in designing complex individual therapies and drug experts in the development of new ARV [1].

At beginning several rule-based systems were introduced, using the knowledge of physicians and also data about mutations previously associated with resistance from clinical trials. This is the case of Rega [6], ANRS [7] and VGI systems [8]. In fact, at the same time the Stanford HIV Drug Resistance Database project enabled the public access to an algorithm known as HIVdb [9]. This approach uses, in addition to rules, a drug penalty score for inferring five levels of resistance. Moreover, this project has a platform (HIValg) for comparing the output of several drug interpretation algorithms; which was used in [10] for determining those relevant mutation patterns responsible of observed discordance among the investigated rule-based approaches.

Subsequently, more accurate computational techniques such as support vector regression were employed in Geno2Pheno [11], which is another web service for drug resistance prediction. Also, different types of neural networks were explored [12-14] where bidirectional recurrent neural networks [15] reported competitive performance in terms of accuracy. Regression models also had been studied; for instance, a standard stepwise linear regression [16] outperformed other genotypic interpretation algorithms publicly available so far, including decision trees, support vector machine and four rule-based algorithms (HIVdb, VGI, ANRS and Rega). Later, in [17] Rabinowitz et al. introduced two regression techniques using convex optimization and perform a comparison against the most relevant approaches at the moment.

On the other hand, in the same year Rhee et al. [18] published the results of five previously proposed predictors including decision trees, linear regression, linear discriminant analysis, neural networks, and support vector regression, using high quality filtered knowledge bases. These historical data is publicly available for experimental comparisons of new algorithms. More recently, in reference [19] was described a linear regression called itemset boosting that works particularly well for predicting the resistance of nucleotide *reverse transcriptase* inhibitors. As well, in [20] least-angle regression was performed to identify *protease* mutations associated with reduced susceptibility to at least one *protease* inhibitor, and least-squares regression was employed in order to quantify the contribution of *protease* mutations to reduced susceptibility. Finally, in [21] the author implements a procedure based on n-grams to generate sequence attributes; where results are complementary to other sequence-based approaches, reporting competitive features in performance.

## 3 A Model Based on Fuzzy Cognitive Maps Theory

In this section we briefly describe the FCM model proposed in [3, 4] which was conceived for studying the causal influence of the *protease* protein positions on the resistance when a mutation occurs. Next, we explain the generalization of this model for any other HIV protein as a tool for describing the drug resistance activity.

As a first step each protein position is represented as a map concept, while another node for denoting the resistance degree to a specific drug is also defined. Afterwards, causal connections among all input concepts are created, representing the interaction (causal influence) among all protein positions. Also, connections between each map concept and the final resistance concept are established. This topology is supported by the fact that there exist relations among not necessarily adjacent positions of the sequence due to the three-dimensional structure of the protein; where a change in the amino acid of a specific position (i.e. mutation) could be relevant for the drug resistance [4]. For better understanding of this scheme, following figure 1 illustrates the general conception of the FCM that results from this stage.

Then, in order to determine the causality among positions and the resistance variable, a learning process based on Swarm Intelligence is carried out. This learning algorithm uses historical data publicly available for finding a causal matrix that minimizes the difference between the reported resistance and the value of the resistance concept (map inference), for all mutations reported [3, 4].

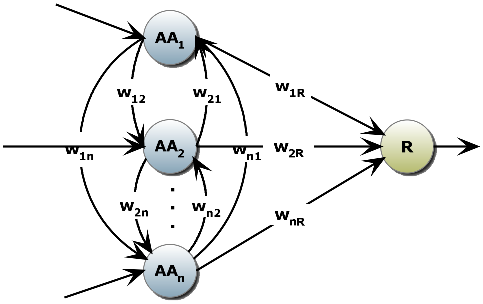


**Fig. 1.** Topology for describing a HIV protein through the FCM theory. Concepts denote positions of the sequence, and links stand for causal connection among amino acids.

It is fair to mention that, with the purpose of reducing the model dimensionality, only sequence positions previously associated with resistance are considered. These positions are selected from both numerical and biological perspective, as a result of available research in this field. As a result, predicting the resistance target from descriptors nodes means to solve the related classification/regression problem as follow. First, the activation value of each concept is taken as the contact energy [22] of corresponding amino acid normalized in the range $[0,1]$. As a second step, the map inference process is triggered and the value of the resistance concept is examined. For example, for a regression perspective this value will denote the normalized degree of resistance; whereas for a classification perspective, a drug-specific cut-off for determining the resistance class (0-susceptible and 1-resistant) should be used.

It is also remarkable that typical FCM can't solve classification problems [23]; however from empirical simulations we noticed that this model was able to outperform other well-known classifiers. In fact, in the following section we carry out an extensive set of experiments for fully exploring the prediction ability of this model against other classifiers for both *protease* and *reverse transcriptase* proteins.

## 4 Simulations and Results

In the present section we study the inference ability of the FCM model for solving the bioinformatics problem enunciated before. To do that, we use historical data associated with 7 *protease* inhibitors and 11 *reverse transcriptase* inhibitors taken from [9]. The *protease* inhibitors used in this work are: Amprenavir (APV), Atazanavir (ATV), Indinavir (IDV), Lopinavir (LPV), Nelfinavir (NFV), Ritonavir (RTV) and Saquinavir (SQV). While two kinds of *reverse transcriptase* inhibitors are used: nucleoside/nucleotide and nonnucleoside inhibitors. The nucleoside/nucleotide ones are: Lamivudine (3TC), Abacavir (ABC), Zidovudine (AZT), Stavudine (D4T), Zalcitabine (DDC), Didanosine (DDI), Emtricitabine (FTC) and Tenofovir (TDF); whereas nonnucleoside are: Delavirdine (DLV), Efavirenz (EFV) and Nevirapine (NVP). Then, we evaluate the FCM model against some approches from literature, for solving the related classification problem having two and three classes.

### 4.1 Solving the Sequence Classification Problem for Two Classes

The idea here is to compare the accuracy of the FCM model against other classifiers for solving the binary classification problem (susceptible and resistant). In all cases we use the cut-off values reported in [9] as thresholds for determining the class of each instance. In addition we use the following parameters settings for the learning algorithm [3,4]: 80 particles, five variable neighborhoods, 200 generations, and the allowed number of generations without progress is set to 40. For comparison we used next methods: Support Vector Machine with linear kernel (SVML), polynomial of degree1 (SVM1), degree 2 (SVM2), degree 3 (SVM3), and radial basis (SVMR); in addition we used a Multilayer Perceptron (MLP) and a Bidirectional Recurrent Neural Network (BRNN), all taken from [15]. To conclude, we consider an Artificial Neural Network (ANN) from [14] and a novel ensemble classifier from [24] called Multi-Expert by Hard Instances (MEHI) which has reported promising accuracies.

Table 1 shows the accuracy from a 10-fold cross-validation process using data of six *protease* inhibitors, corresponding to the complete unfiltered datasets of the Phenosense assay. From these results the following conclusions are drawn: for ATV, SQV, LPV and IDV the investigated model notably outperforms other algorithms, while for RTV and NFV it reports quite competitive results (where MEHI computes the best accuracies). However, it is remarkable that, for ATV the FCM model is able to outperform MEHI in 12 percent points, whereas for drugs RTV and NFV the FCM model and the MEHI algorithm only differs at most in two percent points.

In literature, there are few reports concerning to algorithms for solving this binary classification problem for *reverse transcriptase* inhibitors using datasets with complete sequence. Despite this inconvenient, in [25] Grau et al. proposed a Recurrent Neural Network (RNN) that uses a modified backpropagation through time algorithm for dealing with instances of variable length, allowing handling the complete sequence. Following Table 2 summarizes the comparison accuracies between the FCM model and the RNN approach. Here the FCM model largely outperforms the RNN at most inhibitors, which confirm the suitability of FCMs to deal with this problem.

**Table 1.** Classification accuracy obtained for *protease* inhibitors (two classes)

| Drug | SVML | SVM1 | SVM2 | SVM3 | SVMR | ANN | MLP | BRNN | MEHI | FCM |
|---|---|---|---|---|---|---|---|---|---|---|
| ATV | 0.78 | 0.71 | 0.68 | 0.72 | 0.70 | - | 0.80 | 0.81 | 0.84 | 0.96 |
| SQV | 0.87 | 0.80 | 0.69 | 0.85 | 0.82 | 0.91 | 0.85 | 0.91 | 0.85 | 0.95 |
| LPV | 0.88 | 0.85 | 0.85 | 0.88 | 0.85 | 0.92 | 0.92 | 0.94 | 0.93 | 0.98 |
| RTV | 0.91 | 0.84 | 0.79 | 0.92 | 0.86 | 0.96 | 0.90 | 0.94 | 0.99 | 0.97 |
| IDV | 0.91 | 0.86 | 0.83 | 0.92 | 0.88 | 0.95 | 0.86 | 0.92 | 0.97 | 0.99 |
| NFV | 0.84 | 0.75 | 0.70 | 0.84 | 0.80 | 0.95 | 0.86 | 0.93 | 0.96 | 0.95 |

**Table 2.** Classification accuracy obtained for *reverse transcriptase* inhibitors (two classes)

| Model | 3TC | ABC | AZT | D4T | DDC | DDI | DLV | EFV | FTC | NVP |
|---|---|---|---|---|---|---|---|---|---|---|
| RNN | 0.70 | 0.61 | 0.56 | 0.89 | 0.78 | 0.81 | 0.58 | 0.56 | 0.86 | 0.65 |
| FCM | 0.96 | 0.94 | 0.98 | 0.93 | 0.80 | 0.79 | 0.92 | 0.95 | 0.98 | 0.95 |

### 4.2 Solving the Sequence Classification Problem for Three Classes

In this subsection we extend the experimentation by comparing the FCM approach against other classifiers for solving the related classification problem, now using three classes (susceptible, intermediate and resistant). Besides, the same parameter setting of the FCM learning algorithm used in the above section is adopted. At this moment two kind of datasets are studied; the first ones corresponds to the complete unfiltered dataset, while the second are high quality filtered datasets both available in [9]. For comparison we used following approaches: Random Forest classifier using relative frequency approach (RF1) and Random Forest classifier using a counts method (RF2), Reduced-Error Pruned Tree with relative frequency procedure (REPT1) and Reduced-Error Pruned Tree with counts approach (REPT2), all taken from [21]. Also, we consider: Decision Trees (DT), Neural Networks (NN), Least-Squares Regression (LSR), Support Vector Regression (SVR) and also Least-Angle Regression (LARS), all taken from [18]. Table 3 and 4 show the computed accuracy from a 5-fold cross-validation process for *protease* and *reverse transcriptase* inhibitors.

**Table 3.** Classification accuracy obtained for *protease* inhibitors (three classes)

| | **High quality filtered datasets** | | | | | | **Complete unfiltered datasets** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Drug** | **SVR** | **LSR** | **LARS** | **DT** | **NN** | **FCM** | **REPT1** | **RF1** | **REPT2** | **RF2** | **FCM** |
| APV | 0.82 | 0.81 | 0.81 | 0.77 | 0.74 | 0.61 | - | - | - | - | - |
| ATV | 0.69 | 0.68 | 0.76 | 0.71 | 0.64 | 0.70 | 0.74 | 0.75 | 0.76 | 0.76 | 0.78 |
| IDV | 0.77 | 0.78 | 0.77 | 0.75 | 0.73 | 0.78 | 0.78 | 0.80 | 0.75 | 0.80 | 0.72 |
| LPV | 0.80 | 0.79 | 0.83 | 0.77 | 0.76 | 0.85 | 0.80 | 0.82 | 0.80 | 0.81 | 0.78 |
| NFV | 0.79 | 0.79 | 0.80 | 0.76 | 0.73 | 0.70 | 0.80 | 0.80 | 0.79 | 0.82 | 0.75 |
| RTV | 0.86 | 0.86 | 0.88 | 0.84 | 0.81 | 0.80 | 0.87 | 0.86 | 0.87 | 0.84 | 0.79 |
| SQV | 0.81 | 0.81 | 0.82 | 0.75 | 0.76 | 0.73 | 0.80 | 0.79 | 0.80 | 0.80 | 0.74 |

**Table 4.** Classification accuracy obtained for *reverse transcriptase* inhibitors (three classes)

| | **High quality filtered datasets** | | | | | | **Complete unfiltered datasets** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Drug** | **SVR** | **LSR** | **LARS** | **DT** | **NN** | **FCM** | **REPT1** | **RF1** | **REPT2** | **RF2** | **FCM** |
| 3TC | 0.84 | 0.83 | 0.88 | 0.90 | 0.90 | 0.86 | 0.89 | 0.87 | 0.87 | 0.90 | 0.83 |
| ABC | 0.65 | 0.63 | 0.77 | 0.69 | 0.66 | 0.68 | 0.68 | 0.68 | 0.66 | 0.67 | 0.62 |
| AZT | 0.7 | 0.64 | 0.76 | 0.70 | 0.71 | 0.83 | 0.75 | 0.75 | 0.73 | 0.70 | 0.86 |
| D4T | 0.68 | 0.66 | 0.78 | 0.75 | 0.72 | 0.78 | 0.74 | 0.79 | 0.76 | 0.78 | 0.72 |
| DDC | - | - | - | - | - | - | 0.80 | 0.75 | 0.80 | 0.76 | 0.74 |
| DDI | 0.67 | 0.61 | 0.75 | 0.74 | 0.71 | 0.85 | 0.69 | 0.73 | 0.69 | 0.71 | 0.74 |
| DLV | 0.78 | 0.73 | 0.84 | 0.84 | 0.78 | 0.67 | 0.76 | 0.70 | 0.76 | 0.71 | 0.78 |
| EFV | 0.82 | 0.78 | 0.87 | 0.84 | 0.77 | 0.63 | 0.78 | 0.74 | 0.76 | 0.73 | 0.73 |
| FTC | - | - | - | - | - | - | 0.96 | 0.83 | 0.94 | 0.89 | 0.98 |
| NVP | 0.78 | 0.74 | 0.87 | 0.91 | 0.81 | 0.91 | 0.84 | 0.79 | 0.82 | 0.77 | 0.88 |
| TDF | 0.69 | 0.46 | 0.70 | 0.68 | 0.73 | 0.66 | 0.75 | 0.75 | 0.68 | 0.74 | 0.70 |

From Table 3 and 4 it is observed that the overall performance of all classifiers is reduced regarding to the classification problem using two classes. For example, in the *protease* filtered dataset, the studied algorithm reports better accuracies for IDV and LPV, whereas for the other drugs it computes competitive results. On the other hand, for the *reverse transcriptase* filtered dataset the FCM model performs better for AZT, DDI and NVP; reporting reasonable accuracies for remaining inhibitors, except for non-nucleoside drugs DLV and EFV where the percent of correct classified instances notably decreases. Though, using the complete *reverse transcriptase* datasets, FCM is able to outperform other classifiers for AZT, DDI, DLV, FTC, and also NVP; showing competitive results for remaining inhibitors. The reduction in the performance could be due to inconsistency or imbalanced knowledge bases. Future work will be focused on improving these results by introducing a new classification strategy for FCM which takes into account these issues and uses an alternative topology and stability criteria in the inference process. In addition this study will be extended for solving the related regression problem.

## 5 Conclusions

Understanding the complex behavior of HIV includes the prediction of resistance features to existing drugs. However, predicting phenotype from genotype information involves a challenging sequence classification problem, which has been addressed in literature by using well-known classifiers, but the prediction accuracies are still unsatisfactory. Recently was proposed a model based on Fuzzy Cognitive Maps for analyzing causal patterns among positions on *protease* sequences. While this study was oriented to the knowledge discovering, we noticed that reported prediction accuracies were promising. In this paper we explored this feature, and next aspects are concluded: (i) The FCM model using two prediction classes (susceptible and resistant) significantly outperformed other evaluated classifiers for both *protease* and *reverse transcriptase* datasets, (ii) The FCM model using three prediction classes (susceptible, intermediate and resistant) decreased its performance, although it is competitive in most cases with respect to other classifiers, therefore complementing reported approaches from literature.

## References

1. Tang, M.W., Shafer, R.W.: HIV-1 Antiretroviral Resistance Scientific Principles and Clinical Applications. Drugs 72(9), 1–25 (2012)
2. Beerenwinkel, N., et al.: Computational methods for the design of effective therapies against drug resistant HIV strains. Bioinformatics 21, 3943–3950 (2005)
3. Grau, I., Nápoles, G., León, M., Grau, R.: Fuzzy Cognitive Maps for Modelling, Predicting and Interpreting HIV Drug Resistance. In: Pavón, J., Duque-Méndez, N.D., Fuentes-Fernández, R. (eds.) IBERAMIA 2012. LNCS, vol. 7637, pp. 31–40. Springer, Heidelberg (2012)
4. Nápoles, G., Grau, I., León, M., Grau, R.: Modelling, aggregation and simulation of a dynamic biological system through Fuzzy Cognitive Maps. In: Batyrshin, I., Mendoza, M.G. (eds.) MICAI 2012, Part II. LNCS, vol. 7630, pp. 188–199. Springer, Heidelberg (2013)
5. Kosko, B.: Fuzzy Cognitive Maps. Int. Journal of Man-Machine Studies 24, 65–75 (1986)

6. Laethem, K., et al.: A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. Antiviral Therapy 7, 123–129 (2002)
7. Rousseau, M.N., et al.: Patterns of resistance mutations to antiretroviral drugs in extensively treated HIV-1-infected patients with failure of highly active antiretroviral therapy. Journal of Acquired Immune Deficiency Syndromes 26, 36–43 (2001)
8. Reid, C., et al.: A dynamic rules-based interpretation system derived by an expert panel is predictive of virological failure. Antiviral Therapy 7, S91 (2002)
9. Rhee, S.Y., et al.: Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Research 31, 298–303 (2003)
10. Ravela, J., et al.: HIV-1 Protease and Reverse Transcriptase Mutation Patterns Responsible for Discordances Between Genotypic Drug Resistance Interpretation Algorithms. Journal of Acquired Immune Deficiency Syndromes 33, 8–14 (2003)
11. Beerenwinkel, N., et al.: Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Research 31, 3850–38505 (2003)
12. Draghici, S., Potter, R.B.: Predicting HIV drug resistance with neural networks. Bioinformatics 19, 98–107 (2003)
13. Woods, M., Carpenter, G.A.: Neural Network and Bioinformatic Methods for Predicting HIV-1 Protease Inhibitor Resistance. Technical Report 02215 (2007)
14. Pasomsub, E., et al.: The Application of Articial Neural Networks for Phenotypic Drug Resistance Prediction: Evaluation and Comparison with Other Interpretation Systems. Jpn. Journal of Infectious Diseases 63, 87–94 (2010)
15. Bonet, I., García, M.M., Saeys, Y., Van de Peer, Y., Grau, R.: Predicting Human Immunodeficiency Virus (HIV) Drug Resistance Using Recurrent Neural Networks. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2007. LNCS, vol. 4527, pp. 234–243. Springer, Heidelberg (2007)
16. Wang, K., et al.: Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance. Antiviral Therapy 9, 343–352 (2004)
17. Rabinowitz, M., et al.: Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. Bioinformatics 22, 541–549 (2005)
18. Rhee, S.Y., et al.: Genotypic predictors of human immunodeficiency virus type 1 drug resistance. PNAS 103, 17355–17360 (2006)
19. Saigo, H., Uno, T., Tsuda, K.: Mining complex genotypic features for predicting HIV-1 drug resistance. Bioinformatics 23, 2455–2462 (2007)
20. Rhee, S.Y., et al.: HIV-1 Protease Mutations and Protease Inhibitor Cross-Resistance. Antimicrobial Agents and Chemotherapy 54, 4253–4261 (2010)
21. Masso, M.: HIV-1 Prediction of Human Immunodeciency Virus Type 1 Drug Resistance: Representation of Target Sequence Mutational Patterns via an n-Grams Approach. In: IEEE International Conference on Bioinformatics and Biomedicine, pp. 1–6 (2010)
22. Song, H.J., et al.: An Extension to Fuzzy Cognitive Maps for Classification and Prediction. IEEE Transactions on Fuzzy Systems 19, 116–135 (2011)
23. Miyazawa, S., Jernigan, R.L.: Contacts energies Self-Consistent Estimation of Inter-Residue Protein Contact Energies Based on an Equilibrium Mixture Approximation of Residues. PROTEINS: Structure, Function, and Genetics 34, 49–68 (1999)
24. Bonet, I., et al.: Multi-Classifier Based on Hard Instances- New Method for Prediction of Human Immunodeficiency Virus Drug Resistance. Current Topics in Medicinal Chemistry 13, 685–695 (2013)
25. Grau, I., Nápoles, G., Bonet, I., Garcia, M.M.: Backpropagation Through Time Algorithm for Training Recurrent Neural Networks using Variable Length Instances. Computación y Sistemas 17, 15–24 (2013)

# Meaningful Features for Computerized Detection of Breast Cancer

José Anibal Arias, Verónica Rodríguez, and Rosebet Miranda

Universidad Tecnológica de la Mixteca, Km 2.5 Carretera a Acatlima
CP 69000 Huajuapan de León, Oaxaca, México
{anibal,veromix,rmiranda}@mixteco.utm.mx

**Abstract.** After pre-processing and segmenting suspicious masses in mammographies based on the Top-Hat and Markov Random Fields methods, we developed a mass-detection algorithm that uses gray level co-occurrence matrices, gray level difference statistics, gray level run length statistics, shape descriptors and intensity parameters as the entry of a vector support machine classifier. During the classification process we test up to 63 image features, keeping the 35 most important and obtaining 85% of accuracy score.

**Keywords:** Breast cancer, CADx, Image features, SVM.

## 1 Introduction

Breast cancer is a disease in which malignant cells grow in breast tissue. This type of cancer is more frequent in middle age women (40-49 years-old) [1] and, in Mexico, it is the primary cause of death from malignant tumors among women [7]. Mammography (X-ray picture of the breast) associated with clinical breast examination is the cheapest and most efficient method for early detection of breast cancer. Radiologists make a visual examination of mammographies searching for masses, calcifications, density asymmetries and structure distortions that reveal the presence of cancer. However, it is very difficult to search for abnormalities because of the small differences in the image densities of breast tissue and the vast range of possible abnormalities, so the task remains highly subjective and qualitative, depending mainly on the quality of the mammography and the training and experience of radiologists [10]. This is a risk, especially in third level developed countries, where there are no other diagnosis protocols widely available.

Computer-aided diagnosis (CADx) is a helpful tool that improves diagnostic accuracy assisting radiologists to make correct mammography interpretation. The detection sensitivity without CADx is around 80% and with it up to 90% [6]. The tasks a CADx system should accomplish are:

**Pre-processing.** Noise in the digitized mammogram is reduced and the general image quality is improved. Labels, tape and scanning artefacts, and pectoral muscle are removed.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 198–205, 2013.

**Segmentation.** Suspicious regions are isolated to be later classified as abnormality (true positive) or tissue (false positive).

**Feature Extraction.** Several features are obtained from the suspicious regions.

**Classification.** CADx system declares each detected region as an abnormality or normal breast tissue. Also, in this stage, if the region is an abnormality, their malignant or benign class is determined.

Several CADx systems have been developed for research purposes [15], but there is no report of any commercial system available. We intend to develop one for detection and diagnosis of masses (in a first version, identifying other abnormalities later) and make it available to public health institutions. In this work, we present the last two stages of a CADx system that identifies masses in mammographies. Masses are subtle areas (2-30 mm in diameter) with smooth boundaries and high densities and represent the most difficult type of lesion to detect and characterize.

The paper is organized as follows. In Section 2, we describe several approaches for automated detection and classification of masses in mammograms. The data used in our tests is mentioned in Section 3. The different features generated from suspicious regions are described in Section 4. The classifier and the experimental results are presented in Section 5. Finally, conclusions and future work are given in Section 6.

## 2 Related Work

Several methods have been proposed for mammography mass detection. Excellent state of art reviews are presented in [11] and [2], showing an evaluation of several methods for enhancement of mammographic images, detection and classification of masses.

Rojas and Nandi [13] proposed a three stages method to perform mass detection. The first one is a multilevel adaptative process based on local statistical measure of the pixel intensities and morphological operators to enhance breast structures. In the next stage, the images are segmented by applying thresholding and Gaussian filtering. Finally, the selection of suspicious regions is performed by means of a ranking system that uses 18 shape and intensity features. The method was tested on 57 mammographic images of masses from the MIAS database [17], and achieved a sensitivity of 80% at 2.3 false-positives per image.

An interesting method for reduction of false positives in mass detection is presented by Llado et al. [9]. The basic idea of their approach is the use of Local Binary Patterns for texture descriptions of ROIs. Support Vector Machines (SVM) with a polynomial kernel performed classification of mass and normal breast tissue. Their approach was evaluated on 1792 ROIs extracted from the DDSM mamographic database [5], and reported a mean $Az$ value (area under the ROC curve) of 0.94.

Sampaio et al. [14] proposed a methodology based on Cellular Neural Networks, geostatistic functions and Support Vector Machines. In the first step of their methodology, the images are pre-processed by using Hough Transform, K-means and morphological operators. Identification of suspicious regions is performed by segmentation with Cellular Neural Networks. A SVM classifier that uses shape and texture features is proposed with a sensitivity of 80% at 0.84 false positives per image.

## 3 Database

Our method was tested on a subset of images extracted from the Mammographic Image Analysis Society (MIAS) database [17]. This publicly available digitized database contains left and right breast images in mediolateral oblique (MLO) view that represent the mammograms of 161 patients with ages between 50 and 65. All images were digitized at a resolution of 1024×1024 pixels and at 8-bit gray scale level.

The chosen set corresponds to masses annotated as spiculated, circumscribed or miscellaneous (ill-defined masses). The summary of this dataset by type of mass and density of breast tissue is shown in Table 1.

**Table 1.** Summary of MIAS images used

| | Fatty | Fatty-Glandular | Dense-Glandular | Total |
|---|---|---|---|---|
| Circumscribed | 13 | 8 | 3 | 24 |
| Miscellaneous | 8 | 5 | 2 | 15 |
| Spiculated | 5 | 7 | 7 | 19 |
| Total | | | | 58 |

For decreasing computational cost, all images were reduced by a factor of two. Moreover, the $3 \times 3$ median filter was applied to reduce noise, and labels and pectoral muscle were manually extracted from the images with help of the ImageJ program [12]. With the purpose to filter and enhance the contrast of the possible mass regions, the Top-Hat transform was applied to all images. A disk was used as structural element to filter suspicious regions. The size of the disk was iteratively modified from two pixels to the width of breast area. Then, detection of suspicious regions (ROIs) was done by applying segmentation based on texture and Markov Random Fields. A Gaussian observation model with three texture features of first order: mean, standard deviation, and entropy, was used. In total, 278 ROIs of different sizes were identified, from which, 50 represent suspicious masses, while the other 228, normal tissue. These ROIs are the entry to the classification stage.

## 4 Features

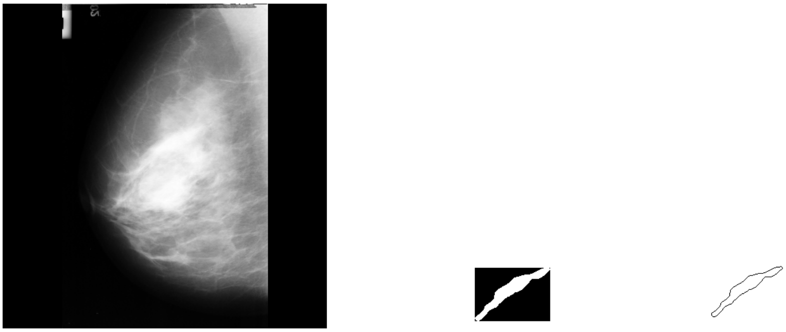
The following stage of mass detection by CADx systems is the feature extraction and selection. The feature space is very large and complex, but only some of features are significant. After years of intensive research, hundreds of features have been proposed. But using many features degrades the performance of the classifiers, so that redundant features should be removed to improve the performance of the classifier. There are basically three types of features: intensity, geometric and texture features. After reviewing many feature evaluation initiatives [8], [18], [20], we chose an important and discriminative subset of 35 features for mass detection.

### 4.1 Intensity Features

Three basic statistics of the detected ROIs were used: skewness, kurtosis and entropy.

### 4.2 Shape Features

Before the extraction of these features, the detected ROIs are binarized and processed to identify their boundaries. In Fig. 1 some examples of results for these processes are shown. Seven features were directly calculated from the pixels in the boundaries and within area of ROIs: perimeter, area, compactness, and the first four central invariant moments.



(a) Original MIAS image (b) Binarized ROI (c) ROI shape

**Fig. 1.** ROI processing for shape features extraction

### 4.3 Texture Features

Texture is the term used to characterize the surface of a given region, and it is one of the main features used in identifying ROIs in an image [3]. In general, texture features can be grouped into three classes based on what they are derived from: Gray-level co-occurrence matrices, Gray-level difference statistics, and Gray-level run length statistics.

**Gray-Level Co-occurrence Matrix (GLCM).** An element of the GLCM matrix $P(i, j, d, \theta)$ is defined as the joint probability that the gray levels $i$ and $j$ occur separated by a distance $d$ and along direction $\theta$ of the image [2]. Four GLCM matrices were calculated from each ROI using $\theta = \{0°, 45°, 90°, 135°\}$ and $d = 1$. From these matrices, the six following features were obtained (and averaged in the four directions): contrast, correlation, variance, energy, entropy and homogeneity.

**Gray-Level Difference Statistics (GLDS).** The GLDS vector is the histogram of the absolute difference of pixel pairs which are separated by a given displacement $\delta$ [19]. Also, to obtain GLDS features, four forms of the vector $\delta$ were considered: $(0, d), (-d, d), (d, 0)$, and $(-d, -d)$. Three textural features were measured and averaged (considering $d = 1$) from these vectors: mean, entropy and variance.

**Gray-Level Run Length Statistics (GRLS).** The GRLS method is based on computing the number of gray-level runs of various lengths [4]. A gray-level run is a set of consecutive and collinear pixel points having the same gray-level value. The length of the run is the number of pixels in the run. For an $M \times N$ run length matrix $p(i, j)$, $M$ is the number of gray levels and $N$ is the maximum run length. In a study [4], four feature extraction functions following the idea of joint statistical measure of gray level and run length gave better performance:

1. Short run low gray level emphasis (SRLGE)

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 \cdot j^2} \tag{1}$$

2. Short run high gray level emphasis (SRHGE)

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot i^2}{j^2} \tag{2}$$

3. Long run low gray level emphasis (LRLGE)

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot j^2}{i^2} \tag{3}$$

4. Long run high gray level emphasis (LRHGE)

$$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 \cdot j^2 \tag{4}$$

where $n_r$ is the total number of runs.

These four features were calculated in four positive directions: $0°, 45°, 90°$ and 135° (16 features) for Test 2. For Test 1 we add seven more features calculated in four directions (28 features): Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-Level Nonuniformity (GLN), Run Length Nonuniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE).

## 5 Classification and Experimental Results

### 5.1 Support Vector Machine (SVM)

SVM classifier [16] is a relative new option for doing classification. It has their roots in the existence of an optimal (in the sense of quadratic convex optimization) hyperplane that separates two classes. Data is projected by means of a kernel function in a high-dimensional space and, in this space, the hyperplane is linear, but their projection back in original space is non-linear. In our experiments we use a radial basis function (RBF) kernel. The fit of the hyperplane to data is controlled by the parameter $\beta$ of the RBF function and the SVM parameter $C$ that controls the width of the classifier's margin.

### 5.2 Experiments

For the experiments, the set of 278 detected ROIs was randomly divided in 25 masses and 114 normal tissue segments for training, and the equivalent for testing. In the first experiment (Test 1) we tested the 63 intensity, shape and texture features described in Section 4; the corresponding results are presented in Table 2. In other experiments we tested different subsets of texture features, and the best results were obtained with the first 35 features mentioned in Section 4.

**Table 2.** SVM classification results using a RBF kernel and the full set of 63 features

| Parameters | Accuracy in training set | Number of support vectors | Accuracy in test set |
|---|---|---|---|
| $\beta = 1, C = 2$ | 92.19 % | 58 | **84.18 %** |
| $\beta = 1, C = 10$ | 98.57 % | 62 | 79.86 % |
| $\beta = 2, C = 2$ | 76.98 % | 16 | 71.95 % |
| $\beta = 0.5, C = 2$ | 98.57 % | 98 | 84.18 % |
| $\beta = 1, C = 1$ | 50.4 % | 10 | 52.52 % |

SVM classifier gave the best results with $\beta = 1$ and $C = 2$; defining 54 support vectors (Table 3). Differences in the results represent the compromise between accuracy in test stage and number of support vectors. We tested different subsets of texture parameters and different kernels, but we are reporting here the best scores.

**Table 3.** SVM classification results using a RBF kernel and the best 35 features

| Parameters | Accuracy in training set | Number of support vectors | Accuracy in test set |
|---|---|---|---|
| $\beta = 1, C = 2$ | 88.49 % | 54 | **84.18 %** |
| $\beta = 1, C = 10$ | 94.25 % | 56 | 78.5 % |
| $\beta = 2, C = 2$ | 49.7 % | 8 | 48.3 % |
| $\beta = 0.5, C = 2$ | 94.25 % | 80 | 84.9 % |
| $\beta = 1, C = 1$ | 74.2 % | 12 | 71.3 % |

## 6 Conclusions and Future Work

We selected and tested some of the simplest and most discriminant features for digital processing of mammographies. After pre-processing and segmenting the ROIs of the MIAS database, SVM classification gives reasonably good accuracy scores with only 35 well known features.

With this framework we can test more shape and texture features, as well as classifiers and combinations among them. We are still far away of our purpuse, but with the future improvement of the different stages, we will be closer to build a working CADx system available for public service.

**Acknowledgments.** This research is supported by the mexican SEP-PROMEP program within the project “Detección, Segmentación e Identificación de Masas en Imágenes de Mamografía”. 103.5/12/4621, (IDCA: 10151 CLAVE: UTMIX-CA-32).

## References

1. Brandan, M., Villaseñor, N.: Detección del cáncer de mama: Estado de la mamografía en México. Cancerología: Revista del Instituto Nacional de Cancerología de México 1(3), 147–162 (2006)
2. Cheng, H., Shi, X., Min, R., Hu, L., Cai, X., Du, H.: Approaches for automated detection and classification of masses in mammograms. Pattern Recognition 20, 646–668 (2006)
3. Chang, T., Kuo, C.: Texture Analysis and classification with tree-structured wavelet transform. IEEE Transactions on Image Processing 2, 429–441 (1993)
4. Dasarathy, B., Holder, E.: Image characterizations based on joint gray-level run-length distributions. Pattern Recognition Letters 12, 497–502 (1991)
5. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, W.P.: The digital database for screening mammography. In: Proceedings of the Fifth International Workshop on Digital Mammography, pp. 212–218. Medical Physics Publishing (2001)
6. Horsch, A., Hapfelmeier, A.: Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies. International Journal of Computer Assisted Radiology and Surgery 6(6), 749–767 (2011)

7. Instituto Nacional de Estadística, Geografía e Informática (INEGI): Estadística a propósito del día mundial contra el cáncer. Boletín de Prensa (Febrero 2010), http://www.inegi.org.mx/inegi/contenidos/espanol/prensa/Contenidos/estadisticas/2010/cancer10.doc
8. Kim, J., Park, H.: Statistical Textual Features for Detection of Microcalcifications in Digitized Mammograms. IEEE Transactions on Medical Imaging 18(3), 231–238 (1999)
9. Lladó, X., Olivier, A.: A textural approach for mass false positive reduction in mammography. Computerized Medical Imaging and Graphics 33(6), 415–422 (2009)
10. Li, H., Liu, Y., Lo, S., Freedman, M.: Computerized Radiographic Mass Detection-Part II: Decision Support by Featured Database Visualization and Modular Neural Networks. IEEE Transactions on Medical Imaging 20(4), 302–313 (2001)
11. Oliver, A., Freixenet, J.: A review of automatic mass detection and segmentation in mammographic images. Medical Image Analysis 14(2), 87–110 (2010)
12. Rasband, W.: ImageJ. National Intitutes of Health, USA (1997), http://imagej.nih.gov/ij
13. Rojas, D., Nandi, K.: Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection. Computerized Medical Imaging and Graphics 32(4), 304–315 (2008)
14. Sampaio, B., Moraes, D.: Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Computers in Biology and Medicine 41(8), 653–664 (2011)
15. Samulski, M., Karssemeijer, N.: Optimizing case-based detection performance in a multiview CAD system for mammography. IEEE Transactions on Medical Imaging 30(4), 1001–1009 (2011)
16. Scholkopf, B., Smola, A.: Learning with Kernels. The MIT Press (2002)
17. Suckling, J.: The Mammographic Image Analysis Society Digital Mammogram Database. Exerpta Medica. International Congress Series 1069, 375–378 (1994)
18. Tang, X.: Texture Information in Run-Length Matrices. IEEE Transactions on Image Processing 7(11), 1602–1609 (1998)
19. Weszka, J., Dyer, C., Rosenfeld, A.: A comparative study of texture measures for terrain classification. IEEE Transactions Syst., Man, Cybern. SMC-6, 269–285 (1976)
20. Yin, F., Giger, M., Doi, K., Vyborny, C., Schmidt, R.: Computerized detection of masses in digital mammograms: investigation of feature-analysis techniques. J. Digital Imaging 7, 18–26 (1994)

# A Novel Right Ventricle Segmentation Approach from Local Spatio-temporal MRI Information

Angélica Maria Atehortúa Labrador, Fabio Martínez,
and Eduardo Romero Castro

CIM@LAB, Universidad Nacional de Colombia,
Ciudad Universitaria Cra 30 N 45-03 Facultad de Medicina Centro de telemedicina,
Bogotá, Colombia
{amatehortual,fmartinezc,edromero}@unal.edu.co
http://cimlaboratory.com/

**Abstract.** This paper presents a novel method that follows the right ventricle (RV) shape during a whole cardiac cycle in magnetic resonance sequences (MRC). The proposed approach obtains an initial coarse segmentation by a bidirectional per pixel motion descriptor. Then a refined segmentation is obtained by fusing the previous segmentation with geometrical observations at each frame. A main advantage of the proposed approach is a robust MRI heart characterization without any prior information. The proposed approach achieves a Dice Score of 0.62 evaluated over 32 patients.

**Keywords:** Right Ventricle Segmentation, Cardiac MRI Cine, Local Motion Models, Structural Information.

## 1 Introduction

Cardiovascular diseases (CVDs) are world wide one of the principal causes of death and disability [1]. An accurate quantification of the right ventricular structure and function has become important to support the diagnosis, prognosis and evaluation of several cardiac diseases and also to complement the typical analysis of the left ventricular function [2,3]. Currently, most common methods, for assessing the heart, are based on quantification and characterization of patterns during a Cardiac Magnetic Resonance Imaging (CMRI) [4]. Such methods are widely used to analyze, diagnose and even prognose certain heart diseases. Among the evaluated heart variables, the most common are the ventricular chamber sizes, volumes and masses at each cardiac phase, ventricular function and correlation flow [5]. A proper RV analysis demands an accurate 3D temporal segmentation, specifically endocardial and epicardial contours. Typically such task is carried out by expert cardiologists who perform manual delineations that may take $18.9 \pm 4\ min$ [6] per case, introducing high inter-and-intra observer variability [3,7]. Therefore, automatic segmentation methods are appealing to obtain more accurate RV temporal-segmentation. Nevertheless, several challenges arise because of the complex RV geometry shape and high non-linear heart motion

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 206–213, 2013.

during diastole and systole transition. In addition, RV fuzzy edges and random acquisition noise make more challenging the RV segmentation [6].

Several state-of-the-art methods have been proposed for automatic RV segmentation, most of them based on the use of strong structural and appearance priors that adjust the shape w.r.t a set of samples. In this sense, these methods use mainly statistical shape models, multi-atlas strategies and deformable approximations [6]. These strategies are strongly dependent on how data is learned to build up the prior. However, accurate quantification of certain variables like the ejection fraction depends on the shape changes, particularly important in pathological cases. In addition, such approaches pay a high price when mapping the prior to the MR since the metrics is usually noisy because of the dependency of intensity variations or the pixel spatial distribution to represent the heart, facts that may lead to inconsistent segmentations [6].

On the other hand, methods with no prior are based on appearance and temporal MRI information. Cocosco et al. [8] describe the segmentation of both the left ventricle (LV) and right ventricle (RV), by a simple temporal RoI estimation of major motions and then a voxel classification is performed between RV and LV using morphological operations. However, the simplicity of the temporal descriptor, a simple subtraction between consecutive frames, turns out to be very noisy. In addition, Wang et al. [9] capture information that is shared during the sequence and merge it with a spatial within-frame descriptor, based on a classical isodata algorithm. Nonetheless, RV segmentation may easily overflow the actual borders.

The main contribution of this work is a fully automatic method that uses no prior at all and delineates the RV endocardium contour in 4D MR sequences. The strategy uses both a heart motion descriptor and an estimation of RV shape for each frame of the sequence. Firstly a robust per pixel motion model is introduced to highlight the edges with major changes along the sequence, under the hypothesis that heart is the organ with larger motion. Afterwards, a conventional isodata strategy estimates the heart shape which is superimposed to the edges computed from the motion estimation. The final delineation is set to the intersection between those edges and estimated heart shape. The following section describes the proposed segmentation approach. In section 3, the evaluation and results. Finally in section 4 is presented the discussion on the results obtained and some conclusions.

## 2 Methodology

The strategy herein proposed is capable of capturing the temporal RV contours from a spatio-temporal MRI characterization. As widely acknowledged, heart motion is the main biomarker in cardiology, allowing by itself an appropriate assessment of cardiac function [10]. Hence, the approach starts by coding temporal MRI information with a bidirectional per-pixel motion descriptor [11]. A coarse heart segmentation is initially obtained from that estimated cardiac motion. This segmentation is corrected using geometrical observations from the

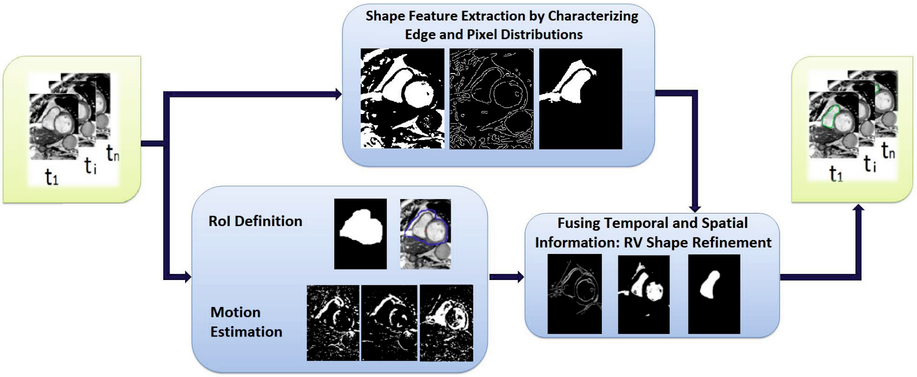

Shape Feature Extraction by Characterizing Edge and Pixel Distributions
t1
ti
tn
RoI Definition
Motion Estimation
Fusing Temporal and Spatial Information: RV Shape Refinement
t1
ti
tn

**Fig. 1.** The proposed method. A motion descriptor is computed for the whole MRI cardiac cycle, which is then adjusted to the edged and spatial estimation found at each frame in the estimated shape.

estimated shape. The pipeline of the proposed approach is illustrated in Figure 1 and described in the following subsections.

## 2.1 Motion Estimation

The heart is the organ whose vital function amounts to a constant motion. The proposed strategy starts by estimating the cardiac movement with a bidirectional local motion descriptor. For doing so, a temporal median sets the elements with less motion during the sequence by recursively updating the frame median, as follows $M_t(x) = M_{t-1}(x) + sgn\left(I_t(x) - M_{t-1}(x)\right)$, where $M_t(x)$ represent the median and $I_t(x)$ the frame at time $t$ for each pixel $x$. Using such recursive median, a likelihood measure $\varDelta_t$ sets those pixels in movement at each instant $t$ as $\varDelta_t = |M_t(x) - I_t(x)|$. This last term is in due turn regularized by a cumulated variance of the motion history, as: $V_t(x) = V_{t-1}(x) + sgn(N \times \varDelta_t(x) - V_{t-1}(x))$. This descriptor is highly noise robust and computes the per-pixel temporal variation that allows to classify the RV. Specifically, At the End of the Diastole, when the heart is maximally expanded, pixel candidates should meet two conditions: the pixel motion is larger than an accumulated temporal variance under the restriction that the movement must span an important percentage of the cardiac cycle. Such relationship is represented by a simple thresholding as $\widehat{BS_t}^{(D)}(x) = \varDelta_t(x) \geq V_t(x)$. In contrast, at the Systole, the heart contraction is maximum and the motion is practically null so that this phase constructs a very steady history of the cardiac flow. After the semilunar valves open, blood flows out the ventricle with an important change that is very likely detected by the motion estimation algorithm. The heart edges are thus calculated from pixels with major motion by comparing the likelihood measure with a learned scalar parameter $\tau$ as: $\widehat{BS_t}^{(S)}(x) = \varDelta_t(x) \geq \tau$.

Classically, local motion descriptors [11] are usually unidirectional recursive algorithms, but in this case the first iteration yielded a very blurry estimation of the heart contour at the End-of-Diastole. As the recursive accuracy depends on the captured motion history, the descriptor is herein bidirectionally run, i.e., forward and backward as $BS_t(x)^{(D,S)} = \alpha\widehat{BS_t}^{(D,S)}(x) + (1-\alpha)\widehat{BS_{N-t}}^{(D,S)}(x)$, where $\alpha$ is an important parameter defined as $\dfrac{t}{N}$ and $N$ is the number of frames. Once motion has been thresholded, morphological operators groups up pixels associated with movement [11].

### 2.2 Shape Feature Extraction by Characterizing Edge and Pixel Distributions

The previous motion estimation produces a coarse shape segmentation and serves also to define a Region of Interest (RoI). The aim of this second phase is to construct another complementary shape approximation, using exclusively spatial observations. A first approximation to such heart shape consisted in finding a RoI that consistently surrounded the heart, as the spatial region with larger temporal motion at each time step. Within such RoI, heart ventricles are estimated from two complementary measurements: a global shape clustering and an edge extraction.

Firstly a global shape description of the ventricles was herein obtained by a classical isodata algorithm [12] that is used to separate the intensities corresponding to the myocardium and the cardiac chambers. The cardiac wall or myocardial tissue is segmented and therefore the right and left heart chambers serve as a reference frame of the right and left ventricles.

On the other hand, ventricle edges are estimated from the MRI RoI by using a conventional Canny filter [13]. In the apical slices however, while the LV is still differentiable, RV edges are very blurred (as shows Figure 2). Overall, edges in apical slices are considered as part of the LV. Estimations of RV edges are performed from the previous motion estimation provided that such edge is not already part of the previously defined LV edges.

### 2.3 Fusing Temporal and Spatial Information: RV Shape Refinement

During certain phases of the cardiac cycle, some boundaries of the heart were not properly segmented. Two fusion strategies were herein implemented to cope with such issue: 1) a first approach fused the spatial information obtained from the temporal information with the edge estimation and the isodata algorithm 2) a second strategy fused the temporal and isodata informations, but using exclusively the left ventricle isodata information. This second approach was particularly useful to segment the right ventricle at the apex level. For the first fusion strategy, the two edge representations are fused by simply summing and normalizing. The final shape is in this case outlined by intersecting both the RV

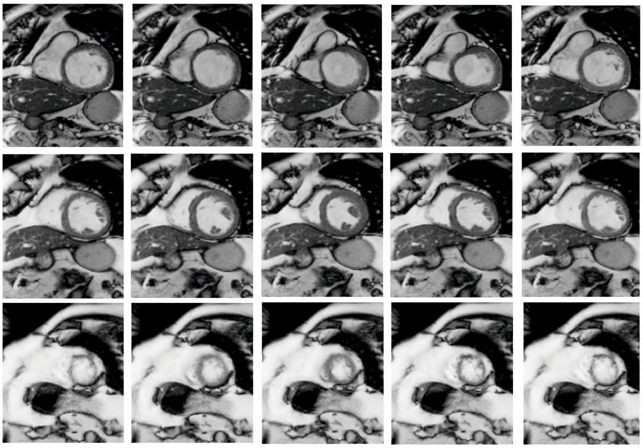


**Fig. 2.** The variability of the RV shape, from basal (top row) to apical (bottom row), and from left to right for the whole cardiac cycle, being the first column the End-of-Diastole and the mid column the End-of-Systole

shape estimated from the isodata information and edges. For the second strategy, it was applied a simple difference between the temporal heart segmentation and the spatial LV segmentation obtained by the isodata algorithm so that the remaining pixels then correspond to the RV. Finally, isolated pixels are always removed by simple opening and closing operators.

### 2.4 Data

The evaluation of the proposed approach was performed over a public Cardiac MRI dataset [3,14] with 32 patients split into two subsets: training and test data set, which are specified by the authors of the dataset. For evaluation, the obtained segmentation was submitted to the RVSC [15] which sends back the results. Training data consisted in a set of 16 cardiac MRI, half split into men and women, with an average age of 51 ± 12 years. For test data was split into 3 women and 13 men cases, respectively, with and average age of 48 ± 18 years. The recorded patients were diagnosed with several cardiac pathologies like myocarditis, ischaemic cardiomyopathy, arrhythmogenic right ventricular dysplasia (ARVD), dilated cardiomyopathy, hypertrophic cardiomyopathy, Aortic stenosis, cardiac tumour, left ventricular and ejection fraction assessment. Each MR sequence was captured in the short-axis with 1.5 Tesla, in a plane resolution of 1.3 mm and a between-slice distance of 8.4 mm. The epicardium and endocardium of 32 MR sequences were manually delineated by an expert cardiologist. Trabeculae and papillary muscles were included in the RV cavity.

## 3 Evaluation and Results

Figure 3 illustrates the good performance of the method in cardiac MRI sequences. The green contour corresponds to the result obtained by the presented

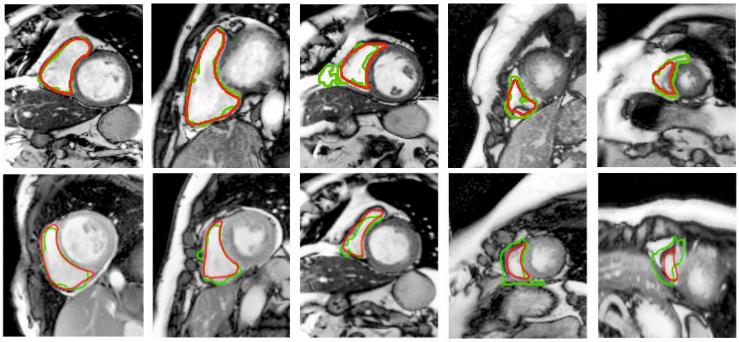


**Fig. 3.** Example of RV segmentations with several cases, including the End-of-Diastole (firts row) and End-of-Systole (second row). The ground truth is the red line and the green line is the automatic segmentation. As expected, better results were observed at the basal slices (first column).

method, while the ground truth is drawn in red. As expected, failures are mainly present in apical images because of the fuzzy borders and small RV area.

Quantitative technical evaluation was performed using the most classical metrics described in the literature: Dice Score (DSC) measure[16] and Hausdorff distance (HD)[17]. An overlap DSC measure is defined as: $DSC(A,B) = \dfrac{2(A \cap B)}{A+B}$, where $A$ and $B$ represent the obtained area and the expert ground truth, respectively. On the other hand, the Hausdorff measure $H(A,B)$ computes the maximum distance between two sets of points as $\max(H(A,B), H(B,A))$ and $H(A,B) = \max\limits_{a \in A} \min\limits_{b \in B} \|a-b\|_2^2$. In this case, each set of points represents the organ surface. This measure allows to indirectly assess the compactness of the segmentation. A clinical performance was also assessed as the ejection fraction (EF).

**Table 1.** Performance of the proposed approach for training data using Dice Score (DSC) and Hausdorff distance (HD) over the Endocardium contour

| | **DSC** | **HD (in mm)** |
|---|---|---|
| | mean (std) | mean (std) |
| **End-of-Diastole (ED)** | 0.66 (0.22) | 20.66 (13.00) |
| **End-of-Systole (ES)** | 0.54 (0.26) | 27.72 (23.45) |

Quantitative results were only evaluated at End-of-Diastole (ED) and End-of-Systole (ES) since these two states are the most important to clinical quantification [18]. As baseline it was taken the work proposed by Wan et. al [9], which until now represents the best strategy to segment the RV without prior. Table 1 summarizes the obtained performance for training data sequences in ED and ES times. The proposed approach clearly outperforms the baseline method in terms of overlapping and compactness in both cardiac states. As expected a much better segmentation is obtained at the ED because the MRI frame quality allows a better quantification. Although, at the ES many times the poor MRI contrast leads to a quite fuzzy RV edges, the proposed approach outperforms the state-of-the-art approach. Table 2 summarizes the performance obtained by our approach over the test data. Although the obtained score errors are slightly larger for the RV segmentations, the proposed approach properly captures the shape variability and is easily adapted to new RV shapes since it only depends on the particular MRI observations.

**Table 2.** Performance of the our RV segmentation method for the Test data set using Dice Score (DM) and Hausdorff distance (HD) over the Endocardium contour

| | **Our approach** | | **Baseline** | |
|---|---|---|---|---|
| | **DSC**<br>mean (std) | **HD (in mm)**<br>mean (std) | **DSC**<br>mean (std) | **HD (in mm)**<br>mean (std) |
| **ED** | 0.72 (0.29) | 16.17 (16.48) | 0.63 (0.32) | 22.89 (25.01) |
| **ES** | 0.51 (0.31) | 27.47 (27.96) | 0.50 (0.34) | 27.99 (24.97) |

Finally, it was calculated the mean error for the ejection fraction, defined as $error = \sum_{p=1}^{N} EFp_{auto} - EFp_{manual}$, where an error of 0.36 was obtained over the whole data set (32 patients). Although the error index shows an acceptable performance, some important noise sources, such as the inter-and-intra high variability of RV shape, the fuzzy edges and the complex heart motion, are not properly captured by our method. Nevertheless, the approach herein presented shows appropiate RV segmentations using an strategy based principally in temporal characterization. This approach outperform state-of-the-art methods that use only appearance and temporal observations for each sequence [8,9].

## 4 Conclusions

In this paper it was introduced a new strategy to segment the right ventricle in MR sequences. The proposed mixed approach uses spatio-temporal observations and produces reliable RV segmentations. A great advantage of the proposed approach is its independency of any prior heart shape, facilitating the capture of dynamic and shape heart variability, which could be associated to specific cardiac pathology. In future work, the method could extend to 3D processing and further validation with a larger data set will be performed.

## References

1. Roger, V., Go, A., Lloyd-Jones, D., Adams, R., Berry I, J.: Heart disease and stroke statisticts - 2012 update: A report from the american heart association. Circulation 125, 200–220 (2012)

2. Haddad, F., Doyle, R., Murphy, D., Hunt, S.: Right ventricular function in cardiovascular disease, part ii: pathophysiology, clinical importance, and management of right ventricular failure. Circulation 117, 1717–1731 (2008)
3. Caudron, J., Fares, J., Lefebvre, V., Vivier, V., Petitjean, C., Dacher, J.: Cardiac mri assessment of right ventricular function in acquired heart disease: factors of variability. Acad. Radiol. 19, 991–1002 (2012)
4. Baur, L.: Magnetic resonance imaging: the preferred imaging method for evaluation of the right ventricle. The International Journal of Cardiovascular Imaging 24, 699–700 (2008)
5. Haddad, F., Hunt, S., Rosenthal, D., Murphy, D.: Right ventricular function in cardiovascular disease, part i: Anatomy, physiology, aging, and functional assessment of the right ventricle. Circulation 117, 1436–1448 (2008)
6. Petitjean, C., Dacher, J.: A review of segmentation methods in short axis cardiac mr images. Med. Image Anal. 15, 169–184 (2011)
7. Caudron, J., Fares, J., Vivier, P., Lefebvre, V., Petitjean, C., Dacher, J.: Diagnostic accuracy and variability of three semi-quantitative methods for assessing right ventricular systolic function from cardiac mri in patients with acquired heart disease. Eur. Radiol., 2111–2120 (2011)
8. Cocosco, C., Niessen, W., Netsch, T., Vonken, E., Lund, G., Stork, A., Viergever, M.: Automatic image-driven segmentation of the ventricles in cardiac cine MRI. J. Magn. Reson. Imaging 28, 366–374 (2008)
9. Wang, C., Peng, C., Chen, H.: A simple and full automatic right ventricle segmentation method for 4-dimensional cardiac MR images. In: RV Segmentation Challenge in Cardiac MRI MICCAI 2012 (2012)
10. Punithakumar, K., Ayed, I.B., Islam, A., Goela, A., Ross, I.G., Chong, J., Li, S.: Regional heart motion abnormality detection: An information theoretic approach. Medical Image Analysis 17(3), 311–324 (2013)
11. Manzanera, A., Richefeu, J.C.: A new motion detection algorithm based on $\sigma$-$\delta$ background estimation. Pattern Recogn. Lett. 28(3), 320–328 (2007)
12. Ridler, T., Calvard, S.: Picture thresholding using an iterative selection method. IEEE Transactions on Systems, Man and Cybernetics 8(8), 630–632 (1978)
13. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8(6), 679–698 (1986)
14. Download MRI Data from RV segmentation Challenge MICCAI 2012 (2012), `http://www.litislab.eu/rvsc/rv-segmentation-challenge-in-cardiac-mri/download`
15. Evaluation Framework Right Ventricle Segmentation (2012), `http://www.litislab.eu/rvsc`
16. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
17. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9), 850–863 (1993)
18. Goetschalckx, K., Rademakers, F., Bogaert, J.: Right ventricular function by MRI. Curr. Opin. Cardiol., 451–455 (2010)

# Advances in Texture Analysis for Emphysema Classification

Rodrigo Nava[1], J. Victor Marcos[2], Boris Escalante-Ramírez[1],
Gabriel Cristóbal[2], Laurent U. Perrinet[3], and Raúl San José Estépar[4]

[1] Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México, Mexico City, Mexico

[2] Instituto de Óptica, Spanish National Research Council (CSIC), Serrano 121, Madrid 28006, Spain

[3] INCM, UMR6193, CNRS & Aix-Marseille University, 31 ch. Aiguier, 13402 Marseille Cedex 20, France

[4] Brigham and Women's Hospital, Harvard Medical School, Boston MA, United States

**Abstract.** In recent years, with the advent of High-resolution Computed Tomography (HRCT), there has been an increased interest for diagnosing Chronic Obstructive Pulmonary Disease (COPD), which is commonly presented as emphysema. Since low-attenuation areas in HRCT images describe different emphysema patterns, the discrimination problem should focus on the characterization of both local intensities and global spatial variations. We propose a novel texture-based classification framework using complex Gabor filters and local binary patterns. We also analyzed a set of global and local texture descriptors to characterize emphysema morphology. The results have shown the effectiveness of our proposal and that the combination of descriptors provides robust features that lead to an improvement in the classification rate.

**Keywords:** Co-occurrence matrices, Emphysema, Gabor filters, LBP, Sparsity, Tchebichef, Texture analysis.

## 1 Introduction

COPD is a progressive and irreversible lung condition, which is characterized by tissue damage. It hinders air from passing through airpaths and causes that alveolar sacs lose their elastic quality, increasing the risk of death. COPD can manifest as either emphysema, bronchitis or both; the former is the most common manifestation that destroys lung parenchyma [1].

Literature recognizes three types of emphysema: *i*) **Paraseptal** (PS) or distal acinar emphysema, which is characterized by destruction of distal airway structures, alveolar ducts, and alveolar sacs. The process is localized around the pleura; *ii*) **Panlobular** (PL) or panacinar emphysema destroys uniformly the entire alveolus, it is predominant in the lower half of the lungs; and *iii*) **Centrilobular** (CL) or centriacinar emphysema, which is the most common type

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 214–221, 2013.

of emphysema. It begins in the respiratory bronchioli and spreads peripherally, most damage is usually contained to the upper half of the lungs.

Spirometry is the gold standard criterion to establish a diagnosis of emphysema. It measures the volume of air that a patient is able to expel from lungs after a maximal inspiration. Nevertheless, this method does not allow to discriminate pathological subphenotypes of emphysema. On the other hand, HRCT is a minimally invasive imaging technique capable of providing both high-contrast and high-resolution details of lungs and airways; it has shown its potential for identifying changes in lung parenchyma and abnormalities associated with emphysema.

Hayhurst et al. [2] showed that Hounsfield Unit (HU) frequency distributions in patients who had CL differed significantly from patients with Normal Tissue (NT). Low-attenuation areas in HRCT images have been found to represent macroscopic and microscopic changes due to emphysema. Such areas are determined using the density mask method, which measures the amount of emphysematous lung by calculating the percentage of voxels lesser than a threshold; commonly, the threshold lies somewhere between -910 and -980 HU.

Texture-based classification of lung HRCT images may provide new insights towards the construction of a reliable computer-aided diagnosis system. New methods include features extracted using local binary patterns [3]. A simpler alternative based on kernel density estimation of local histograms has been proposed in [4]. A different approach was proposed in [5] where the authors used meta-data to label lung samples, whereas in [6], the Riesz transform was used to obtain textural features in interstitial lung abnormalities but it has not been tested in analysis of emphysema subtypes. However, researchers have analyzed texture in HRCT images using simple descriptors. In this work, we claim that the combination of both global and local descriptors will provide robust features because global characteristics and local information are encode simultaneously. Thus, an improvement in the classification rate can be attained.

This paper is organized as follows: the datasets are described in Section 2. In Section 3 we defined a set of global and local descriptors used in the present study and provided their mathematical foundations. The results are presented in Section 4. Finally, our work is summarized in Section 5.

## 2 Material

We used two datasets labeled by experienced pulmonologists: the **Bruijne and Sørensen dataset** (BS) was provided by Prof. Dr. Bruijne and Dr. Sørensen [3]. It consists of 168 non-overlapping annotated ROIs of size $61 \times 61$ pixels and belong to three types of patterns: NT=59, CL=50, and PS=59; and **Brigham and Women's Hospital dataset** (BWH). This dataset was provided by researchers from the Brigham and Women's Hospital using a subset of the COPDGene study [4]. 1337 ROIs from 267 CT scans were randomly selected; the distribution per pattern is: NT=370, PS=184, PL=148. BWH includes three subtypes of CL patterns (mild, moderate, and severe): CL1=170, CL2=287, and CL3=178 respectively. The size of the samples was chosen to fit the physical extent of emphysema within the secondary lobule corresponding to 31 × 31 pixels.

## 3 Methods

We propose the combination of Complex Gabor Filters (CGF) and Local Binary Patters (LBP) for a better characterization of emphysema; the former are global descriptors, whereas the latter are local descriptors. Additionally, a wide set of texture descriptors have been analyzed. To assign a given patch to one of several emphysema patterns, we used a methodology composed of three stages: *i*) feature extraction with global and local descriptors; *ii*) dimensionality reduction using Kernel-Fisher Discriminant Analysis (KFDA); and *iii*) classification with $k$-Nearest Neighbors (kNN). In the following paragraphs we summarize the main characteristics of the descriptors used in the current study.

**Complex Gabor Filters** [7] are defined as the product of Gaussian functions and complex sinusoids. They are band-pass filters that constitute a complete but non-orthogonal basis set and their shape match with psychophysical properties of receptive fields [8]. They can be divided into two parts: $g_e\left(x,y\right) = K\exp\{-\frac{1}{2}(\frac{\tilde{x}^2+\gamma^2\tilde{y}^2}{\alpha^2})\}\cos\left(2\pi u_0\tilde{x}\right)$, which is an even filter, whereas $g_o\left(x,y\right) = K\exp\{-\frac{1}{2}(\frac{\tilde{x}^2+\gamma^2\tilde{y}^2}{\alpha^2})\}\sin\left(2\pi u_0\tilde{x}\right)$ is an odd filter. $K$ represents a normalizing constant, $u_0$ is the central frequency, $(\alpha,\gamma)$ are the constants of the Gaussian envelope along $x$ and $y$-axes respectively. $\tilde{x} = x\cos\theta - y\sin\theta$, $\tilde{y} = x\sin\theta + y\cos\theta$, and $\theta$ denotes the orientation. Further filtering parameters were tuned by following the design constraints recommended in [9].

We used a bank made of 24 filters distributed in 4 scales (s) and 6 orientations; for each of them, we computed $E_{(s,\theta)} = I(x,y) \star g_{e(s,\theta)}(x,y)$ and $O_{(s,\theta)} = I(x,y) \star g_{o(s,\theta)}(x,y)$ where $I(x,y)$ is the given patch and the $\star$ indicates convolution. Then, we extracted the magnitude coefficients $(M_{(s,\theta)}(x,y))$ as:

$$M_{(s,\theta)}(x,y) = \sqrt{E^2_{(s,\theta)}(x,y) + O^2_{(s,\theta)}(x,y)} \tag{1}$$

Since $M_{(s,\theta)}(x,y)$ are considered as random variables, we extracted the mean $(\mu)$, the standard deviation $(\sigma)$, the skewness $(\varUpsilon)$, and the kurtosis $(\varPsi)$ from them to characterize the response of any image and build a feature vector, $\overline{f_{CGF}}$, as follows:

$$\begin{aligned}\overline{f_{CGF}} = &\big[\mu_{(0,0)},\, \sigma_{(0,0)},\, \varUpsilon_{(0,0)},\, \varPsi_{(0,0)},\, \ldots, \\ &\mu_{(s-1,\theta-1)},\, \sigma_{(s-1,\theta-1)},\, \varUpsilon_{(s-1,\theta-1)},\, \varPsi_{(s-1,\theta-1)}\big]\end{aligned} \tag{2}$$

**Log-Gabor Filters** (LGF) [10] are defined in frequency domain as Gaussian functions shifted from the origin; they have a null DC component and can be split into radial and angular filters: $\hat{G}(\rho,\theta) = \exp\{-\frac{1}{2}[\frac{\log(\frac{\rho}{u_0})}{\log(\frac{\alpha_\rho}{u_0})}]^2\}\exp\{-\frac{1}{2}[\frac{(\theta-\theta_0)}{\alpha_\theta}]^2\}$, where $(\rho,\theta)$ represent the polar coordinates, $u_0$ is the central frequency, $\theta_0$ is the orientation angle, $\alpha_\rho$ and $\alpha_\theta$ determine the scale and the angular bandwidth respectively. We applied setting recommendations that appear in [9] and computed the feature vectors, $\overline{f_{LGF}}$, by convolving a bank of 24 log-Gabor filters distributed in 4 scales and 6 orientations with the input images and then we followed the procedure presented in Eq. (2).

**Sparse Gabor Coding** (SGC). Gabor filters provide redundant representations, which may hamper classification tasks. As proposed first by [11], this problem may be solved using a greedy algorithm. This approach corresponds to first choosing a single filter, $\Phi_i$, that best fits the image, $I(x, y)$, along with a suitable coefficient $a_i$, such that the *single* source $a_i\Phi_i$ is a good match to the image: $i = \arg\max_j(\langle \frac{I(x,y)}{\|I(x,y)\|}, \frac{\Phi_j}{\|\Phi_j\|}\rangle)$, where $\langle\cdot,\cdot\rangle$ represents the inner product.

The associated coefficient is the scalar projection: $a_i = \langle I(x, y), \frac{\Phi_i}{\|\Phi_i\|^2}\rangle$. Knowing this choice, the image can be decomposed as: $I(x, y) = a_i\Phi_i + \mathbf{R}$ where $\mathbf{R}$ is the residual image. We repeat this 2-step process on the residual until some stopping criterion is met. This procedure is known as the matching pursuit algorithm, which has proven to be a good approximation for natural images [12]. Measuring the ratio of extracted energy in the images, 256 edges were on average enough to extract 90% of the energy of whitened images on all datasets. We thus used this set of sparse coefficients as the input vector for the classification framework.

**Gray-level Co-occurrence Matrices** (GLCM) were proposed by Haralick [13]. They evaluate spatial relationship among gray levels. Each pixel in an image $I(x, y)$ is assigned to one of $N_g$ gray levels. The GLCM matrix consist of a set of $\{P_{ij}|i, j = 1, \ldots, N_g\}$ values. $P_{ij}$ represents the number of occurrences of two pixels with gray levels $i$ and $j$ separated by a distance $d$ in the direction of the angle $\theta$. The GLCM's elements are normalized, providing the relative frequency of occurrence for a pair of gray levels.

The element $p\,(i, j)$ denotes the probability of finding the pair of levels $(i, j)$ in the image, which is obtained as: $p(i, j) = P_{ij}(\sum_{i,j}^{N_g} P_{ij})^{-1}$. 10 features were chosen to capture texture properties: energy, contrast, correlation, homogeneity, entropy, autocorrelation, dissimilarity, cluster shade, cluster prominence, and maximum probability. In our study, $N_g$ was set to 8 according previous works focused on texture analysis [14]. $d$ was set to 1 while four different angle values were assessed: $0, 45, 90$, and 135 degrees. Thus, a total of 40 descriptors (10 statistical features for each of the four orientations) were obtained for each texture.

**Discrete Tchebichef Moments** (DTM) [15] are computed by projecting the image $I(x, y)$ onto the set of Tchebichef polynomial kernels. DTM provides a unique representation of the image in the spanned Tchebichef space. The moment $T_{pq}\,(p, q = 0,\, 1,\, \ldots,\, N - 1)$ of order $s = p + q$ is defined as:

$$T_{pq} = \frac{1}{\tilde{\rho}(p, N)\tilde{\rho}(p, N)} \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} \tilde{t}_p(x)\tilde{t}_q(y)I(x, y) \tag{3}$$

where $\tilde{t}_p(x)$ and $\tilde{t}_q(x)$ are scaled Tchebichef polynomials and $\rho(n, N)$ is its squared norm. $T_{pq}$ quantifies the correlation between the image, $I(x, y)$, and the kernel $\tilde{t}_p(x)\tilde{t}_q(y)$. Hence, this magnitude will be higher for images characterized by repetitive patterns occurring at a similar rate to the kernel. The following feature evaluates the similarity between the image and the varying patterns implemented by $s$-order Tchebichef kernels: $T(s) = \sum_{p+q=s} |T_{pq}|$, $(s = 0, 1, \ldots, 2N-2)$. The analysis based on DTM yields a feature vector of length $2N-1$ to describe texture attributes.

**Local Binary Patterns** [16] are based on the idea that textural properties within homogeneous regions can be mapped into patterns, which represent micro-features. LBP uses a $3 \times 3$ square mask called "texture spectrum". The values in the square mask are compared with the central pixel, those ones lesser are labeled with "0" otherwise they are labeled with "1". The labeled pixels are multiplied by a fixed weighting function according with their positions to form a chain. Afterward, the values of the eight pixels are summed to obtain a label: $LBP_{P,R}(g_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p$, where $\{g_p | p = 0, \ldots, P-1\}$ are the values of the neighbors. The comparison function $s(x)$ is defined as a Heaviside function: $s(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$

**Uniform Local Binary Patterns** ($LBP_{P,R}^{uni}$) [17]. Over 90% of LBP patterns can be described with few spatial transitions, which are the changes (0/1) in a pattern chain. Ojala introduced the measure $U(LBP_{P,R}(g_c)) = |s(g_{p-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)|$, which corresponds to the number of spatial transitions. So that, the uniform LBP ($LBP_{P,R}^{uni}$) can be obtained as:

$$LBP_{P,R}^{uni}\left(g_c\right) = \begin{cases} \sum_{p=0}^{P-1} s\left(g_p - g_c\right) \text{ if } U\left(LBP_{P,R}\left(g_c\right)\right) \leq 2 \\ P+1 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{4}$$

after the process is completed; a labeled image, $L\left(x,y\right)$, is generated and the pixel-wise information is encoded as a histogram, $H_i$.

### 3.1 Multi-class Kernel Fisher Discriminant Analysis

It must be considered that the size of a training set should be exponentially increased with the dimensionality of the input space. Since the previous methods generate high-dimensional feature vectors and a limited dataset is available, we used KFDA [18], which maps original data into a new feature space preventing overfitting. Let $\boldsymbol{X}^1 = \{x_1^1, x_2^1, \ldots, x_{l_1}^1\}, \ldots, \boldsymbol{X}^C = \{x_1^C, x_2^C, \ldots, x_{l_C}^C\}$ be feature vectors from $C$ classes and let $\boldsymbol{K}(m,n)$ be the kernel matrix defined as $\boldsymbol{K}(m,n) = k(X_m, X_n)$ where $\boldsymbol{X} = \bigcup_{i=1}^{C} \boldsymbol{X}^i$. We used the Gaussian kernel, $k(x,y) = \exp\{-\frac{1}{2}\frac{\|x-y\|^2}{a^2}\}$, $a = 333$.

The "between scatter matrix" is defined by $\boldsymbol{P} = \sum_{j=1}^{C} l_j(\mu_j - \mu)(\mu_j - \mu)^T$ with $\mu_j = \frac{1}{l_j}\sum_{\forall n \in X^j} \boldsymbol{K}(m,n)$ and $\mu = \frac{1}{l}\sum_{\forall n} \boldsymbol{K}(m,n)$. The "within class scatter matrix" is defined by $\boldsymbol{Q} = \boldsymbol{K}\boldsymbol{K}^T - \sum_{j=1}^{C} l_j \mu_j \mu_j^T$; since $\boldsymbol{Q}$ must be a positive definite matrix, we used $\boldsymbol{Q} = \boldsymbol{Q} + r\boldsymbol{I}$ to guarantee that $\boldsymbol{Q}$ is positive definite. Finally, $\alpha^*$ is built with the $C-1$ largest eigenvalues of $\boldsymbol{Q}^{-1}\boldsymbol{P}$ and the projection can be computed as: $y = \boldsymbol{K}\alpha^*$.

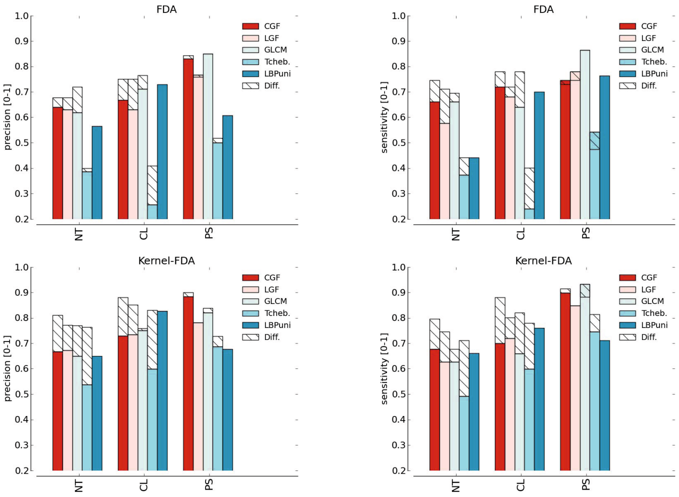


**Fig. 1.** BS classification rates (Three classes). The first row shows the results using FDA whereas in the second row, the results using KFDA are shown. Note that in almost all the cases, the extended methods ($Diff$), which are built by concatenating a single descriptor and its corresponding $LBP^{uni}_{P,R}$ histogram, achieved higher rates.

## 4 Experiments and Results

Parameter selection is a fundamental step in any classification problem; we used 10-fold cross-validation to estimate global parameters resulting in $k = 20$ neighbors in the kNN classifier as the best case. Then, we applied leave-one-out cross-validation to measure Specificity (Sp), Sensitivity (S), and Precision (P). We carried out a comparison of each method using both Fisher Discriminant Analysis (FDA) and KFDA, (see Fig. 1 for BS and Fig. 2 for BWH). Since KFDA generates non-linear boundaries among classes, the classification rates are better than those achieved with FDA. Furthermore, we compared each method with its extended version, which is built by concatenating a single descriptor and its corresponding $LBP^{uni}_{P,R}$ histogram into a single sequence to represent a mixture descriptor.

We computed the $F_1$-Score $= 2 * \frac{P*S}{P+S}$ for each algorithm and measured the accuracy of the tests. For BS dataset, our proposal, CGF + $LBP^{uni}_{P,R}$, achieved the highest $F_1$-Score with 0.8637. A straightforward comparison with the work of Bruijne and Sørensen [3] is not possible because they reported a classification rate using patches of $31 \times 31$ pixels as the best case. Here, we used larger patches, which implies the risk of including different lobes that might have different emphysema and might decrease the classification performance. Using the BWH dataset, our proposal also achieved the highest $F_1$-Score with 0.6899. Mendoza et al. [4] reported a $F_1$-Score of 0.6440 using the kernel density estimation approach.

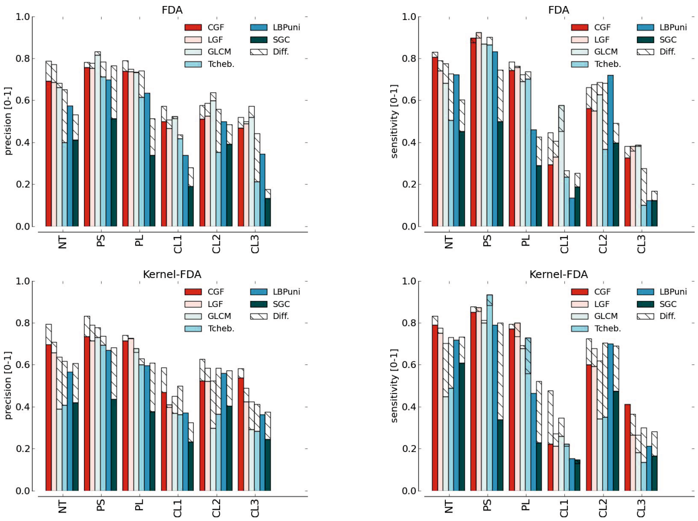


**Fig. 2.** BWH classification rates (Six classes). The first row shows the results using FDA while in the second row the classification rates using KFDA are shown. The extended methods, $Diff$, achieved higher rates than single texture approaches.

## 5 Conclusions

We proposed a novel approach to quantify emphysema patterns based on global and local descriptors to form a single sequence that represent any given texture patch. This approach simultaneously encodes global characteristics with local information that leads to better classification rates. Additionally, we analyzed six texture descriptors and compared their performance. Since the size of extended descriptors increases exponentially, we applied KFDA via the kernel trick to avoid computing a mapping function. This procedure resulted in an improvement of the classification rates.

**Acknowledgments.** This work has been sponsored by the grant UNAM PAPIIT IN113611. R. Nava gives a special thank to CONACYT for the doctoral scholarship 167161; V. Marcos thanks the programme ”Juan de la Cierva” from the Spanish Ministry of Economy and Competitiveness; and R. San José was supported by NHLBI grants K25 HL104085-03 and 1R01HL116931-01.

## References

1. Galban, C., Han, M., Boes, J., Chughtai, K., Meyer, C., Johnson, T., Galban, S., Rehemtulla, A., Kazerooni, E., Martínez, F., Ross, B.: Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. Nat. Med. 18(11), 1711–1715 (2012)

2. Hayhurst, M., Flenley, D., Mclean, A., Wightman, A., Macnee, W., Wright, D., Lamb, D., Best, J.: Diagnosis of pulmonary emphysema by computerised tomography. The Lancet 324, 320–322 (1984)
3. Sørensen, L., Shaker, S., de Bruijne, M.: Quantitative analysis of pulmonary emphysema using Local Binary Patterns. IEEE Trans. Med. Imag. 29(2), 559–569 (2010)
4. Mendoza, C., Washko, G., Ross, J., Diaz, A., Lynch, D., Crapo, J., Silverman, E., Acha, B., Serrano, C., Estepar, R.: Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. In: 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 474–477 (2012)
5. Sørensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J., de Bruijne, M.: Texture-based analysis of COPD: A data-driven approach. IEEE Trans. Med. Imag. 31(1), 70–78 (2012)
6. Depeursinge, A., Foncubierta–Rodriguez, A., Van de Ville, D., Müller, H.: Multiscale lung texture signature learning using the riesz transform. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 517–524. Springer, Heidelberg (2012)
7. Gabor, D.: Theory of communication. J. Inst. Elec. Eng (London) 93III, 429–457 (1946)
8. Daugman, J.: Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J. Opt. Soc. Am. A 2, 1160–1169 (1985)
9. Nava, R., Escalante-Ramírez, B., Cristóbal, G.: Texture image retrieval based on log-gabor features. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 414–421. Springer, Heidelberg (2012)
10. Field, D.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A 4(12), 2379–2394 (1987)
11. Perrinet, L.U., Samuelides, M., Thorpe, S.J.: Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. Neurocomputing 57C (2002)
12. Perrinet, L.U.: Role of homeostasis in learning sparse representations. Neural Computation 22(7), 1812–1836 (2010)
13. Haralick, R., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst., Man, Cybern., Syst. SMC-3(6), 610–621 (1973)
14. Randen, T., Husøy, J.H.: Filtering for texture classification: A comparative study. IEEE Trans. Pattern Anal. Mach. Intell. 21, 291–310 (1999)
15. Marcos, V., Cristóbal, G.: Texture classification using Tchebichef moments. J. Opt. Soc. Am. A 30(8), 1580–1591 (2013)
16. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: 12th International Conference on Pattern Recognition - Conference A: Computer Vision Image Processing (IAPR), vol. 1, pp. 582–585 (1994)
17. Ojala, T., Pietikäinen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24(7), 971–987 (2002)
18. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.: Fisher discriminant analysis with kernels. In: IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing IX, pp. 41–48 (1999)

# Cervical Cell Classification Using Features Related to Morphometry and Texture of Nuclei

Juan Valentín Lorenzo-Ginori[1], Wendelin Curbelo-Jardines[1], José Daniel López-Cabrera[1], and Sergio B. Huergo-Suárez[1]

[1] Center for Studies on Electronics and Information Technologies,
Universidad Central "Marta Abreu" de Las Villas (UCLV), Cuba
juanl@uclv.edu.cu

**Abstract.** The Papanicolaou test is used for early prediction of cervical cancer. Computer vision techniques for automating the microscopy analysis of cervical cells in this test have received great attention. Cell segmentation is needed here in order to obtain appropriate features for classification of abnormal cells. However, accurate segmentation of the cell cytoplasm is difficult, due to cell overlapping and variability of color and intensity. This has determined a growing interest in classifying cells using only features from the nuclei, which are easier to segment. In this work, we classified cells in the pap-smear test using a combination of morphometric and Haralick texture features, obtained from the nucleus gray-level co-occurrence matrix. A comparison was made among various classifiers using these features and data dimensionality reduction through PCA. The results obtained showed that this combination can be a promising alternative in order to automate the analysis of cervical cells.

**Keywords:** Papanicolaou test, features, texture, dimensionality reduction, classifiers.

## 1 Introduction

Cervical cancer is, after breast cancer, the most common form of this disease among the female population. Early detection of this has contributed to a considerable reduction of the associated mortality rate. The Papanicolaou test [1] is the standard procedure currently used for early prediction of cervical cancer. In this test, microscopy analysis of the so-called pap-smear, a sample of cervical cells appropriately stained, is analyzed in a microscope to detect abnormal cells, which can be considered precursors of the disease. However, there are some drawbacks associated to microscopy analysis of the pap-smear by human experts: some rate of false negatives appears due to subjectivity, routine and tiredness of the analysts. This has determined a growing interest in developing automated analysis procedures using computer vision techniques. A typical image from the standard Papanicolaou test is shown in Figure 1.

The analysis of pap-smears through computer vision requires prior image segmentation, in order to extract appropriate features to classify the cells. A set of 20 morphometric features, half of them related to cells' cytoplasm have been used to classify

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 222–229, 2013.

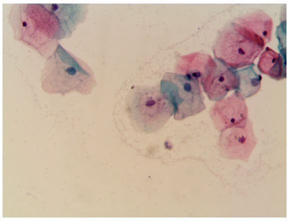


**Fig. 1.** A typical pap-smear image. The nuclei appear as dark spots within the clearer, coloured cytoplasm.

the cervical cells in seven well defined classes, from which a benchmark database has been built [2]. Three out of the seven cervical cells classes are considered as normal and four of them are abnormal, thus raising two problems: classification in seven classes, and binary classification in normal-abnormal. This work addresses the latter.

A formal attempt to classify cervical cells using different classifiers and the previously mentioned features from nuclei and cytoplasm, with feature selection using genetic algorithms, is shown in [3]. Other classification approaches can be found in [4] and [5], using features that are mostly associated to cells' morphometry, and the latter using also four textural features, in this case applied to a technique called liquid-based cervical smears, that differ from the standard pap test to which the present work is devoted. In regard to segmentation, some works address the problem of finding an appropriate way to segment the cytoplasm [6], however, it has been found that the cell cytoplasm is very difficult to segment with good accuracy, due to overlapping of cells and their variability of color and intensity. On the other hand, the cell nuclei usually appear better defined in the images, and this allows improving segmentation accuracy. Therefore, some works have been devoted to nuclei segmentation in pap-smears [7], [8]. This situation has led to conduct research on the possibility of classifying cells using only features obtained from their nuclei as in [9], where nine nuclei morphometric features were used.

In this work, we explore the possibility of improving classification of cells in the pap-smear using information from the nuclei only, but additionally including texture features. This was motivated by the fact that staining in pap smears makes visible the chromatin textural patterns in the nuclei and this information is used by cytopathologists to classify the cells [10].

Here a method using morphological image processing [11] was employed to calculate, from a given nucleus image, its gray-level co-occurrence matrix (GLCM) and the associated Haralick features associated to texture [12]. These features were used, together with the previously known morphometric ones, to construct a combined feature matrix. Dimensionality reduction using principal components analysis (PCA) was also employed, given the relatively large amount of features obtained. Four classifiers were used: linear, Mahalanobis distance, K-nearest-neighbors (KNN) and support vector machines (SVM). Comparison among classifiers' performance was made through statistical methods [13] and results showed that using this combination of features, the binary classification results can be improved.

This article is organized as follows: in section 2, the main characteristics of the cells used in the experiments are described, as well as the methods used to calculate the texture features and the experiments to test the classifiers' performance. We summarize and discuss the main results in section 3, and conclusions are exposed in section 4.

## 2 Materials and Methods

The cervical microscopy cell images used in this work were obtained from the Herlev database [2], which contains 917 annotated images, each with a manually segmented version that can be used as ground-truth. The different classes to which they belong are shown in Table 1.

**Table 1.** Cervical cells in the Herlev image database

| Class | Category | Cell type | Number of cells | Sub-totals |
|---|---|---|---|---|
| 1 | Normal | Superficial squamous epithelial | 74 | |
| 2 | Normal | Intermediate squamous epithelial | 70 | |
| 3 | Normal | Columnar squamous epithelial | 98 | 242 |
| 4 | Abnormal | Mild squamous non-keratinizing dysplasia | 182 | |
| 5 | Abnormal | Moderate squamous non-keratinizing dysplasia | 146 | |
| 6 | Abnormal | Severe squamous non-keratinizing dysplasia | 197 | |
| 7 | Abnormal | Squamous cell carcinoma in situ intermediate | 150 | 675 |

### 2.1 Texture Analysis

The set of features employed for classification is shown in Table 2, which included both morphometric and Haralick texture features. In order to obtain the latter, the GLCM from the cell nuclei was calculated. Once the nuclei had been segmented (an operation that is out of the scope of this paper), an image represented by a square matrix containing the texture pattern of the nucleus region was obtained. The problem associated to acquiring this matrix was formulated in terms of obtaining the largest *x*-*y* oriented square that can be inscribed in the nucleus region. This operation was performed through morphological image processing in the following way:

1. Obtain a binary mask which represents the location of the segmented nucleus.
2. Successively erode the nucleus mask with a square structuring element (SE) increasing its side length one pixel per iteration, until the nucleus mask disappears.
3. Go back one step and obtain the square SE that, when its side is enlarged just one pixel, completely erodes the nucleus mask. Using the erosion of the nucleus mask with this SE, pick one pixel from the resulting binary image.
4. With the same SE used in step (3), perform a dilation of this pixel. The result will be a largest inscribed square in the selected nucleus.
5. Use this binary square as a mask and perform an array multiplication with the nucleus image to obtain a square matrix in grayscale containing the nucleus texture.
6. Expand linearly the intensity of this image to the maximum interval allowed, in order to enhance its contrast.

This process is illustrated in Figure 2. The co-occurrence matrix of this image [12] was calculated for three pixel offset values (1, 4, 7) and four spatial orientations (0, π/4, π/2, 3π/4), forming a three dimensional array in which the spatial orientation corresponds to dimension 3. A final co-occurrence matrix was obtained by selecting the maximum along dimension 3, as we are interested in the most significant co-occurrence values independently of their associated spatial orientation.

**Table 2.** Morphometric and texture features

| **Nuclei features/morphometric** | |
|---|---|
| Mean intensity | Area |
| Maximum intensity | Perimeter |
| Minimum intensity | $(\text{Area})^{1/2}$/Perimeter |
| Solidity | Entropy (grayscale image) |
| Eccentricity | |
| **Nuclei features/Haralick coefficients** | |
| Autocorrelation | Sum of squares: Variance |
| Contrast | Sum average |
| Correlation | Sum variance |
| Cluster prominence | Sum entropy |
| Cluster shade | Difference variance |
| Dissimilarity | Difference entropy |
| Energy | Information measure of correlation 1 |
| Entropy | Information measure of correlation 2 |
| Homogeneity 1 | Inverse difference (INV) |
| Homogeneity 2 | Inverse difference normalized (INN) |
| Maximum probability | Inverse difference moment norm. |

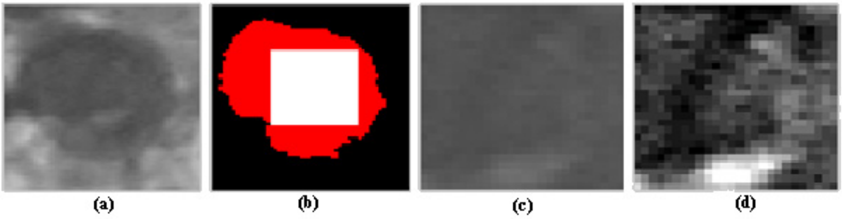


**Fig. 2.** A sample is obtained from a cell's nucleus. (a) Grayscale image from the Herlev database, (b) the corresponding nucleus mask (red) and the inscribed square (white), (c) the grayscale sample obtained using the mask and (d) the sample after enhancing its contrast.

After this, the set of Haralick features was calculated, finally forming a feature matrix of size $N_c \times (22+9)$, where $N_c$ is the amount of cells contained in each class in Table 1.

### 2.2 Cell Classification

The cell classification process had two purposes: determining if using texture features meant a statistically significant improvement in classification accuracy and comparing the performance of various classifiers. Classes 3 and 4 from the Herlev database (see Table 1) were used to perform an abridged evaluation. The indexes of classifiers' effectiveness used were: sensitivity (se), specificity (sp), positive and negative predictive values (pp and np) and the F-measure (harmonic mean of se and sp) with emphasis in the last two. The classifiers evaluated were: linear, Mahalanobis distance, *k* nearest neighbors (KNN) and (after testing some SVM options), a Gaussian radial basis function kernel SVM with σ=2. In all cases, dimensionality reduction (DR) by principal components analysis (PCA) was employed. Several values of DR were employed and the best among them was used when comparing the classifiers.

An m-fold cross-validation (m=20) was performed in which the indexes of effectiveness were calculated. The series of indexes values were used for determining, using the Friedman test [13], if there was a statistically significant difference in two situations: (1) among the alternatives of features used: morphometric only, texture only or both, with various alternatives of dimensionality reduction, and (2) among the various classifier algorithms employed.

## 3 Results and Discussion

### 3.1 Classification Performance for Different Alternatives of Features

After performing several tests, we determined the most favorable values for PCA data reduction as well as the classifier with best performance. Table 3 shows the values of the indexes of effectiveness for the Mahalanobis classifier, with morphometric features only, without PCA data reduction and with DR to seven features, as well as using texture only and all the features, the latter two with 17 features after DR. Notice that the highest pn and F-measure values were obtained using all features with DR to

17 features. The Friedman test was realized using results from a 20-fold cross-validation, with the four data reduction alternatives as related samples. The result for the Mahalanobis classifier is shown in Table 4, in which the higher rank was obtained for the all-features case, DR to 17, for the F-measure (p<0.05). Further pair-wise analysis using the Wilcoxon signed rank test tended to confirm the superiority of this alternative, although this result is somewhat limited due to correlation in the training data. Results were similar for the np index, and for the rest of the classifiers, the best performance in most cases appeared when using all features with DR to 17, regarding F-measure and np. Image resolution and offset values could also affect these results.

### 3.2 Determining the Best Classifier

After performing the previous experiments, we made a comparative evaluation of the various classifiers, again with an m-fold cross-validation with m=20, now using the best alternative, i. e., both morphometric and texture features with PCA and 17 features after DR. The Friedman test was used again, for which we made the classification using for all the classifiers the same grouping of cells in training and test sets in the cross-validation. The corresponding results are shown in Table 5. The ranks obtained suggest that the SVM classifier was the best. However, further pair-wise Wilcoxon test showed that the statistically significant differences are among SVM-Mahalanobis and KNN-linear, the former pair being better in comparison to the second, with no statistically significant difference between the paired methods. However, there is an important difference in terms of computer time, as shown in Table 6. A binary classification for the whole Herlev database was made, and its results, although inferior, were consistent with those from the abridged experiment.

**Table 3.** Results for the Mahalanobis classifier

| | **Morphometric features** | | **Texture features** | **Combined features** |
|---|---|---|---|---|
| | without DR | DR to 7 | DR to 17 | DR to17 |
| np | 0,895 | 0,823 | 0,677 | **0,910** |
| F_measure | 0,916 | 0,917 | 0,854 | **0,971** |

**Table 4.** Results of the Friedman test for the for the Mahalanobis classifier, $p < 0.05$

| **Alternative** | **Mean rank, F** | **Mean rank, np** |
|---|---|---|
| Morphometric features without DR | 2, 80 | 3,10 |
| Morphometric, 7 features after DR | 2,23 | 2,03 |
| Haralick, 17 features after DR | 1,00 | 1,00 |
| Combining all features, DR to 17 | **3,98** | **3,88** |

**Table 5.** Results of the Friedman test for the F-measure obtained with the different classifiers using the combined features with DR, $p < 0.05$

| Classifier | Mean rank, F | Mean rank, np |
|---|---|---|
| Linear, combined, DR | 2,03 | 2,33 |
| Mahalanobis, combined, DR | **4,00** | 3,25 |
| KNN, morphometric, no DR | 1,00 | 1,00 |
| SVM, combined, DR | 2,98 | **3,43** |

**Table 6.** Computer time for the various classifiers

| Classifier | Time, s |
|---|---|
| Linear | 0,152 |
| Mahalanobis | 0,151 |
| SVM | 14,619 |
| KNN | 0,155 |

## 4 Conclusions

In this work, binary classification of cells in the Papanicolaou test was performed using features from the cells' nuclei only. Morphometric features of the nuclei were calculated firstly. Then a square sub-image from each cell nucleus was extracted using morphological image processing, and its GLCM and the associated Haralick features were calculated. These were combined with the morphometric data to build a feature matrix, whose dimensionality was reduced through PCA. An abridged m-fold cross-validation experiment using classes 3 and 4 described in Table 1 was made. Results showed a more accurate classification in terms of the negative predictive value and the F-measure in comparison to using morphometric data only. From the classifiers tested: linear, KNN, Mahalanobis and SVM, the latter two showed better results. Classification results using the whole database, although inferior compared to the abridged experiment, tended to confirm the advantages of using also nuclei texture features. Evaluation of the classifiers was made using statistical hypothesis testing.

The results obtained showed advantages from using also texture features when classifying cells in the Papanicolaou test using data from the nuclei only. This suggests a number of alternatives to be evaluated in future work, for example: exhaustive search for the best offsets in the GLCM, new methods to extract texture features like morphological granulometry, using kernel PCA or feature selection methods for dimensionality reduction, and trying other classifying algorithms.

**Acknowledgments.** The authors acknowledge M.D. Irene Rodríguez-Santos for her collaboration and Ph.D. Julián L. Cárdenas-Barreras for reviewing the manuscript.

This work received support from the Canadian International Development Agency (CIDA) Project "Biomedical Engineering Education in Cuba", from the Institutional University Co-operation (IUC) program of the Flemish University council (VLIR) and from the Center for Studies in Electronics and Information Technologies at UCLV.

## References

1. Papanicolaou, G.N.: A new procedure for staining vaginal smears. Science 95, 438–439 (1942)
2. Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B.: Pap-smear Benchmark Data For Pattern Classification. Presented at the Nature Inspired Smart Information Systems, NiSIS (2005)
3. Marinakis, Y., Dounias, G., Jantzen, J.: Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. Comput. Biol. Med. 39, 69–78 (2009)
4. Mat-Isa, N.A., Mashor, M.Y., Othman, N.H.: An automated cervical pre-cancerous diagnostic system. Artif. Intell. Med. 42, 1–11 (2008)
5. Huang, P.-C., Chan, Y.-K., Chan, P.-C., Chen, Y.-F., Chen, R.-C., Huang, Y.-R.: Quantitative Assessment of Pap Smear Cells by PC-Based Cytopathologic Image Analysis System and Support Vector Machine. In: Zhang, D. (ed.) ICMB 2008. LNCS, vol. 4901, pp. 192–199. Springer, Heidelberg (2007)
6. Kale, A., Aksoy, S.: Segmentation of Cervical Cell Images. Presented at the IEEE 2010 International Conference on Pattern Recognition, August 23 (2010)
7. Sobrevilla, P., Montseny, E., Vaschetto, F., Lerma, E.: Fuzzy-based analysis of microscopic color cervical pap smear images: nuclei detection. Int. J. Comput. Intell. Appl. 9, 187–206 (2010)
8. Plissiti, M.E., Nikou, C., Charchanti, A.: Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering. IEEE Trans. Inf. Technol. Biomed. 15, 233–241
9. Plissiti, M.E., Nikou, C.: Cervical Cell Classification Based Exclusively on Nucleus Features. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part II. LNCS, vol. 7325, pp. 483–490. Springer, Heidelberg (2012)
10. Patten, S.F.: Diagnostic cytopathology of the uterine cervix. S. Karger AG, Basel (1978)
11. Soille, P.: Morphological Image Analysis: Principles and Applications. Springer (2004)
12. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621 (1973)
13. Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30 (2006)

# Study of Electric and Mechanic Properties of the Implanted Artificial Cardiac Tissue Using a Whole Heart Model*

Sándor Miklos Szilágyi[1,2], László Szilágyi[3], and Béat Hirsbrunner[1]

[1] University of Fribourg, Fribourg, Switzerland
[2] Petru Maior University of Tîrgu-Mureş, Romania
szsandor72@yahoo.com
[3] Budapest University of Technology and Economics, Budapest, Hungary

**Abstract.** This study focuses on the effects of artificial cardiac tissue in the excitation-contraction process of the ventricular muscle. We developed a spatio-temporal computerized model of the whole heart that handles half millimeter sized compartments using 1 microsecond time step. We employed the effect of muscle fiber direction, laminar sheets, depolarization period and other parameters. The artificial tissue differs from the normal one in several ways, so their describing parameters are also modified. In our simulation the depolarization wave (DW) conduction speed of the artificial tissue was decreased by up to 3 times. In presence of a two centimeter wide and 2 mm thick artificial tissue slice, the maximal depolarization delay was 38 msec. Large ventricle size, low conducting speed and spaciousness of the injured ventricular tissue are the main generating factors of arrhythmia, while the location of the artificial tissue has secondary importance.

**Keywords:** ventricle modeling, geometry estimation, interpolation techniques.

## 1 Introduction

In developed countries, cardiac failure induced by myocardial infarction, despite several decades of research, still represents an important mortality factor. In spite of the advances in surgical techniques, immune system suppression and post-operative health-care, the life-saving and -extending effect of cardiac transplantation remains limited by shortage of proper donors and weakened immune system [1]. The development of proper artificial cardiac tissue (ACT) may eliminate both limiting factors [2].

As a first organ transplant substitution method, scientists have developed cell-based therapies, where special myocardial cells were injected into the tissue involved in infarction [3]. This approach was hardened by a massive apoptosis

* This work has been funded by the Scientific Exchange Program NMS-CH, "Rou: Swiss Research Fellowships", Project code 12031.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 230–237, 2013.

of the injected cells (about 90%) and the low rate of successfully differentiation into cardiomyocytes [4].

Nowadays the construction of a properly functioning “bio-artificial human heart” is still far away, but it became possible to create of three-dimensional (3D) muscle equivalents that can be useful for cardiac regeneration. As a result of continuous progress in the last two decades, the transplantation of contractile ACT and the replacement of degenerated tissue areas represent an important alternative to the whole organ transplantation [1]. Motivated to develop minimally invasive procedures, physicians' aim is to create biocompatible, non-immunogenic heart muscle that has similar morphological and functional properties as healthy cardiac tissue.

The myocardium, due to the cardiac progenitor cells [5], has a limited ability to recover after a serious injury, so the artificial tissue has to be surgically attached to the damaged area [6]. In order to introduce the new tissue in the ordinary work process, physicians have to create a proper capillary network [7]. The main obstacle of the integration of artificial tissue into the organ represents the cell apoptosis, caused by insufficient oxygen level in newly introduced tissue [8]. In the presence of this pathological condition, several parts of the artificial tissue may modify their electrical and mechanical properties that can develop altered depolarization and repolarization waves, causing rhythm irregularities [9]. The dysfunction of electrical impulse propagation may develop cardiac arrhythmias that perturb pump activity [10].

In the last decade several mathematical models and intelligent computational methods were developed in order to perform real-time computerized simulations of the whole heart, creating a useful tool to study cardiac dynamics [10,11]. These simulations have many advantages: they are not perturbed by data acquisition errors, the simulated values of all internal variables may be visualized, the size and nature of artificial tissue may be studied before the real intervention and the simulation may be stopped at any time for further improvements [12].

In the following we present the main benefits and dangers of the artificial tissue implantation procedure. Our goal is to establish a modeling platform that can a priori show the expected results of a future implantation. The negative effects of an eventually inaccurate operations can also a priori analysed.

The main goal of this paper is to model the onset of possible rhythm problems that can endanger the patient's life. The rest of the paper is organized as follows: Section 2 gives a detailed description of the cardiac excitation and contraction in presence of artificial tissue. Section 3 presents and discusses several aspects of modified depolarization and cardiac pump functionality, and the results of simulations. In Section 4, the conclusions are formulated.

## 2 Materials and Methods

### 2.1 Modeling Background

In our study we used a multi-level modeling technique that is visualized in Fig. 1. Each modeling level and the main descriptor parameters were described in

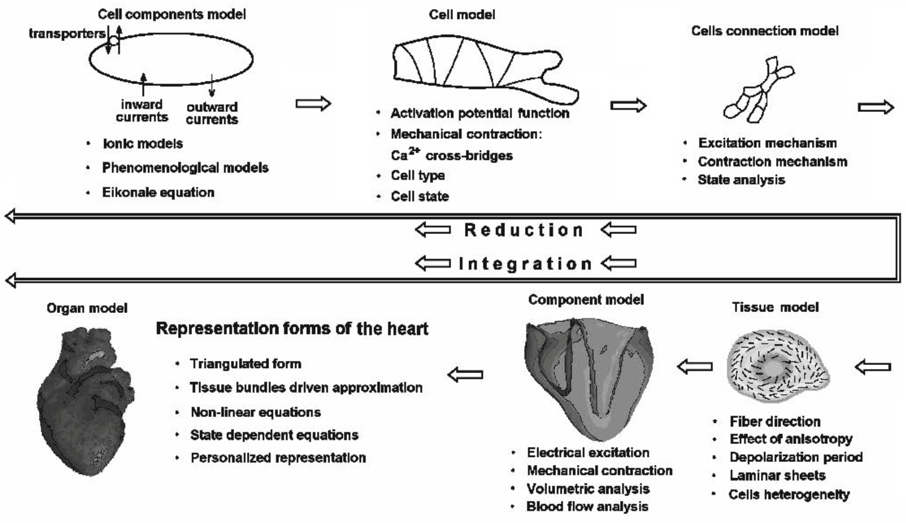


**Fig. 1.** The hierarchical structure of heart modeling, with all possible levels from individual cells to whole organ

our earlier work [12]. As described in the above cited paper, the main modeling levels are: cell (type and state), cell connection, compartments, cardiac tissue (type, structure and state), component and whole organ.

In our simulation the lowest entity is considered a compartment that has a homogenous structure. These entities may contain only one cardiac cell type. Each compartment may be in normal or pathological state, which describes its electrical and mechanical behavior. All of these modeling units have unique activation potential function, mechanical contraction rules and a connection model that determines the propagation of the depolarization wave and the mechanical contraction of the cells. Moreover, the Purkinje system of the ventricles is also constituted by such compartments. These Purkinje units are also included in the connection system of the compartments.

As we move toward integration, we have to define the main properties of each integration element, such as: tissue, component and whole organ. The basic tissue parameters, such as fiber direction, anisotropy, depolarization period, laminar sheets and cell inhomogeneity are determined by its consisting compartments. The used component models enable us to determine the electrical excitation and mechanical contraction of the heart chambers, thus supporting the volumetric analysis and blood flow analysis for the given component or for the whole organ.

## 2.2 The Main Properties of Artificial Cardiac Tissue

The native cardiac tissue is a mixed structure of cardiac myocytes, fibroblasts, smooth muscle cells, endothelial cells and macrophages. The population of each cell type varies by state, age, gender and tissue's position. In our model the ratio and nature of excitable cells is defined in tissue properties, such as: fiber spatial orientation, level of anisotropy, shape of the activation potential that determines the depolarization period, laminar sheets, cells and structure inhomogeneity [12]. Some biological properties are reflected indirectly in our model, for example the number and spreading manner of fibroblasts are not expressed directly but introduced in the level of conduction speed and level of anisotropy. These simplifications allow a much higher simulation speed, while the most important biological parameters are not altered notably.

In artificial tissues, the cellular complexity is drastically lower than in normal cardiac tissue. Several cell types are not present at all, and the overall structure is more homogeneous. The cellular complexity can be enhanced by using unpurified cardiac cell populations, but the cell arrangement cannot be controlled sufficiently. A partially controlled cellular arrangement not always yields better electro-mechanical properties instead of substantially higher cellular diversification.

The mechanical properties of the artificial tissues significantly differ from of the native ones. They usually are more sensitive to calcium regulation and may produce a lower absolute force than native cardiac tissue.

Nowadays all engineered cardiac tissues suffer from the absence of vascularization and perfusion. It is known that tumors cannot reach more than 3 mm diameter without capillarization. In case of cardiac muscle that makes permanent effort during contraction, the presence of sufficient nutrients and oxygen is imperial, so individual sheets cannot be thicker than 2 mm. These layers can be partially vascularized after the implantation [13]. The construction of a thicker tissue layer demands the development of an inter-layer capillary system that nowadays is a challenging physiological task.

The connection area between the implanted artificial tissue slice and native myocardial cell allows significantly slower depolarization propagation than both native and implanted tissue.

### 2.3 Details of the Simulation

The compartment-based simulation uses an adaptive spatio-temporal resolution, so a 0.5 mm spatial and 1 $\mu$sec temporal resolution is used in case of depolarization, and a significantly lower resolution (up to 5 mm and 1 msec) in resting phase. The instantaneous resolution, given for each region and time segment separately, depends on the derivative of the action potential (AP) function, connections of the simulated compartment, nature of studied phenomena and some restrictions implied by hardware or total simulation time considerations. The highest spatial and temporal resolution is needed at the depolarization wave's front line that propagates in an inhomogeneous and fast conducting tissue due to the fast voltage rise caused by fast sodium current [14]. The used spatio-temporal resolution may vary in time due other important factors, such as: simulation of various pathological cases especially arrhythmias, fragmentation of the depolarization front line and presence of spacious low- or non-conducting isles.

The above presented compartment-based representation was used to simulate the electrical and mechanical behavior of all of cardiac cell types. The internal state of each compartment is modeled separately, which allows their investigation during the whole simulation. The connections among compartments are a priori determined, but its properties may vary with both space and time, so we can properly model the propagation of the depolarization wave and the mechanical contraction of the compartments in almost all circumstances.

Several time- and state-dependent tissue-related parameters were involved in our model that greatly influences the behavior of the compartment groups, such as: fiber direction, level of anisotropy, average depolarization period, laminar sheets and spontaneous cell inhomogeneity. The above mentioned parameters were deduced for each compartment separately from the simulation circumstances. The study of these parameters enables us to determine the electrical excitation and mechanical contraction of the cardiac muscle, thus supporting the volumetric analysis for atria and ventricles.

The tissue level excitation mechanism is based on Fast's work [15], but their results were transformed into compartment compatible data, considering each compartment as a secondary generator element, while the activation potential applied for ventricular tissue compartments was determined by using the Luo-Rudy II (LR) ventricular cell model [16,17]. Each compartment may generate a depolarization wave if any adjacent elements are repolarized; otherwise, the propagation is swooned [12]. The LR model accounts for dynamic changes of ionic concentrations, so it can properly handle several pathological cases. Although it contains few dozens of parameters instead of several hundreds used in newer ventricular models [11], the propagation of depolarization wave in the artificial tissue can be simulated properly.

During the simulation of a healthy cardiac activity, we employed the effect of: muscle fiber direction (the ratio between longitudinal and transversal conductivity varies from 2 to 10), normal and minimal depolarization period (considered 80-250 msec), laminar sheet effect (in-sheet transversal conduction 2-5 times faster than trans-sheet conduction), and cell inhomogeneity (using conduction speed differences for base-apex gradient (5%-20%), transmural epicardial-endocardial gradient (5%-35%), left-right ventricular gradient (5%-15%)).

For pathological cases, normal parameter values were no longer maintained. In our simulation the depolarization wave (DW) conduction speed of the injured-but still functioning-tissue was decreased by up to 20 times.

The simulation of various pathological circumstances, of the artificial tissue region or its barriers was performed using altered parameters [18,19]. For example the effect of various anatomical modifications were considered as: muscle fiber direction (the ratio between longitudinal and transversal conductivity varies from 1 to 3), normal and minimal depolarization period (considered 70-350 msec), laminar sheet effect (in-sheet transversal conduction 1-2.5 times faster than trans-sheet conduction), and cell inhomogeneity (using conduction speed differences for base-apex gradient (0%-25%), transmural epicardial-endocardial gradient (0%-50%), left-right ventricular gradient (0%-25%)).

## 3 Results

Fig. 2(left) presents the depolarization time of the ventricles in presence of a serious injury, covered by a two millimeter wide slice of artificial tissue. The injury is situated in the left paraseptal location. The depolarization time was calculated from the excitation moment of the atrio-ventricular (AV) node-HIS bundle system. The dark area visualizes the later excited tissue.

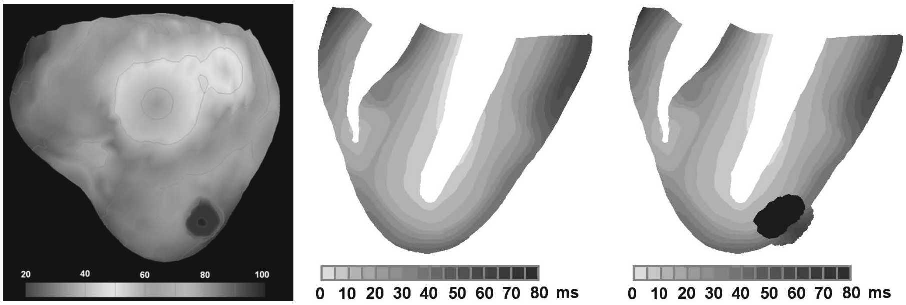


**Fig. 2.** (left) The depolarization time of the ventricles in presence of a 2 mm wide artificial tissue, implanted in top of the injured region situated in the left anterior paraseptal location. The depolarization time was determined from the excitation moment of the AVnode-HIS bundle system, and expressed in msec; (middle and right) The sectional representation of the depolarization time in ventricles. In the left side a normal ventricle is presented, while in the right side an anatomically similar ventricle had a serious injury, covered by a two millimeter wide slice of artificial tissue. The depolarization time was determined from the excitation moment of the AVnode - HIS bundle system. The sectional representation of the depolarization time in ventricles. In the left side a normal ventricle is presented, while in the right side an anatomically similar ventricle had a serious injury, covered by a two millimeter wide slice of artificial tissue. The depolarization time was determined from the excitation moment of the AVnode - HIS bundle system.

In the middle and right sections of Fig. 2, we get an insight into the depolarization phenomena of the inner ventricular structure, solved by a sectional representation. As shown in the figure, the depolarization time was counted from the excitation moment of the AV-node-HIS bundle system. In the middle section of Fig. 2 a healthy ventricular tissue is presented, while in the right section the ventricular tissue is seriously damaged, and is covered by a functioning slice of artificial tissue.

Table 1 presents the simulation results for healthy and injured tissue. The total depolarization time of the ventricle in presence of a 2 cm wide and 2 mm thick artificial tissue, situated in the right posterior region can reach 109 msec (counted from the excitation moment of the AV-node-HIS bundle system), while in case of a healthy ventricle the total depolarization time was 77 msec. The maximal depolarization delay highly depends on the size of the artificial tissue and slightly from the injured region. The 38 msec maximal delay was

**Table 1.** Simulated physiological parameters

| Case study | Simulation parameters of the ventricles | | |
|---|---|---|---|
| | Studied phenomena | Healthy tissue | Injury and artificial tissue |
| 1 | Total depolarization time | 77 msec | 109 msec |
| 2 | Maximal depolarization delay | 0 msec | 38 msec |
| 3 | Contractile efficiency | 100 % | about 65 % |
| 4 | Maximal heartbeat rate (beats/min) | 300 | about 250 |

determined for a 2 cm wide ACT situated in the posterior region, while the subjacent tissue was completely isolator. The contractile power of an injured ventricle with a 2 cm wide infarcted area in best case may reach half of the normal value. Our simulations show a 15% contractile power increase due the presence of ACT. Due to the presence of artificial tissue, de maximal heart bit rate was reduced by 50 beats per minute. It is important to mention that an altered shape of depolarization and repolarization may induce arrhythmias. As the heart rate becomes higher, the additional risk may drastically increase. The maximal heart beat was determined from the increase of the total cardiac depolarization-repolarization cycle duration. We expected that ACT cells repolarize at least as fast as the middle (m cells) of the ventricular tissue.

In case of high heart rate the delayed excitation of the ACT may induce irregular depolarization process that can develop various arrhythmias. The level of delayed excitation of ACT highly depends on its size and the nature of the subjacent tissue.

## 4 Conclusions

We created a simulation environment to show the effects of artificial tissue. The simulation was performed for a 2 cm wide and 2 mm thick artificial tissue. From the results of this simulation we concluded that the artificial tissue may enhance the cardiac pumping function, but also may increase the chance to develop arrhythmias. Numerical calculation confirms that occurring arrhythmia may develop ventricular fibrillation. This deadly phenomenon is promoted by diverse factors, such as: inhomogeneity in ventricular tissue, high excitation frequency, presence of accessory pathways, slow depolarization (due to thick walls) or repolarization and greater than normal ventricular size. Computerized simulation represents a non-invasive visualization tool than helps us understand the inner cardiac process in normal and pathological cases, and may help to select the most endangered patients that can enhance the efficiency of health care.

## References

1. Kofidis, T., Akhyari, P., Boublik, J., Theodorou, P., Martin, U., Ruhparwar, A., et al.: In vitro engineering of heart muscle: Artificial myocardial tissue. J. Thorac. Cardiov. Sur. 124, 63–69 (2002)

2. Venugopal, J.R., Prabhakaran, M.P., Mukherjee, S., Ravichandran, R., Dan, K., Ramakrishna, S.: Biomaterial strategies for alleviation of myocardial infarction. J. R. Soc. Interface 9(66), 1–19 (2012)
3. Heldman, A.W., Hare, J.M.: Cell therapy for myocardial infarction: Special delivery. J. Mol. Cell Cardiol. 44, 473–476 (2008)
4. Robey, T.E., Saiget, M.K., Reinecke, H., Murry, C.E.: Systems approaches to preventing transplanted cell death in cardiac repair. J. Mol. Cell Cardiol. 45, 567–581 (2008)
5. Jawad, H., Lyon, A.R., Harding, S.E., Ali, N.N., Boccaccini, A.R.: Myocardial tissue engineering. Brit. Med. Bull. 87, 31–47 (2008)
6. Miyagawa, S., Roth, M., Saito, A., Sawa, Y., Kostin, S.: Tissue-engineered cardiac constructs for cardiac repair. Ann. Thorac. Surg. 91, 320–329 (2011)
7. Radisic, M., Park, H., Gerecht, S., Cannizzaro, C., Langer, R., Vunjak-Novakovic, G.: Biomimetic approach to cardiac tissue engineering. Phil. Trans. R. Soc. B 362, 1357–1368 (2007)
8. Hool, L.C.: Acute hypoxia differentially regulates K(+) channels. Implications with respect to cardiac arrhythmia. Eur. Biophys. J. 34, 369–376 (2005)
9. Bueno-Orovio, A., Cherry, E.M., Fenton, F.H.: Minimal model for human ventricular action potentials in tissue. J. Theor. Biol. 253, 544–560 (2008)
10. Cherry, E.M., Fenton, F.H.: Visualization of spiral and scroll waves in simulated and experimental cardiac tissue. New J. Phys. 10, 125016 (2008)
11. ten Tusscher, K.H.W.J., Bernus, O., Hren, R., Panfilov, A.V.: Comparison of electrophysiological models for human ventricular cells and tissues. Prog. Biophys. Mol. Bio. 90, 326–345 (2006)
12. Szilágyi, S.M., Szilágyi, L., Benyó, Z.: A patient specific electro-mechanical model of the heart. Comput. Meth. Prog. Bio. 101, 183–200 (2011)
13. Leor, J., Aboulafia-Etzion, S., Dar, A., Shapiro, L., Barbash, I.M., Battler, A., Granot, Y., Cohen, S.: Bioengineered cardiac grafts: A new approach to repair the infarcted myocardium? Circulation 102(III), 56–61 (2000)
14. Cherry, E.M., Greenside, H.S., Henriquez, C.S.: A space-time adaptive method for simulating complex cardiac dynamics. Phys. Rev. Lett. 84, 1343–1346 (2000)
15. Fast, V.G., Rohr, S., Gillis, A.M., Kleber, A.G.: Activation of cardiac tissue by extracellular electrical shocks: formation of 'secondary sources' at intercellular clefts in monolayers of cultured myocytes. Circ. Res. 82, 375–385 (1998)
16. Luo, C.H., Rudy, Y.: A dynamic model of the cardiac ventricular action potential I. Simulations of ionic currents and concentration changes. Circ. Res. 74, 1071–1096 (1994)
17. Luo, C.H., Rudy, Y.: A dynamic model of the cardiac ventricular action potential. II. Afterdepolarizations, triggered activity, and potentiation. Circ. Res. 74, 1097–1113 (1994)
18. Rădoiu, D., Enăchescu, C., Adjei, O.: A systematic approach to scientific visualization. Eng. Comput. 23, 898–906 (2006)
19. Enăchescu, C.: Neural networks for function approximation. In: Int. Conf. Bio-Inspired Comput. Meth. Used for Difficult Problem Solving (BICS 2008), pp. 84–89 (2008)

# Adaptive H-Extrema for Automatic Immunogold Particle Detection

Guillaume Thibault, Kristiina Iljin, Christopher Arthur, Izhak Shafran, and Joe Gray

Oregon Health & Science University (OHSU),
3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA

**Abstract.** Quantifying concentrations of target molecules near cellular structures, within cells or tissues, requires identifying the gold particles in immunogold labelled images. In this paper, we address the problem of automatically detect them accurately and reliably across multiple scales and in noisy conditions. For this purpose, we introduce a new contrast filter, based on an adaptive version of the H-extrema algorithm. The filtered images are simplified with a geodesic reconstruction to precisely segment the candidates. Once the images are segmented, we extract classical features and then classify using the majority vote of multiple classifiers. We characterize our algorithm on a pilot data and present results that demonstrate its effectiveness.

**Keywords:** Adaptive H-extrema, Mathematical morphology, Immunogold particle detection, Pattern recognition.

## 1 Introduction

Immunogold staining (IGS) is a technique used in electron microscopy (EM) to localize a molecule of interest – target molecule. This often achieved by attaching a primary antibody to the molecule of interest, which is then linked to the immunogold particle through a secondary antibody. After the gold particles are attached to the target molecules in this manner, the specimen is imaged using an electron microscope where the gold particles appear as “dark spots” due to the high electron density (see image 1). The IGS allows indirect visualization of target proteins and their approximate locations (the distance between primary antibody and immunogold is in range 15 to $30nm$). The immunogold particles are extremely small and so the IGS is typically employed in studies where cells or tissues are imaged at high resolution. The high resolution images in such studies are manually tagged, which is a time consuming process.

In this paper, we describe a new scheme to automatically detect the immunogold particles in high resolution images. We first explain the challenges in this problem (section 2), and then describe a new adaptive version of $H - extrema$, a mathematical morphology algorithm (section 3.1) for accurately detecting the particles in all conditions. In Section 4, we evaluate our method empirically to understand its capabilities and limitations and report results.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 238–245, 2013.

## 2 Problem Description

Immunogold particles appear as dark spots under good imaging conditions. However, the acquired EM images vary significantly depending on conditions of image acquisition, which are difficult to control precisely. These variations makes it difficult to detect or locate the immunogold particles in the image. The most common variation is the magnification of the EM images; figure 1 (left and middle) shows the same group of golds acquired with two different magnifications. The change in magnification, not only impacts the scale of the objects in the view, but also the shape and the intensity profile of the gold. Moreover, as shown in Figure 1, the intensity of the image is effected by the presence of relatively larger (dark) structures in the close neighborhood. Another factor that influences the quality of the image relates to noise or fuzziness, as shown in Figure 1 (right), arising from variations in specimen preparation, image acquisition, clustering of particles in the same location and the nature of organic tissue. The above mentioned variations can substantially affect the appearance of the gold particles, making the task of automatic detection challenging.

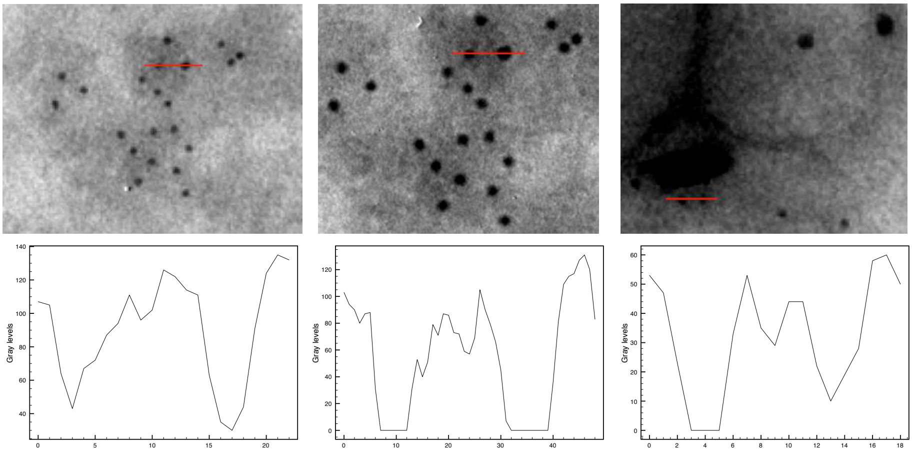


**Fig. 1.** Example of set of golds acquired at different magnifications (top left and middle), with different contrasts in the same original image (top middle and right), and their corresponding intensity profiles along the red segments

## 3 Our Approach

Let $f : \begin{cases} E \to \mathcal{T} \\ \mathbf{x} \mapsto f(\mathbf{x}) \end{cases}$ be a gray-levels image, where $E \subset \mathbf{Z}^2$ is the support space of pixels and the image intensities are discrete values which range in a closed set $\mathcal{T} = \{t_1, t_2, ..., t_N\}$, $\Delta t = t_{i+1} - t_i$, e.g., for a 8 bits image we have $t_1 = 1$, $N = 256$ and $\Delta t = 1$. Let us assume also that image $f$ is segmented into its

$J$ flat zones $R_j[f]$ (i.e., connected regions of constant value): $E = \cup_{j=1}^{J} R_j[f]$, $\cap_{j=1}^{J} R_j[f] = \emptyset$. The size (surface area) of each region is $s(j) = |R_j[f]|$ ($|.|$ the cardinal). Hence, we consider that each zone $R_j[f]$ has associated a constant gray-level intensity $g(j)$.

### 3.1 Review: H-Extrema

H-extrema, a mathematical morphology algorithm [1,2], is a powerful non-linear filter to detect structures with certain intensity profile. The algorithm is comprised of two distinct algorithms – the h-minima and its dual operation the h-maxima. The h-minima (resp. maxima) detects dark (resp. bright) patterns with a intensity range of at least $h$. A constant $h$ is added (resp. subtracted) to the original image $f$. The new image with $f+h$ (resp. $f-h$) is used in an over (resp. under) geodesic reconstruction, $OverRec(f, f+h)$ (resp. $UnderRec(f, f-h)$). In effect, the algorithm erases all dark (resp. bright) patterns with an intensity range lower than $h$, retaining all other structures (flat zone with a lower/higher gray level value than its neighborhood), as illustrated in Figure 2($b$). So each local extremum in the resulting image corresponds to a local extremum in the original image with at least a dynamic range of $h$.
The main inconvenience of h-extrema is the fixed value of $h$ that is added to or subtracted from the entire image. The fixed value doesn't take into account any local information, and hence it is not optimal for our task of gold detection. Figures 2 ($b$ and $c$) illustrate this weakness. Often, gold particles close to dark areas are merged with the neighborhood, and thus erased.

### 3.2 Adaptive H-Extrema

We introduced a simple new adaptive version of h-extrema, which we refer to as *A-Extrema*, where for each pixel we adapt the value of the additive parameter $h$ according to its neighborhood.

First a filter $\mathcal{F}$ is applied on the original image $f$, in order to get a new simplified image $G_f$, smoothed and containing only (preferably) global variations. Next, in the case of a-minima, for each pixel $\mathbf{x}$ of $f$, the value added is computed according to a function $\mathcal{A}$ and the corresponding value of $\mathbf{x}$ in $G_f$, $\mathcal{A}(G_f(\mathbf{x}))$. Finally the same over reconstruction is performed, as in h-minima. The algorithm 1 enumerates all the steps:

```
Data: Image $f$, filter $\mathcal{F}$, function $\mathcal{A}$
Result: Result image $Amin_f$
begin
    $G_f \leftarrow \mathcal{F}(f)$ ;
    $Add_f \leftarrow f + \mathcal{A}(G_f)$ ; $[\Leftrightarrow \forall \mathbf{x}, Add_f(\mathbf{x}) \leftarrow f(\mathbf{x}) + \mathcal{A}(G_f(\mathbf{x}))]$ ;
    $Amin_f \leftarrow OverRec(f, Add_f)$ ;
end
```

**Algorithm 1**. A-minima algorithm.

In our application, we chose the filter $\mathcal{F}$ to be an alternate sequential filter (ASF, alternation of openings and closings with structuring elements of increasing sizes

[1]) because it simplifies the image without being affected by small local patterns. Moreover, ASF is known for its insensitivity to noise, which is useful in our application. For the ASF, we use disk type as our structuring elements, whose maximum radius equal to gold particles sizes so that we can erase them while computing the global variations.

For the function $\mathcal{A}$, we compute a percentage. Thus, in our adaptive algorithm $h$ is computed as a proportion of the global variation computed from the neighborhood of each pixel according to $\mathcal{F}$. Thus by design, in dark areas a low $h$ value is employed, whereas a high $h$ value is employed in bright areas.

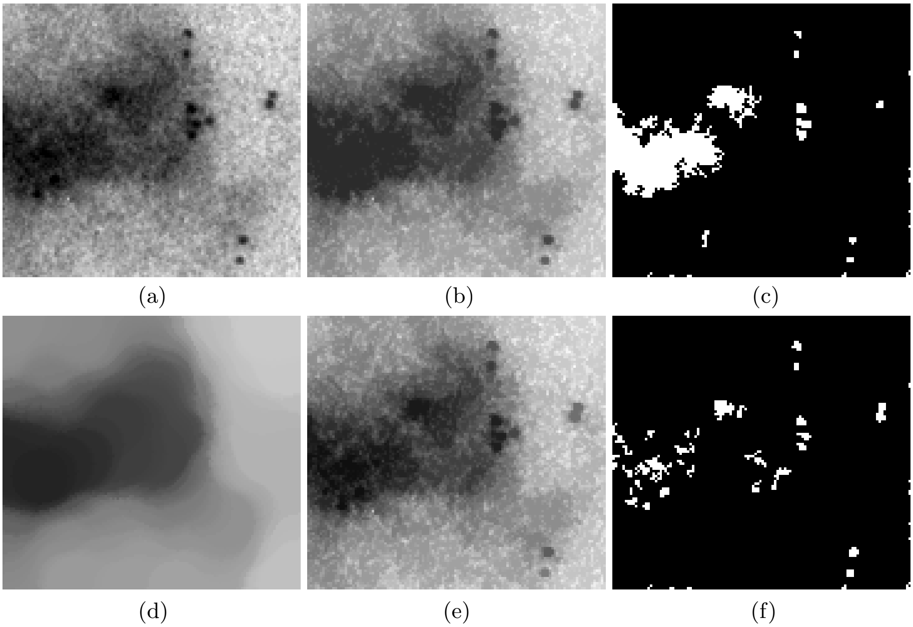


**Fig. 2.** The original image $(a)$ and the various processes: the 43-minima result $(b)$ and its local minima in white $(c)$, the ASF result $(d)$, the a-minima result for 43% of the ASF $(e)$ and its local minima in white $(f)$

The image 2 $e$ shows the A-minima result. The percentage for the function $\mathcal{A}$ needs to be computed empirically. In our example, we collected the statistics of contrasts in gold particles and found that the dynamic range was at least 43%. This is compared with $h$-minima results using the optimal value, $h = 43$. The figure illustrates the advantage of our $A - minima$ algorithm, which preserves gold particles with higher fidelity in both bright areas with high contrast and dark areas with low contrast. Thus, the combination of the filter $\mathcal{F}$ and the function $\mathcal{A}$ is effective in preserving the gold particles under different image contrasts.

**Remarks:**

- We preferred the ASF to classical mean or Gaussian filters, because it is not affected by noise and small artifacts.
- This new adaptive method can be adapted for the dual operation, A-maxima, to detect peaks.
- The function $\mathcal{A}$ can be generalized to larger class of functions. For example, in case of the detection of dark patterns in sub-exposed images, $\mathcal{A}$ can be modified as $Add_f \leftarrow f + Invert(G_f) \times p$ to be more effective.

### 3.3 Detection of Gold Particles

The A-minima preserves the sufficiently contrasted dark patterns and removes all other patterns with a lower contrast, thus provides a simplified image $Amin_f$. On the image $Amin_f$, we apply a new ASF in order to estimate the new global variations and then we compute the difference between the ASF result and the simplified image in order to extract all candidates: $C_f \leftarrow ASF(Amin_f) - Amin_f$. Each candidate is then isolated and characterized with 17 features:

- Geometrical Features: the surface and 3 radii (maximum, minimum and average). All these physical measures are "real“ values estimated according to measurements on the image and the magnification.
- Texture:
  - Intensity Features: the average, the median and the range of intensity. Note, since the candidates are identified after applying $A-minima$, their dynamic range is guaranteed to be more than the minimum specified in the algorithm. The average and median provides more information about the intensity and shape.
  - Basic moments.
  - Contrasts.
- Shape (indexes) [3]:
  - Circularity: according to radii and the inscribed disk. Gold particles are expected to have a circular shape.
  - Besicovitch symmetry: Even though the image of gold particles may be merged, they still respect symmetry along a certain axis. Candidates with no symmetry are unlikely to be gold particles.
  - Gaussian: sigma of the best fitted circular $2D$ Gaussian and the residual error of the fit. As illustrated in Figure 1, gold particles can be easily approximated using Gaussians.

For each candidate, we compute the above features and experimented with three different types of classifiers for identifying the gold particles.

- Logistic Regression [5] (RL) is a linear regression function particularly well-suited to binary classification problems, allowing a variety of complex features.
- Random Forests [6] (RF) is a powerful, state-of-the-art classifier, consisting of an ensemble of trees.

- Neural network [7] (NN) is non-linear classifier, whose parameters are learned using back-propagation to minimize the cost function such as average squared loss.

These three classifiers have very different strengths, thus we expect different types of errors. In our approach, we combine the results from all the three classifiers using majority vote.

## 4 Experiments and Results

### 4.1 EM Images

The EM images were acquired from a realistic biological experiment. SKBR3 breast cancer cells were prepared for immuno-electronmicroscopy using Tokuyasu's method as previously described in [4]. Briefly, cells were chemically fixed in 4% paraformaldehyde in PHEM buffer, washed and embedded in 12% gelatin. After solidification, cell pellets were cut in small blocks and infiltrated in 2.3M sucrose. Blocks were mounted on specimen pins, frozen in liquid nitrogen and ultra-thin $80nM$ sections were cut with Leica cryoultramicrotome. Primary antibody recognizing protein disulfide isomerase localized in endoplasmic reticulum was selected since it has been previously shown to work in immunogold labeling for TEM (e.g. shown in [8,9]) and includes incubation with bridging antibody (rabbit-anti-mouse IgG) and 5 nm protein A gold particles (from Dr. George Posthuma, UMC-Utrecht, the Netherlands), followed by contrasting in uranyloxalate and uranylasetate-methylcellulose. Imaging was performed using iCorr microscope (FEI).

### 4.2 Results

We evaluate our method using a data set of images where all the immunogold particles are manually annotated by experts. The data set consisted of 14 images, containing approximately 8500 gold particles. The evaluation was performed using a leave one out cross validation: an image is discarded from the data set, all golds from the remaining images are used to train the classifier, the discarded image is then processed and the result evaluated. The same process is performed for all the images, thus for a data set of $N$ images, each image is used 1 time for validation and $N-1$ times for training. The figure 3 *left* show the results for all the magnifications available. We can observe that our method provides particularly good sensitivity (nearly 100%) for magnifications from 1 to $\sim 3nm/pixel$, with a comparable specificity, which indicates that false alarms are minimal, and only gold particles are detected. Moreover, the area under ROC is equal to 0.9797, which demonstrates that our algorithm is effective in this task. For magnifications greater than $4nm/pixel$, the performances decrease rapidly. This is because at this magnification, a gold is represented by approximately 4 pixels, too few pixels to extract the relevant information accurately and robustly. At this resolution, they could easily be confused with noise in the image.

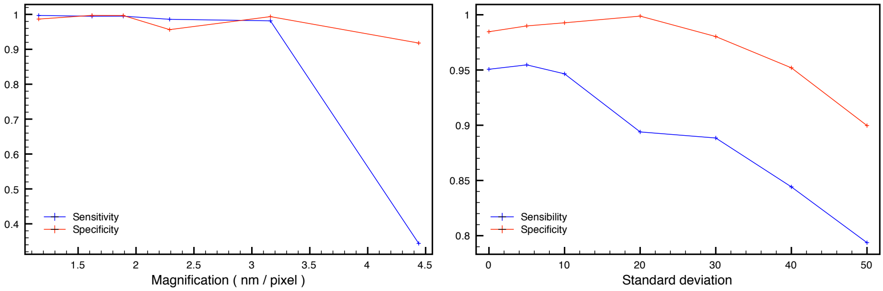
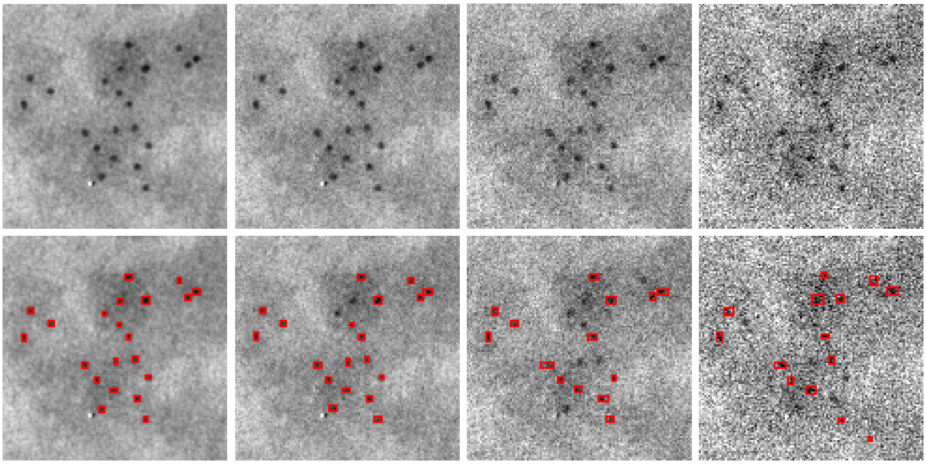


**Fig. 3.** The sensitivity/specificity obtained by the classifiers: without noise and according to the magnification (left); with artificially added noise and for a fixed magnification of 3.16$nm/pixel$ (right)



(a) No noise (b) $SD = 10$ (c) $SD = 20$ (d) $SD = 40$

**Fig. 4.** Examples of detections with additive Gaussian noise at different standard deviations

But in EM imaging, it is particularly frequent to acquire images altered with Gaussian noise. In order to evaluate the algorithm's noise sensitivity, we artificially added Gaussian noise at different levels with respect to its standard deviation. These experiments were performed on images with a magnification of 3.16$nm/pixel$, the current limit of accurate detections of gold particles. The performance of our algorithm is not significantly altered until the noise reaches a standard deviation of 10. Figure 3 (*right*) illustrates that from 10 to 20 the sensitivity decrease but not the specificity in spite of the fact that images are already extremely degraded[1] (see 4). At higher standard deviations, the specificity starts to decrease and the number of false positive increases.

[1] Experts do not analyze images with such quantity of noise.

## 5 Conclusions

In this paper, we developed a complete pipeline for automatically detecting immunogold particles. First we introduced a new adaptive version of h-extrema which filters contrasted patterns according to their dynamic and neighborhood: it preserves contrasted patterns even in really low contrasted neighborhood. Then this new method was successfully applied to simplify the images and find all potential candidates. Each candidate was then classified using machine learning algorithms. The results on a data set of annotated images show that our method detects immunogold particles in EM images with high accuracy, both high sensitivity and specificity, even in highly noisy images.

**Acknowledgments.** This work was funded by Keck Foundation and NSF award 1027834. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NSF or Keck Foundation.

## References

1. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, London (1982)
2. Grimaud, M.: New measure of contrast: the dynamics. In: Paul, D., Gader; Edward, R., Dougherty; Jean, C. (eds.) Proc. SPIE, Image Algebra and Morphological Image Processing III, vol. 1769, pp. 292–305 (1992)
3. Thibault, G., Fertil, B., Navarro, C., Pereira, S., Cau, P., Levy, N., Sequeira, J., Mari, J.-L.: Shape and Texture Indexes: Application to Cell Nuclei Classification. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) 27(1) (2013)
4. Jan, W., Slot, H.J.: Geuze: Cryosectioning and immunolabeling. Nature 10, 2480–2491 (2007)
5. Berkson, J.: Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association 39 (1944)
6. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
7. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 115–133 (1944)
8. Karreman, M.A., van Donselaar, E.G., Gerritsen, H.C., Verrips, C.T., Verkleij, A.J.: VIS2FIX: A high-speed fixation method for immuno-electron microscopy. Traffic 12(7), 806–814 (2011)
9. Karreman, M.A., Agronskaia, A.V., van Donselaar, E.G., Vocking, K., Fereidouni, F., Humbel, B.M., Verrips, C.T., Verkleij, A.J., Gerritsen, H.C.: Optimizing immuno-labeling for correlative fluorescence and electron microscopy on a single specimen. Journal of Structural Biology (2012)

# Improving Dysarthria Classification by Pattern Recognition Techniques Based on a Bionic Model

Eduardo Gonzalez-Moreira, Diana Torres, Carlos A. Ferrer, and Yusely Ruiz

Center for Studies on Electronics and Information Technologies,
Central University of Las Villas, Cuba
{moreira,dtb,cferrer,yuselyr}@uclv.edu.cu

**Abstract.** The goal of this research is to use a bionic model to enhance classification of Dysarthria. The model based on the main features of the mammalian olfactory system is the initial stage of the recognition process. The bionic model aimed to achieve an enhancement in the separation ability of the dysarthric features. The recognition performance obtained by four different pattern recognition algorithms using the bionic model to improve the features is shown and discussed. The results indicated that bionic model had clear influence on classification performance of well-known techniques using dysarthria database as case study. We regard the results of this study as a promising initial step to the use of bionic model as a recognition improvement function.

**Keywords:** pattern recognition, bionic model, dysarthria.

## 1 Introduction

Nowadays one of the persistent problems in the study of pattern recognition is the efficient description of relevant features and the best selection of the artificial intelligent network to the classification tasks. Neural network theory is an old research topic that has been widely used in recent years. Computational power developments and mathematics of complexity have made the field succeed during the last years, and made a significant approach to simulate complex biological systems.

As one of the most important sensory modalities in the sensory systems of mammals, olfactory nervous systems have attracted many researchers during the last years. Some models have been developed to emulate the functions of olfactory nervous systems [1, 2]. The olfactory nervous system is relatively simple and well-known functionally and morphologically, and is an interesting system for understanding the cognitive processes performed in the brain. Even when some brain processes remains unclear, many aspects of olfaction, such as the mechanisms of reception and central processing or the nature of the stimuli have been fairly extensively studied [3, 4].

In helping to understand the olfactory information processing many mathematical models that mimic the main feature of the olfactory system have been applied to pattern recognition, often with remarkable results [5–7]. Therefore this is the approach followed in this research, where a novel methodology for dysarthria classification is designed and implemented.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 246–253, 2013.

Dysarthria is a term associated with a group of neurological diseases caused by lesions in the peripheral or central nervous system. Different speech perturbations are associated to the type and location of the lesions, which are correlated with both: the kind of dysarthria and the brain damage. The speech is one of the mechanisms that are more sensitive to lesions in the nervous system due to the precise coordination and timing required for normal speech production. Therefore, the study of the speech in patients suffering this pathology can reveal important information for assessment and treatment, increasing the reliability and effectiveness of the diagnosis process.

The goal of the present research is to introduce a bionic model, inspired on the mammalian olfactory system, to improve the classification process through enhancement of the data under analysis. The bionic model is formed from a bulbar model which mimics the behavior of excitatory mitral and inhibitory granule cells, and a 3-layered cortical model, which emulates the structure and behavior of the piriform cortex [8, 9]. Unsupervised bionic model allows for the weightings of the input acoustic measures to be determined on the basis of the inherent nonlinear regularities of the input data space. Unsupervised learning is particularly advantageous when no gold standard exists for classification. It is hoped that by using a bionic model, which is trained using few samples, will provide an upgrading to the data resulting in a better classification result.

## 2 Bionic Model

The bionic model emulates the main structural features of the olfactory system mimicking two main parts, bulb and cortex, and how they are connected via feedforward and feedback channels (**Fig. 1**). The researchers look for a balance between the wish for realism when comparing theoretical and computational results with experiment, and the need for abstraction and simplification of the biological complexity for a mathematical analysis and computer simulation.

The olfactory bulb can be viewed as the first central olfactory relay station extracting specific stimulus features, a function characteristic of the primary sensory areas in the brain [10, 11]. The cellular structure of the bulb is well established and in this work, the olfactory bulb was modeled using a simple approximation of excitatory mitral and inhibitory granule cells. The activity of mitral cells was spatially distributed such that odorants were represented in the bulb model by a distributed pattern of mitral cell activity [12, 13]. Mitral cells adjacent to each other project to the same or neighboring glomerulus. Among these models, the dynamics of every neural ensemble is described using a second order differential equation (Eq. (1)), based on physiological experiments of the olfactory system [14]:

$$\frac{1}{ab}\left[x_i''(t) + (a+b)x_i'(t) + abx_i(t)\right] = \sum_{j\neq i}^{N}\left[W_{ij}q\left(x_j(t), g_j\right)\right] + I_i(t) + \varepsilon(t) \qquad (1)$$

Here $i = 1, \ldots, N$, where $N$ is the number of channels, $x_i(t)$ indicates the state variable of $i$th neural ensemble, $x_j(t)$ represent the state variable of $j$th neural ensemble, which is connected to the $i$th, $W_{ij}$ represents the connection strength between them. $I_i(t)$ is an input function which stands for the external input to the $i$th channel. $\varepsilon(t)$ is noise or spontaneous neural activity. The parameters *a* and *b* reflect two rate constants. $q(x_j(t),g_j)$ is a static nonlinear sigmoid function derived from the Freeman model [1] and *g* represents the maximum asymptote of the sigmoid function, also experimentally obtained from biological trials. However, the exact form of these relations is not essential to the system behavior, as long as the shape is qualitatively conserved. Since granule cells do not have axons, they are modeled using a larger linear range, and thus a less strong nonlinear threshold effect than for mitral cells.

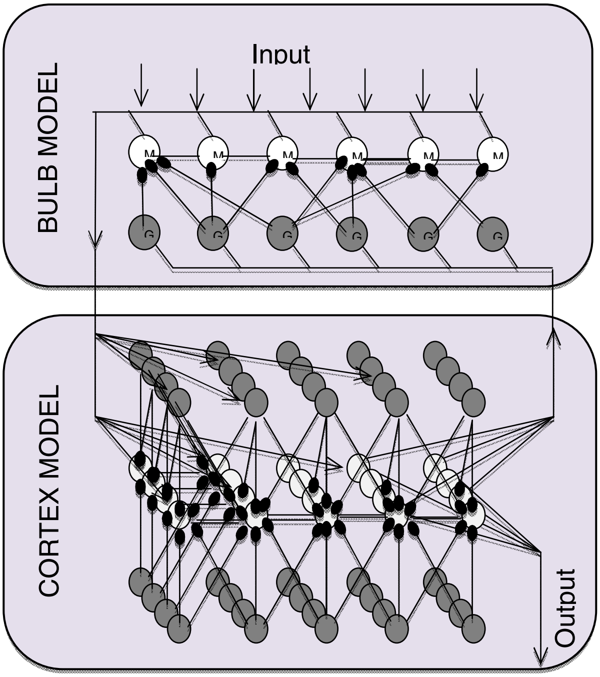


**Fig. 1.** Bionic model structure consists of two main parts (bulb model and cortex model) and their connections

Basically, the piriform cortex structures belonging to the allocortex are thinner and structurally less complex (having three cortical layers) than the neocortex [15, 16]. The architecture of the cortical model was based on the 3-layered structure of the olfactory piriform cortex, using similar network connectivity, but relatively simple model nodes, representing populations of neurons. The two sets of inhibitory nodes have two different time constants and slightly different connectivity to the excitatory nodes. All connections were modeled with distance dependent time delays for signal propagation, corresponding to the geometry and fiber characteristics of the real cortex.

Similar to previously described for bulb model, the time evolution for a cortical network of *N* neural ensemble is given by a set of coupled nonlinear first order differential equations for all internal states, *u*. With external input, coming from the bulb, *I(t)*, noise or spontaneous neural activity *ε(t)*, characteristic time constant *c*, and connection weights $W_{ij}$ between units $i$ and $j$, separated with a time delay $\delta_{ij}$, the neural activity for each ensemble, $u_i$, is given by:

$$\frac{1}{c}[cu_i'(t) + u_i(t)] = \sum_{j\neq i}^{N}\left[W_{ij}q_j\left(u_j(t-\delta_{ij})\right)\right] + I_i(t) + \varepsilon(t) \qquad (2)$$

The continuous sigmoid function, $q_j(u_j)$, represents the input-output transfer function, experimentally determined by Freeman [1]. With gain parameter $g_j$, determines the slope, threshold, and amplitude of the curve for $j$th ensemble and $D$ a normalization constant, $q_j$, is described by:

$$q_j = Dg_j\left\{1 - exp\left[\frac{-exp(u_j)-1}{g_j}\right]\right\}) \qquad (3)$$

It has been previously shown that the model displays major characteristics of the olfactory cortex dynamics. A typical EEG (encephalogram) time series from the cat olfactory cortex is shown in **Fig. 2**, together with a simulated EEG trace using the current model. Oscillatory and aperiodic dynamic behavior was shown to improve the performance by reducing the recall (convergence) time in associative memory tasks.

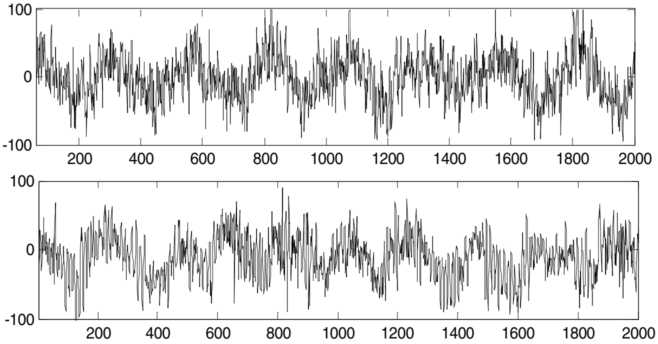


**Fig. 2**. Simulated (top) and real (bottom) EEG, showing the complex dynamics of cortical structures. The real trace is from cat olfactory bulb (data courtesy of Walter J. Freeman), whereas the simulated trace is from a simulation with the current cortical model. The x-axis shows milliseconds, and the y-axis is in microvolts.

While time constants, signal velocities, and other system parameters are determined by physiological constraints, the connection weights should be adjusted properly for the best performance of the model. The bionic model presents two learning processes: Hebbian associative learning and habituation. These learning processes exist in a subtle balance and their relative importance changes at various stages of the memory process. The memory basins and attractors are formed via Hebbian learning under reinforcement, while the impact of environment noise, which includes the background inputs without any information, is reduced by habituation.

The learning processes are applied to the bulb model and to the cortex model. According to modified Hebbian rule, each pair of nodes co-activated by the stimulus have their lateral connections strengthened. The nodes with activities larger than the mean on the layer are considered as activated ones and strengthened with the Hebbian coefficient. In contrast, those with activity less than the mean are not considered to be activated ones and these connections are decreased by the Habituation coefficient and the simulation period. A bias coefficient is defined in the modified Hebbian learning rule to avoid the weight space saturation. These processes are applied in bulbar mitral layer and middle cortical layer.

## 3 Case Study

This study was carried out using speech datasets that contain records from eight types of dysarthria among which are: Spastic Dysarthria, Flaccid Dysarthria, Ataxic Dysarthria, Hyperkinetic Dysarthria (organic voice tremor, chorea and dystonia), Hypokinetic Dysarthria (Parkinson disease) and Mixed Dysarthria (Amyotrophic Lateral Sclerosis). For the particular case study analyzed in this research each kind of dysarthria defines one class in the classification task and comprehend an average of 14 subjects taken from 2 databases corresponding with different levels of the severity of the illnesses. The kind of severity of the sickness is annotated in the databases, where a total of 38, both perceptual and acoustic features, are also given. The first dataset was created including 62 patients [17]. The second dataset was a selection of dysarthric speakers from a database collected by Aronson and colleagues [18]. A total of 14 normal subjects were used as control to contrast the differences between the pathologic and normal speech.

The perceptual and acoustical measures contained in both databases were obtained from 3 utterances that provide more information about these diseases with less computational and storage requirements [19]. The utterances consisted of the sustained phonation of the vowel /a/, the repetition of the syllable /PA/TA/KA/ and the reading of the passage: 'The Grandfather' [20].

Therefore a total of 127 samples were analyzed to classify subjects into nine distinct groups, eight dysarthric groups and a control group. From a total of 38 features, 36 (25 perceptual and 11 acoustic) were selected to provide information about the condition of the speech mechanism of dysarthric patients [21]. However, a multidimensional analysis of dysarthric speech revealed that not all of the 36 features provided valuable information about the dysarthric groups [17].

From this point of view, some linear analysis techniques (clustering of variables, linear discriminant analysis, best first) were applied in order to obtain a reduction in the number of features of the data [17, 22]. The resultant number of features was set to 20, 16 and 12 after dimensionality reduction, following a commonly used rule in pattern recognition that states a ratio of at least 10 cases per input observation. Finally four versions of the original databases were created with 36, 20, 16 and 12 features respectively.

## 4 Results

In this paper, we preprocess the four databases with the bionic model prior to the classification stage (**Fig. 3**). The bionic model transforms input features according to its nonlinear dynamics. This means that original features entering the mitral layer are modified inside the bionic model, and delivered by the middle cortical layer to the next stage.

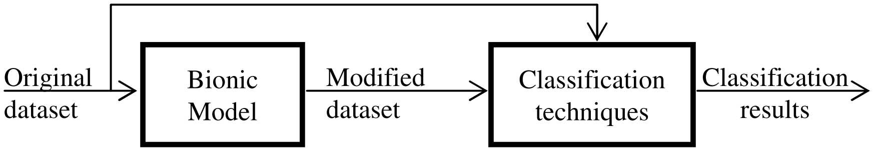


**Fig. 3.** Experimental setup for evaluating the influence of the bionic model performance in the final classification results using dysarthria databases

In the final classification stage four well-known techniques like Support Vector Machines (SVM), Bayesian Network (BN), decision trees (J48) and Naive Bayes (NB) were applied. These machine learning algorithms are included into the data mining software WEKA [23]. The databases are randomly reordered and then split into $n$ folds of equal size. The leave-one-out cross-validation method is applied. In our case $n$ is equal to the number of examples ($n$ = 127), and in each iteration of the cross-validation method, one fold is used for testing and the other $n$-1 folds is used for training the classifier.

**Table 1.** Percent of correct clasification of various methods with original and modified datasets, along with their corresponding MSSS as an improvement index

| Methods | Dataset dimensions | | | |
|---|---|---|---|---|
| | 12 | 16 | 20 | 36 |
| SVM + original data | 79.50 | 79.10 | 81.20 | 80.30 |
| SVM + modified data | 81.10 | 79.50 | 79.50 | 82.80 |
| **MSSS (%)** | **7.80** | **1.91** | **-9.04** | **12.69** |
| J48 + original data | 66.10 | 66.00 | 68.50 | 62.20 |
| J48 + modified data | 72.40 | 72.50 | 70.10 | 68.50 |
| **MSSS (%)** | **18.58** | **19.12** | **5.08** | **16.67** |
| BN + original data | 80.20 | 81.10 | 81.80 | 80.10 |
| BN + modified data | 83.50 | 85.50 | 81.10 | 83.30 |
| **MSSS (%)** | **16.67** | **23.28** | **-3.85** | **16.08** |
| NB + original data | 70.00 | 71.60 | 66.10 | 64.90 |
| NB + modified data | 73.20 | 72.90 | 68.50 | 65.40 |
| **MSSS (%)** | **10.67** | **4.58** | **7.08** | **1.42** |

The four datasets containing objective and perceptual judgments were evaluated with respect to the percentage of correct classification provided by each classifier using bionic model to pre-process the data and without using it. Moreover a skill score based on mean squared error (MSSS) is used as an improvement index in order to evaluate the impact of the bionic model in the final classification results. MSSS is defined as one minus the ratio of the squared error for the classification with modified datasets to the squared error for the classification with original datasets. The result of the assessment process revealed that the bionic model allowed an improved classification rate over the original databases in almost every trial, as shown in **Table 1**.

These results show that the bionic model implementation is appropriate to enhance the classification of the dysarthric groups studied, providing a better percentage of correct classification in almost every trial. In only two cases the bionic models worsened classification results. Particularly two of the four algorithms take significant advantage from the modified features introduced by the bionic model in the initial state of the data processing, J48 and BN. In general, modified dataset based on the bionic model outperforms the performance of almost every classification techniques.

## 5 Conclusions

In this work, a bionic model mimicking the main features of the olfactory system has been analyzed, and its performance to improve dysarthric classification was shown. Our bionic model is constructed from two principal parts: a bulb model and a cortex model. The bulb model is composed of mitral and granule cells, whereas the cortical model mimics the 3-layered structure of the mammalian piriform cortex.

The analyses performed have revealed that the model has the capacity to learn complex patterns due to Hebbian modification of the connection strengths of the M-sets excitatory synapses in the bulb and the excitatory units on the cortical middle layer. The improvement in the classification of the dysarthric databases was shown.

The current digital computers are the bottleneck for the models based on biological sensorial systems due to the time required to solve ODE by numerical integration. Nevertheless, even when considering the memory requirements and the computational time demanded, the bionic model still cannot compete against conventional classification methods. The results obtained with this research have become a promising initial step into the use of the bionic model to improve other patter recognition tasks.

**Acknowledgment.** The authors would like to thank Julian Cardenas Barrera, Guang Li, Hans Liljenström and Walter J. Freeman for their support to this research.

## References

1. Freeman, W.J.: Mass action in the nervous system. Academic Press, New York (1975)
2. Shepherd, G., Brayton, R.: Computer simulation of a dendrodendritic synaptic circuit for self- and lateral-inhibition in the olfactory bulb. Brain Res. 175, 377–382 (1979)
3. Freeman, W.J., Skarda, C.: Spatial EEG patterns, non-linear dynamics and perception: the neo-sherringtonian view. Brain Res. Rev. 10, 147–175 (1985)
4. Aronsson, P., Liljenstrom, H.: Effects of non-synaptic neuronal interaction in cortex on synchronization and learning. Biosystems 63, 43–56 (2001)

5. Liljenström, H.: Stability and instability in autonomous systems. In: Advances in Cognitive Neurodynamics – Proceedings of the International Conference on Cognitive Neurodynamics, Shanghai, November 17-21, pp. 661–665 (2008)
6. Kozma, R., Freeman, W.J.: The KIV Model of Intentional Dynamics and Decision Making. Neural Networks 22(3), 277–285 (2009)
7. Gonzalez-Moreira, E., Li, G., Ruiz, Y.: A tea classification method based on an olfactory system model. In: Advances in Cognitive Neurodynamics, Hangzhou, pp. 747–751 (2008)
8. Gonzalez-Moreira, E., Liljenstrom, H., Ruiz, Y., Li, G.: A biological inspired model for pattern recognition. J. Zhejiang Univ. Sci. B 11, 115–126 (2010)
9. Gonzalez-Moreira, E., Ruiz, Y.: A bionic model inspired on the olfactory system. In: V Latin American Congress on Biomedical Engineering (CLAIB 2011), Havana (2011)
10. Doty, R.: Handbook of olfaction and gustation. Dekker, New York (2003)
11. Lowe, G.: Electrical signaling in the olfactory bulb. Curr. Opin. Neurobiol. 13, 476–481 (2003)
12. Mori, K., Nagao, H., Yoshihara, Y.: The olfactory bulb: coding and processing of odor molecule information. Science 286, 711–715 (1999)
13. Leon, M., Johnson, B.: Olfactory coding in the mammalian olfactory bulb. Brain Res. Rev. 42, 23–32 (2003)
14. Freeman, W.J.: Nonlinear gain mediating cortical stimulus-response relations. Biol. Cybernet. 33, 237–247 (1979)
15. Liljenstrom, H.: Modeling the dynamics of olfactory cortex using simplified network units and realistic architecture. Int. J. Neural Syst. 2, 1–15 (1991)
16. Kowianski, P., Lipowska, M., Morys, J.: The piriform cortex and the endopiriform nucleus in the rat reveal generally similar pattern of connections. Folia Morphol. 58, 9–19 (1999)
17. Castillo Guerra, E.: A modern approach to dysarthria classification. Ph.D. Thesis at UNB, Canada (2002)
18. Aronson, A.E.: Dysarthrias: differential diagnosis (1993)
19. Darley, F.L., Aronson, A.E., Brown, J.R.: Differential diagnostic patterns of dysarthria. J. Speech Hear Res. 12, 246–249 (1969)
20. Darley, F.L., Aronson, A.E., Brown, J.R.: Motor speech disorders, 3rd edn. W.B. Saunders Company, Philadelphia (1975)
21. Castillo Guerra, E., Lovey, D.F.: A modern approach to dysarthria classification. In: 25th Annual lntemational Conference of the IEEE EMBS, Cancun (2003)
22. Torres, D., Acebo, L., Gonzalez-Moreira, E., Ferrer, C.A.: Uso de la señal de voz en la clasificación de disartrias. Presented at the RECPAT 2012 (2012)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. Sigkdd Explor. 11 (2009)

# A Comparison of Different Classifiers Architectures for Electrocardiogram Artefacts Recognition

Carlos R. Vázquez-Seisdedos[1], Alexander A. Suárez-León[1],
and Joao Evangelista-Neto[2]

[1] Center for Neurosciences Studies, Image and Signal Processing,
Biomedical Engineering Departament, Universidad de Oriente, Santiago de Cuba
{cvazquez,aasl}@fie.uo.edu.cu
[2] Amazon State University and UniNorte/Laurente Manaus, Brazil
joao_evangelista_neto@yahoo.com

**Abstract.** Applying heart rate variability (HRV) analysis on ambulatory ECG monitoring is a very useful decision support tool for cardiovascular diagnosis. The presence of non-valid beats (artefacts) on the RR interval time-series affects the diagnosis accuracy using this technique. Despite the importance of artefacts recognition prior to exclusion, no paper was found characterizing quantitatively the performance of, on the one hand, the extracted features and, on the other hand, the clustering methods on artefacts recognition for HRV analysis. In this paper we evaluate the performance of several combinations of three feature extraction methods and four clustering methods (based on machine learning techniques) for the artefacts beats recognition on the ECG signal. The trade-off between performance indexes suggests the use of a non-linear principal component analysis as feature extraction method and a multilayer perceptron (MLP) as clustering method, with sensitivity, specificity and positive-predictive-value (PPV) equal to respectively 95 %, 95.9 % and 98 %.

**Keywords:** ECG, artefact detection, artificial neural networks, feature extraction, classifier.

## 1 Introduction

The ambulatory monitoring of electrocardiogram (ECG) during daily activities plays an important role in the diagnosis but presents the challenge of information loss due to the occurrence of technical and physiological artefacts that distort the ECG signal. Typically, more than 80 000 heartbeats per channel are recorded during 24 hours; so many computer-based methods for automatic ECG analysis have been studied for a long time. ECG recognition is a difficult problem even with the aid of a computer, because ECG waveforms may differ significantly even for the same beat type taken from the same patient. The architecture for morphological recognition of beats in ECG includes several stages as showed in figure 1. The core is the classifier composed by features extraction (FE) and clustering stages, both based on computational intelligence techniques.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 254–261, 2013.

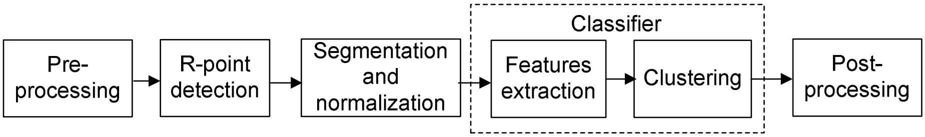


**Fig. 1.** Stages of an ECG-signal classifier system

Some FE methods on ECG are based on:

1. Morphologic features extracted from signal [1], [2]: amplitudes, interval durations or areas of waves or specific segments.
2. Statistical parameters in time domain (mean, standard deviation, maximum, minimum, self-correlation-coefficients, histogram, etc) as well as in frequency domain (QRS-complex-energy, power spectral density).
3. The use of mathematical models to represent ECG wave and segments [3], like autoregressive models, linear prediction coefficients and curve fitting.
4. The use of transforms: (a) Principal Component Analysis (PCA) [4], (b) Discrete Cosine Transform (DCT) [5], (c) Wavelet Transform [3], (d) Time-frequency distributions and (d) Hermite functions, among others.

Several artificial neural networks (ANN) based clustering methods that automatically classify heart beats have been proposed in the last years. Multilayer Perceptron (MLP) is one of the most referred [1]. Other clustering methods (in descending order) are support vector machine (SVM), learning vector quantization (LVQ) and radial basis functions (RBF).

In order to validate the HRV analysis [6], it should be verified that each detected R-point corresponds to a complete beat resulting from sinus node depolarization without any type of atrioventricular blockade. Otherwise, the beat will be considered as an artefact located in the corresponding positions of the RR time series, and it should be excluded of the analysis. The heart beat artefacts can have either a physiological (e.g. arrhythmias) or a technical (e.g. spikes and noise) origin.

Although there is an extensive diversity of publications about arrhythmia recognition [7-9], no publication was found characterizing quantitatively the performance of the FE and clustering methods for recognition of heart beat artefacts. There is a recent work [10] that compares several FE methods according to simplicity, accuracy and positive predictive value, but only on the qualitative point of view and not including artefacts beats. This paper does not analyze the execution time or other indexes, neither others FE methods as popular as DCT and linear PCA.

In a previous work [11], three FE methods were characterized, using an MLP network as a gold standard for clustering. The higher performance corresponded to nonlinear PCA, also named kernel PCA (KPCA). The previous research left the following question: will another cluster method exist with a better performance?

The aim of this work is to validate the performance of three FE methods combined with four clustering approaches to detect non-valid beats (artefacts beats) for HRV analysis.

# 2 Methodology

## 2.1 Data

The MIT-BIH arrhythmia database is used for training and validation. This database consists of 48 30-min two-lead recordings (series 100 and 200) sampled at 360 Hz, for a total of 24 hours [12]. The development platform was MATLAB 7.7.

The beat classes' global distribution from this database has 110288 beats: 75056 normal beats and 35232 artefact beats. Thus, around the 70% of the beats were classified as normal beats (resulting from sinus node). There are 17 classes of beats (ectopic, left and right bundle blocks, and others) that are grouped in a class: artefact (ARTF). Every normal beat belongs to the normal (NORM) class.

Initially, a partial evaluation using 4000 beats (2000 for training and 2000 for validation) was made in order to find the best clustering method for this sample. Then, a global evaluation was made for the entire database using the clustering method found. Of the 4000 beats belonging to different database records, were chosen 2000 for each class according to the following criteria: from each record with more than 50 beats of NORM class (40 records), 50 beats were randomly chosen. For the beats of the ARTF class the criterion is the following: from each record with more than 375 beats (25 records) from ARTF class, 80 beats were randomly chosen.

## 2.2 Stages of the Classification System Beats

**Preprocessing:** To eliminate baseline drift and high frequency noise, a bandpass filter is used, consisting of a high-pass filter (Butterworth, zero-phase, 6th order, cutoff frequency equal to 0.6 Hz) in cascade with a low-pass filter (Butterworth, zero-phase, 12th order, cutoff frequency equal to 45 Hz). Subsequently, the average value was eliminated so that the signal is converted into a signal of unit variance. This standardization is performed to achieve invariance with respect to the amplitude, for any beat.

**R Peak Detection:** Because the R peak detection has been broadly described, no further discussion on this subject is pursued in this paper. In [13] there is an extensive review of recent approaches for R peak detection. Any R peak detector with demonstrated robustness can be used. In this work, R peak annotations were used for each beat, which is equivalent to employing an infallible algorithm to estimate the R peaks on the ECG signal. Thus, the results depend only on the clustering methods and not on the R peak detection approach.

**Segmentation:** An asymmetric window of fixed size around the R peak was used. The length of the window was equal to 235 samples (i.e., the maximum value) including the R peak point. For the selection of the number of samples to the right and to the left around the R peak, the mean of the PR/QT is used for minimal and maxima values, yielding to 39 % @ 360 Hz, resulting in approximately 39% of the samples located on the left (92 samples) and 61% on the right (142 samples).

**Classifiers:** The classifier consist on the combination of three FE algorithms: DCT, PCA and KPCA with four types of machine learning techniques (MLP, LVQ, RBF, and SVM). From FE stage, it is possible to obtain the following number of components:

| DCT | PCA | KPCA |
|---|---|---|
| $1 \le K \le V_L$ | $1 \le K \le V_L$ | $1 \le K \le M$ |

Where, $V_L$ is the vector length (in samples), $M$ is the vector number and $K$ is the number of components for FE stage. In this case, $V_L$ is equal to 235, $M$ is equal to 2000, and $K$ is equal to 10, 15 and 20 components (generating 12 classifiers in total).

To train the MLP classifiers, a network was created with hidden layer architecture: $n$ - $2n$ - 1, i.e., $n$ input neurons, $2n$ neurons in the hidden layer and one output neuron (architectures: 10 - 20 - 1, 15 - 30 - 1 and 20 - 40 – 1). The activation functions are hyperbolic tangent and linear in the hidden and output layers, respectively.

For the LVQ classifier, $n$ neurons in the input layer and two neurons in the output layer (2 classes) are employed while for the RBF classifier uses a simple algorithm to search the optimal dispersion parameter in the range (0.1 - 10). Each classifier has $n$ radial basis neurons and one linear output neuron.

The SVM classifier is the only one for which the number of beats ($N$) is reduced to 1000 due to high computational cost. Thus, one might expect a lower performance, because only a quarter of the available set is employed.

The performance evaluation for each classifier was carried out by computing three indexes: Specificity (Sp), Sensitivity (Se) and Positive Predictive Value (PPV), from the confusion table defined for the NORM and ARTF classes (Figure 3):

**Estimated Classes**

|  |  |  |
|---|---|---|
| **NORM** | TP | FP |
| **ARTF** | FN | TN |
|  | **NORM** | **ARTF** |

**Real classes**

$$Sp(\%) = \frac{TN}{TN + FP} \times 100 \qquad Se(\%) = \frac{TP}{TP + FN} \times 100, \qquad PPV(\%) = \frac{TP}{TP + FP} \times 100$$

**Fig. 2.** Confusion matrix for each classifier. TP: true positive, TN: true negative, FP: False positive, FN: false negative, Se: Sensitivity, Sp: Specificity, PPV: positive predictive value.

### 2.3 Validation and Comparation of Classification Methods

To test whether the differences between classifiers are statistically significant, we used the McNemar's Test, based on the calculation of McNemar statistic, defined as

$$z = \frac{|n_{01} - n_{10}| - 1}{\sqrt{n_{01} + n_{10}}} \tag{1}$$

Where:
$n_{01}$: number of miss-classified samples by A but not by B.
$n_{10}$: number of miss-classified by B but not by A.

The statistic $|z|$ was calculated for all possible combinations of each pair of classifiers. The null hypothesis H0 (the classifiers have the same error) can be rejected with an error probability of 0.05 if $|z| > 1.96$. The alternative hypothesis H1 is that the classifiers have different errors, i.e., the differences in performance indexes are statistically significant.

# 3 Results and Discussion

## 3.1 Partial Evaluation

Tables 1 to 6 show the performance indexes of the four classifiers with 10, 15 and 20 components (features) both for training and validation.

**Table 1.** All Classifiers using DCT (Training)

| **Training** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp* (%)** | | | ***Se* (%)** | | | ***PPV* (%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| DCT+ MLP | 97.8 | 97.1 | 97.4 | 97.7 | 97.8 | 97.6 | 97.8 | 97.1 | 93.2 |
| DCT+LVQ | 75.8 | 83.6 | 79.4 | 91.6 | 79.5 | 95.6 | 79.5 | 84.8 | 76.2 |
| DCT+RBF | 87.2 | 87.6 | 90.8 | 86.6 | 87.3 | 93.6 | 87.3 | 88.0 | 84.0 |
| DCT+SVM | 72.8 | 80.9 | 82.0 | 80.1 | 76.0 | 81.7 | 76.0 | 82.3 | 73.8 |

**Table 2.** All Classifiers using DCT (Validation)

| **Validation** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp* (%)** | | | ***Se* (%)** | | | ***PPV* (%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| DCT+ MLP | 93.3 | 93.2 | 94.6 | 94.2 | 96.3 | 96.6 | 93.2 | 93.3 | 94.6 |
| DCT+LVQ | 71.8 | 80.8 | 75.2 | 91.7 | 88.9 | 96.2 | 76.2 | 82.0 | 79.2 |
| DCT+RBF | 83.9 | 82.9 | 89.9 | 85.9 | 89.7 | 92.5 | 84.0 | 83.8 | 90.0 |
| DCT+SVM | 72.5 | 77.5 | 79.8 | 79.6 | 81.2 | 81.0 | 73.8 | 77.9 | 79.6 |

**Table 3.** All Classifiers using PCA (Training)

| **Training** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp* (%)** | | | ***Se* (%)** | | | ***PPV* (%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| PCA+ MLP | 96.9 | 98.6 | 98.5 | 97.8 | 97.7 | 97.4 | 97.0 | 98.6 | 98.5 |
| PCA+LVQ | 80.2 | 79.0 | 84.7 | 93.8 | 93.1 | 95.4 | 82.8 | 81.3 | 86.4 |
| PCA+RBF | 90.1 | 90.4 | 93.5 | 89.4 | 90.5 | 93.0 | 90.2 | 90.6 | 93.6 |
| PCA+SVM | 82.8 | 84.4 | 84.4 | 79.2 | 84.6 | 85.9 | 83.2 | 85.4 | 85.6 |

**Table 4.** All Classifiers using PCA (Validation)

| Validation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp*** **(%)** | | | ***Se*** **(%)** | | | ***PPV*** **(%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| PCA+ MLP | 94.2 | 95.9 | 95.9 | 94.5 | 96.2 | 94.9 | 94.2 | 95.9 | 95.8 |
| PCA+LVQ | 78.1 | 81.9 | 82.2 | 91.3 | 94.1 | 93.8 | 80.4 | 84.1 | 83.9 |
| PCA+RBF | 89.2 | 88.9 | 92.3 | 88.5 | 88.5 | 91.5 | 89.0 | 88.7 | 92.1 |
| PCA+SVM | 81.4 | 82.5 | 82.1 | 81.3 | 84.7 | 84.2 | 81.0 | 82,6 | 82.2 |

**Table 5.** All Classifiers using KPCA (Training)

| Training | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp*** **(%)** | | | ***Se*** **(%)** | | | ***PPV*** **(%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| KPCA+MLP | 96.8 | 98.4 | 98.3 | 98.0 | 98.6 | 98.8 | 96.9 | 98.4 | 98.3 |
| KPCA+LVQ | 84.4 | 84.2 | 93.9 | 89.1 | 93.1 | 90.0 | 85.3 | 85.7 | 90.5 |
| KPCA+RBF | 87.6 | 91.8 | 90.1 | 88.8 | 88.6 | 94.9 | 87.9 | 91.7 | 90.7 |
| KPCA+SVM | 79.3 | 89.8 | 91.1 | 84.4 | 86.7 | 87.5 | 81.4 | 90.2 | 91.3 |

**Table 6.** All Classifiers using KPCA (Validation)

| Validation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | ***Sp*** **(%)** | | | ***Se*** **(%)** | | | ***PPV*** **(%)** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| KPCA+MLP | 94.0 | 96.8 | 96.0 | 94.9 | 96.3 | 95.9 | 94.0 | 96.8 | 96.0 |
| KPCA+LVQ | 83.0 | 82.9 | 89.5 | 88.0 | 90.4 | 91.9 | 83.6 | 83.9 | 89.6 |
| KPCA+RBF | 87.0 | 91.2 | 88.6 | 88.3 | 88.5 | 92.7 | 87.0 | 90.8 | 88.9 |
| KPCA+SVM | 79.8 | 88.6 | 89.2 | 84.5 | 87.4 | 89.0 | 80.4 | 88.2 | 88.9 |

From the above results, it is evident that the MLP classifier gives the best results for all features extraction variants.

### 3.2 Global Evaluation

Table 7 shows the results of the evaluation for the entire database (110192 beats). Only, 96 beats were excluded: the first and the last of each record. It is evident that for 10 components in the feature vector, the PCA + MLP method outperforms in specificity and positive predictive value to the other two methods, although it is less sensitive than the DCT + MLP and KPCA + MLP, in that order. For 15 and 20 components, the KPCA + MLP method is better than DCT + MLP and PCA + MLP, showing indexes greater than (or equal to) the best shown by the other two methods.

**Table 7.** All Classifiers using MLP

| Validation | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Classifier** | **DCT** | | | **PCA** | | | **KPCA** | | |
| | **10** | **15** | **20** | **10** | **15** | **20** | **10** | **15** | **20** |
| ***Sp*** **(%)** | 93.9 | 95.4 | 96.2 | 95.4 | 95.9 | 96.7 | 94.6 | 96.4 | 96.7 |
| ***Se*** **(%)** | 94.6 | 95.2 | 94.7 | 93.9 | 94.8 | 93.9 | 94.4 | 95.2 | 95.4 |
| ***PPV*** **(%)** | 97.1 | 97.8 | 98.1 | 97.8 | 98.0 | 98.4 | 97.4 | 98.2 | 98.4 |

The experiments show that KPCA has a higher performance than PCA and DCT. It can be explained by the capability of nonlinear PCA algorithms to capture nonlinear correlations between the data. It leads to an excellent trade-off in to preserve the biggest information with a minimum number of features.

The values of statistic $|z|$ are shown in Table 8 for all possible combinations of each pair of classifiers. The value for each pair of classifiers is obtained by intercepting the row and column of the table. In all cases, the null hypothesis has to be rejected meaning that differences in the performance indexes for each classifier are statistically significant among methods, validating the results.

**Table 8.** Results of McNemar's Test for all classifiers and beats the database. The grouping method in MLP is all cases.

| | **PCA10** | **PCA15** | **PCA20** | **DCT10** | **DCT15** | **DCT20** | **KPCA10** | **KPCA15** |
|---|---|---|---|---|---|---|---|---|
| **PCA10** | | | | | | | | |
| **PCA15** | 12.5 | | | | | | | |
| **PCA20** | 7.1 | 5.6 | | | | | | |
| **DCT10** | 100.8 | 106.3 | 104.2 | | | | | |
| **DCT15** | 154.7 | 159.9 | 158.8 | 74.3 | | | | |
| **DCT20** | 170.9 | 176.2 | 173.7 | 88.1 | 28.1 | | | |
| **KPCA10** | 208.5 | 210.5 | 209.6 | 140.3 | 80.2 | 60.9 | | |
| **KPCA15** | 218.8 | 222.3 | 221.5 | 151.0 | 98.5 | 83.7 | 25.3 | |
| **KPCA20** | 193.3 | 197.1 | 195.5 | 119.3 | 53.6 | 30.6 | 32.7 | 58.1 |

It was not possible to compare the results with other studies about artefacts recognition, because there are no other publications to our knowledge about this particular topic.

## 4 Conclusions

The trade-off between performance indexes suggests the use of the non-linear principal component analysis as feature extraction method and a multilayer perceptron as clustering method. In spite of its high runtime, it can be implemented with reasonable resources taking into account the current computer technologies. The future improvement and optimization of KPCA algorithm could ensure its practical application with a greater efficiency and speed, for example, using programmable devices to accelerate the calculation of principal components.

## References

1. Hosseini, H.G., Luo, D., Reynolds, K.J.: The comparison of different feed forward neural network architectures for ECG signal diagnosis. Medical Engineering & Physics 28, 372–378 (2006)
2. Mitra, S., Mitra, M., Chaudhuri, B.B.: Pattern defined heuristic rules and directional histogram based online ECG parameter extraction. Measurement 42, 150–156 (2009)
3. Engin, M.: ECG Beat Classification Using Neuro-fuzzy Network. Pattern Recognition Letters 25, 1715–1722 (2004)
4. Joy, M.R., Chakraborty, C., Ajoy, K.R.: A two-stage mechanism for registration and classification of ECG using Gaussian mixture model. Pattern Recognition 42, 2979–2988 (2009)
5. Acir, N.: A support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems. Expert Systems with Applications 31, 150–158 (2006)
6. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology: “Heart rate variability. Standards of measurement, physiological interpretation, and clinical use”. European Heart Journal 17, 354–381 (1996)
7. Ceylan, R., Özbay, Y.: Comparison of FCM, PCA and WT techniques for classification ECG arrhythmias using artificial neural network. Expert Systems with Applications 33, 286–295 (2007)
8. Patra, D., Das, M.K., Pradhan, S.: Integration of FCM, PCA and Neural Networks for classification of ECG arrhythmias. IAENG International Journal of Computer Science 36(3) (2009)
9. Özbay, Y., Ceylan, R., Karlik, B.: A Fuzzy Clustering Neural Network Architecture for Classification of ECG Arrhythmias. Computers in Biology and Medicine 36, 376–388 (2006)
10. Karpagachelvi, S., Arthanari, M., Sivakumar, M.: ECG Feature Extraction Techniques-A Survey Approach. International Journal of Computer Science and Information Security 8(1), 76–80 (2010)
11. Neto, J.E., Suárez-León, A.A., Vázquez Seisdedos, C.R., López-Mora, N.A., Leite, J.C., Oliveira, R.C.L.: Métodos de extracción de características en el ECG: análisis comparativo. In: V Latin American Congress on Biomedical Engineering, CLAIB 2011, Cuba. IFMBE Proceedings, vol. 33, pp. 858–861 (2011) (in Spanish), `http://link.springer.com/content/pdf/10.1007%2F978-3-642-21198-0_218`
12. Moody, G.B., Mark, R.G.: The MIT-BIH Arrhythmia Database on CD-ROM and software for use with it. Computers in Cardiology 17, 185–188 (1990)
13. Köhler, B.U., Hennig, C., Orglmeister, R.: The Principles of Software QRS Detection. Reviewing and Comparing Algorithms for Detecting this Important ECG Waveform. IEEE Engineering in Medicine and Biology, 42–57 (2002)

# Comparing Binary Iris Biometric Templates Based on Counting Bloom Filters

Christian Rathgeb and Christoph Busch

da/sec Biometrics and Internet Security Research Group
Hochschule Darmstadt, Darmstadt, Germany
{christian.rathgeb,christoph.busch}@h-da.de

**Abstract.** In this paper a binary biometric comparator based on Counting Bloom filters is introduced. Within the proposed scheme binary biometric feature vectors are analyzed and appropriate bit sequences are mapped to Counting Bloom filters. The comparison of resulting sets of Counting Bloom filters significantly improves the biometric performance of the underlying system. The proposed approach is applied to binary iris-biometric feature vectors, i.e. iris-codes, generated from different feature extractors. Experimental evaluations, which carried out on the CASIA-v3-Interval iris database, confirm the soundness of the presented comparator.

## 1 Introduction

Iris biometric recognition [2] is field-proven as a robust and reliable biometric technology. The iris's complex texture and its apparent stability hold tremendous promise for applying iris recognition in diverse application scenarios, such as border control or forensic investigations [12]. Daugman's algorithm [3], forms the basis of the vast majority of today's iris recognition systems, which report (true positive) identification rates above 99% and equal error rates less than 1%: (1) at enrollment an image of a subject's eye is acquired; (2) in the pre-processing step the boundary of the pupil and the outer iris are detected and the iris (in the approximated form of a ring) is "un-rolled" to obtain a normalized rectangular iris texture; (3) feature extraction is applied in order to generate a highly discriminative binary feature vector called iris-code; (4) at the time of authentication pairs of iris-codes are efficiently compared by calculating the Hamming distance between them, where template alignment is performed within a single dimension, applying a circular shift of iris-codes, to compensate against head tilts of a certain degree. While most approaches to iris recognition algorithms focus on extracting highly discriminative iris-codes, potential improvements within comparators are frequently neglected.

The contribution of this work is the proposal of a binary biometric comparator based on Counting Bloom filters (CBFs) [1,5]. In the presented scheme iris-codes are transformed to sets of CBFs which enables an enhanced biometric comparison, yielding a significant improvement in biometric performance.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 262–269, 2013.

In addition the generic comparator does not require a re-enrollment of registered subjects, i.e. it can be integarted to any existing iris recognition system.

This paper is organized as follows: related work is summarized in Sect. 2. The proposed comparator based on CBFs is described in detail in Sect. 3. Experiments are presented in Sect. 4 and conclusions are drawn in Sect. 5.

## 2 Template Comparison in Iris Recognition

Focusing on iris recognition, a binary representation of biometric features offers two major advantages:

1. Rapid authentication (even in identification mode).
2. Compact storage of biometric templates.

Comparisons between binary biometric feature vectors are commonly implemented by the simple Boolean exclusive-OR operator (XOR) applied to a pair of binary biometric feature vectors, masked (AND'ed) by both of their corresponding mask templates to prevent occlusions caused by eyelids or eyelashes from influencing comparisons. The XOR operator $\oplus$ detects disagreement between any corresponding pair of bits, while the AND operator $\cap$ ensures that the compared bits are both deemed to have been uncorrupted by noise. The norms ($|| \cdot ||$) of the resulting bit vector and of the AND'ed mask template are then measured in order to compute a fractional Hamming distance (*HD*) as a measure of the (dis-)similarity between pairs of binary feature vectors {codeA, codeB} and the according mask bit vectors {maskA, maskB} [3]:

$$HD = \frac{||(\text{codeA} \oplus \text{codeB}) \cap \text{maskA} \cap \text{maskB}||}{||\text{maskA} \cap \text{maskB}||}. \tag{1}$$

Note that for the dis-similarity metrics the score for a genuine comparison (i.e. both codes stemming from the same source) is expected to be low. Apart from the fractional Hamming distance several other techniques of how to compare iris-codes have been proposed. To obtain a representative user-specific iris template during enrollment Davida *et al.* [4] and Ziauddin and Dailey [13] analyze several iris-codes. Davida *et al.* propose a majority decoding where the majority of bits is assigned to according bit positions in order to reduce *HD*s between genuine iris-codes. Experimental results are omitted. Ziauddin and Dailey suggest to assign weights to each bit position, defining the stability of bits at according positions. Hollingsworth *et al.* [6] examined the consistency of bits in iris-codes resulting from different parts of the iris texture. The authors suggest to mask out so-called “fragile” bits for each subject, where these bits are detected from several iris-code samples. In experiments the authors achieve a significant performance gain. Obviously, applying more than one enrollment sample yields better recognition performance, however, commercial applications usually require single sample enrollment as the operational constraints can not tolerate an extended capture process duration. Rathgeb *et al.* [10,11] have demonstrated that incorporating preliminary comparison scores, which are obtained during the

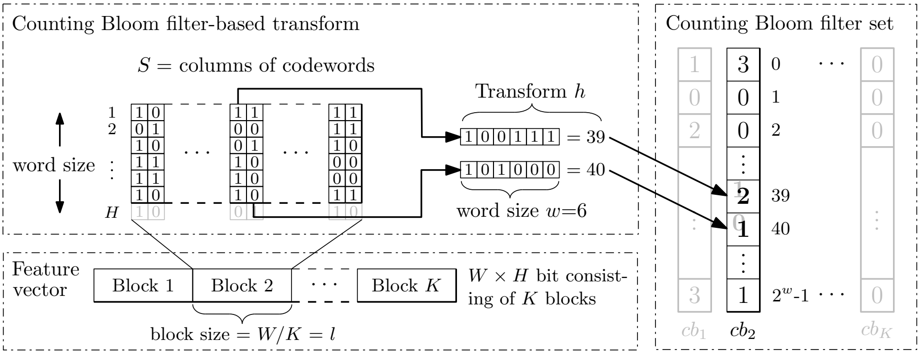


**Fig. 1.** Proposed Counting Bloom filter-based transform: highlighted codewords increment in $cb_2$ the element at index 39 and 40

alignment process, significantly increases biometric performance. *HD* scores are expected to decrease towards an optimal alignment, i.e. the distance between the lowest and highest score as well as the overall distribution yielded by scores at different shifting positions, indicates (non-)genuine comparisons. Typically, minor improvements do not lead to significant performance gain with respect to accuracy. On the other hand, more complex comparison techniques do not provide a rapid comparison of biometric templates, yielding a trade-off between computational effort and recognition accuracy.

## 3 Counting Bloom Filter-Based Comparator

Basically, a Bloom filter $b$ is a bit array of length $n$, where initially all bits are set to 0 [1]. In order to represent a set $S$ a Bloom filter traditionally utilizes $k$ independent hash functions $h_1, h_2, ..., h_k$ with range $[0, n-1]$. For each element $x \in S$, bits at positions $h_i(x)$ of Bloom filter $b$ are set to 1, for $1 \leq i \leq k$. To test if an element $y$ is in $S$, it has to be checked whether all position of $h_i(y)$ in $b$ are set to 1. If this is the case, it is assumed that $y$ is in $S$ with a certain probability of false positive. If not, clearly $y$ is not a member of $S$, hence, traditional Bloom filters are suitable for any application where a distinct probability of false positive is acceptable. In a Counting Bloom filter $cb$, which has first been introduced by Fan et al. [5], the array positions are extended from being a single bit to being an integer counter.

In the following subsections, the CBF-based transform, which is depicted in Fig. 1, and the corresponding comparison technique are described in detail.

### 3.1 Counting Bloom Filter-Based Transform

In the proposed system CBFs are utilized in order to achieve an alignment-free representation (to a certain degree) of iris-codes. For this purpose the original concept of CBFs is adapted in two ways:

**Algorithm 1.** Construction of CBF-based template.

```
for $j = 0 \rightarrow K - 1$ do                          ▷ process each block of the feature vector
    for $i = j * l \rightarrow j * (l + 1)$ do                  ▷ process each codeword within a block
        $cb_j[h(x_i)] \leftarrow cb_j[h(x_i)] + 1$    ▷ increment the CBF at the according position
    end for
end for
```

1. A single trivial transform $h$ is utilized instead of numerous hash function.
2. A fixed number of exactly $l$ elements are inserted into an according CBF.

Generic iris recognition systems [2] extract binary feature vectors based on a row-wise analysis of normalized iris textures, i.e. iris-codes typically represent two-dimensional binary feature vectors of width $W$ and height $H$ (see Fig. 2 (e)-(f)). In the proposed scheme $W \times H$ iris-codes are divided into $K$ blocks of equal size, where each column consists of $w \leq H$ bits. In case $w < H$, columns consist of the $w$ upper most bits, i.e. features originating from outer iris bands, which are expected to contain less discriminative information, are ignored and not represented in the CBF. Subsequently, the entire sequence of columns of each block is successively transformed to according locations within CBFs, that is, a total number of $K$ separate Bloom filters of length $n = 2^w$ form the template of size $K \cdot 2^w$. The transform is implemented by mapping each column in the iris-code to the index of its decimal value, which is shown for two different codewords (=columns) as part of Fig. reffig:system, for each column $x \in \{0,1\}^w$, the mapping is defined as,

$$h(x) = \sum_{j=0}^{w-1} x_j \cdot 2^j. \qquad (2)$$

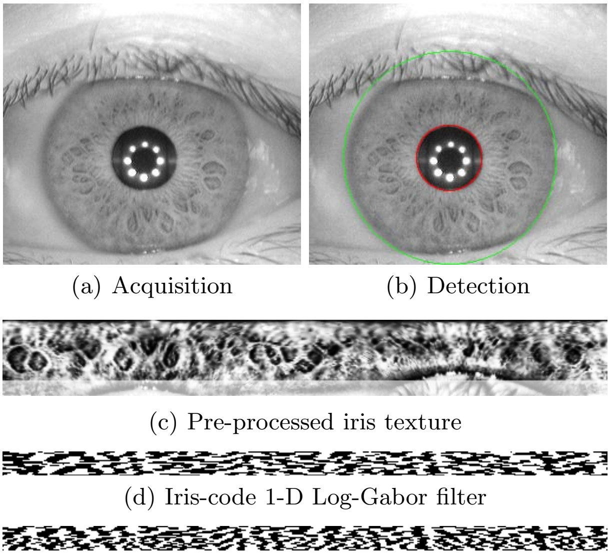
The entire process of constructing a set of CBFs which represents a distinct iris-code is described in Algorithm 1. The representation is alignment-free, i.e. generated templates (=sets of CBFs) do not need to be aligned at the time of comparison. Equal columns within certain blocks (=codewords) increment identical indexes within CBFs, i.e. self-propagating errors caused by an inappropriate alignment of iris-codes are eliminated (radial neighborhoods persist).

### 3.2 Comparison in Transformed Domain

The dissimilarity $DS$ between two CBFs $cb$ and $cb'$ of length $n$,$n = 2^w$, is defined as the sum of difference at each index of both CBFs,

$$DS(cb, cb') = \sum_{j=1}^{n} |cb_j - cb'_j|/2l, \qquad (3)$$

Obviously, $DS$ requires more computational effort compared to $HD$, however, $DS$ does not have to be computed at numerous shifting positions. In order to incorporate masking bits obtained at the time of pre-processing, columns of iris-codes which are mostly affected by occlusions must not be mapped to Bloom filters, i.e. a separate storage of bit masks is not required.



(a) Acquisition

(b) Detection

(c) Pre-processed iris texture

(d) Iris-code 1-D Log-Gabor filter

(e) Iris-code Ma *et al.*

**Fig. 2.** Iris processing chain: applied pre-processing and feature extraction algorithms

# 4 Experiments

Performance is estimated in terms of false non-match rate (FNMR) at a targeted false match rate (FMR) and equal error rate (EER). In accordance to the International Standard ISO/IEC IS 19795-1 [7] the FNMR of a biometric system defines the proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample. By analogy, the FMR defines the proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template. As score distributions overlap EERs are obtained, i.e. the system error rate where FNMR = FMR.

## 4.1 Experimental Setup

Experiments are carried out using the CASIA-v3-Interval iris database[1]. At pre-processing the iris of a given sample image is detected, un-wrapped to an enhanced rectangular texture of 512 × 64 pixel, shown in Fig. 2 (a)-(d).

In the feature extraction stage custom implementations[2] of two different iris recognition algorithms are employed where normalized iris textures are divided into stripes to obtain 10 one-dimensional signals, each one averaged from the

[1] The Center of Biometrics and Security Research, `http://www.idealtest.org`

[2] USIT – University of Salzburg Iris Toolkit v1.0, `http://www.wavelab.at/sources/`

**Table 1.** Original performance (in %) for both feature extractors (*HD* comparator)

| Aligorithm | 1-FNMR @ FMR=0.01 | EER |
|---|---|---|
| 1-D Log Gabor | 95.03 | 1.58 |
| Ma *et al.* | 96.16 | 1.19 |

**Table 2.** 1-FNMRs @FMR=0.01 (in %) for different configurations of the comparator

| Algorithm | Word size | Block size $l$ (bits) | | | | |
|---|---|---|---|---|---|---|
| | $w$ (bits) | $2^5$ | $2^6$ | $2^7$ | $2^8$ | $2^9$ |
| 1D Log Gabor | 10 | 95.75 | 94.32 | 88.43 | 66.24 | 31.81 |
| | 9 | 94.98 | 94.27 | 89.36 | 64.45 | – |
| | 8 | 93.65 | 93.91 | 87.97 | – | – |
| Ma *et al.* | 10 | 98.15 | 96.11 | 93.40 | 82.65 | 60.71 |
| | 9 | 97.80 | 94.88 | 91.30 | 76.21 | – |
| | 8 | 97.08 | 93.40 | 87.92 | – | – |

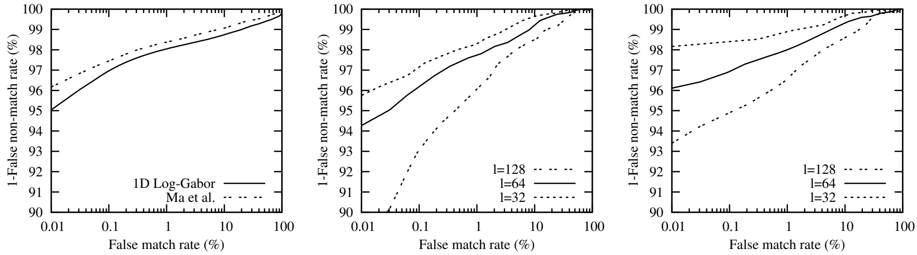
pixels of 5 adjacent rows (the upper 512 × 50 rows are analyzed). The first feature extraction method follows an implementation by Masek [9] in which filters obtained from a Log-Gabor function are applied. A row-wise convolution with a complex Log-Gabor filter is performed on the texture pixels and the phase angles of resulting complex values are discretized into 2 bits generating a binary code of 512 × 20 = 10240 bit. The second feature extraction algorithm was proposed by Ma *et al.* [8]. Within this algorithm a dyadic wavelet transform is performed on 10 signals obtained from the according texture stripes. For two selected subbands minima and maxima above an adequate threshold are located, and a bit-code of $512 \times 20 = 10240$ bits is extracted. Sample iris-codes generated by both feature extraction methods are shown in Fig. 2 (e)-(f). iris-code are divided into upper and lower 512 × 10 halfs as these represent real and complex values or minima and maxima extracted from different subbands, respectively.

### 4.2 Performance Evaluation

The biometric performance of the original systems, in which $HD$-based iris-code comparisons are performed at ± 8 circular bit shifts, are shown in Table 1. The corresponding receiver operation characteristic (ROC) curves are plotted in Fig. 3 (a). For both feature extraction techniques practical performance rates are achieved, yielding EERs of 1.58% and 1.19%, respectively. With respect to the proposed CBF-based comparator, Table 2 and Table 3 summarize obtained 1-FNMRs at target FMRs of 0.01% and EERs for different word sizes $w$ and block sizes $l$ for both feature extraction algorithms. As can be seen, a choice of large block sizes implies a greater loss of local information (original positions of codewords) and causes a drastic decrease in biometric performance. From the

**Table 3.** EERs (in %) for different configurations of the proposed comparator

| Algorithm | Word size | Block size $l$ (bits) | | | | |
|---|---|---|---|---|---|---|
| | $w$ (bits) | $2^5$ | $2^6$ | $2^7$ | $2^8$ | $2^9$ |
| 1D Log Gabor | 10 | 1.21 | 1.75 | 2.49 | 4.54 | 7.87 |
| | 9 | 1.34 | 1.77 | 3.02 | 4.74 | – |
| | 8 | 1.42 | 1.93 | 3.17 | – | – |
| Ma *et al.* | 10 | 0.88 | 1.56 | 2.54 | 4.10 | 6.90 |
| | 9 | 1.04 | 1.61 | 2.70 | 4.62 | – |
| | 8 | 1.09 | 1.67 | 3.22 | – | – |



(a) Original (b) 1D-LogGabor $w = 10$ (c) Ma *et al.* $w = 10$

**Fig. 3.** ROC curves for (a) the original $HD$-based comparator and the proposed algorithm for (b) the 1D Log-Gabor feature extractor and (c) the algorithm of Ma *et al.* for different settings of block sizes and a word size of $w = 10$

obtained results it is clear that rotations of ± 8 bits, which significantly affect original systems, are compensated. For both feature extraction algorithms performance is gained for different configurations, achieving best results at word size of $w = 10$ and a block size of $l = 32$, obtaining EERs of 1.21% and 0.88%, respectively. The according ROC curves for a word size of $w = 10$ are depicted in Fig. 3 (b)-(c). Significant improvement is obtained compared to the original system, while the proposed scheme does not require re-enrollment or any adaption of the original iris-codes. CBFs can be stored in addition to iris-code records or efficiently calculated at the time of comparison.

## 5 Conclusions

In this work an advanced binary biometric comparator based on counting Bloom filters has been introduced. Compared to a conventional, $HD$-based comparison, within the proposed approach iris-codes are transformed to sets of CBFs, prior to comparison. Additional computational effort is limited since the CBF-based representation enables an alignment-free comparison. The system is evaluated on a publicly available dataset where it gains biometric performance for different feature extraction techniques, confirming the soundness of the presented approach.

**Acknowledgment.** This work has been partially funded by the European FP7 FIDELITY project (SEC-2011-284862) and the Center of Applied Security Research Darmstadt (CASED).

## References

1. Bloom, B.: Space/time tradeoffs in hash coding with allowable errors. Communications of the ACM 13(7), 422–426 (1970)
2. Bowyer, K., Hollingsworth, K., Flynn, P.: Image understanding for iris biometrics: A survey. Computer Vision and Image Understanding 110(2), 281–307 (2007)
3. Daugman, J.: How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 21–30 (2004)
4. Davida, G., Frankel, Y., Matt, B.: On enabling secure applications through offline biometric identification. In: Proc. IEEE Symp. on Security and Privacy, pp. 148–157. IEEE (1998)
5. Fan, L., Cao, P., Almeida, J., Broder, A.Z.: Summary cache: a scalable wide-area web cache sharing protocol. IEEE/ACM Transactions on Networking 8(3), 281–293 (2000)
6. Hollingsworth, K.P., Bowyer, K.W., Flynn, P.J.: The best bits in an iris code. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 964–973 (2009)
7. ISO/IEC TC JTC1 SC37 Biometrics. ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework. International Organization for Standardization and International Electrotechnical Committee (March 2006)
8. Ma, L., Tan, T., Wang, Y., Zhang, D.: Efficient iris recognition by characterizing key local variations. IEEE Transactions on Image Processing 13(6), 739–750 (2004)
9. Masek, L.: Recognition of human iris patterns for biometric identification. Master's thesis, University of Western Australia (2003)
10. Rathgeb, C., Uhl, A., Wild, P.: Shifting score fusion: On exploiting shifting variation in iris recognition. In: Proc. 26th ACM Symposium on Applied Computing, pp. 1–5. ACM (2011)
11. Rathgeb, C., Uhl, A., Wild, P.: Iris-biometric comparators: Exploiting comparison scores towards an optimal alignment under gaussian assumption. In: Proc. 5th Int'l Conf. on Biometrics, pp. 1–6. IEEE (2012)
12. Rathgeb, C., Uhl, A., Wild, P.: Iris Biometrics: From Segmentation to Template Security. Advances in Information Security, vol. 59. Springer (2012)
13. Ziauddin, S., Dailey, M.: Iris recognition performance enhancement using weighted majority voting. In: Proc. 15th Int'l Conf. on Image Processing, pp. 277–280. IEEE (2008)

# Improving Gender Classification Accuracy in the Wild

Modesto Castrillón-Santana*, Javier Lorenzo-Navarro,
and Enrique Ramón-Balmaseda

SIANI
Universidad de Las Palmas de Gran Canaria, Spain
mcastrillon@siani.es

**Abstract.** In this paper, we focus on gender recognition in challenging large scale scenarios. Firstly, we review the literature results achieved for the problem in large datasets, and select the currently hardest dataset: The Images of Groups. Secondly, we study the extraction of features from the face and its local context to improve the recognition accuracy. Different descriptors, resolutions and classifiers are studied, overcoming previous literature results, reaching an accuracy of 89.8%.

**Keywords:** gender recognition, local context, head and shoulders, LBP, HOG, in the wild.

## 1 Introduction

Gender is a valid demographic characteristic for different applications that has recently attracted commercial attention in the context of audience analysis and advertisement.

Different approaches have tackled the problem of automatic gender recognition. Most recent works have basically considered the face pattern to solve the problem [2,3,14]. Other approaches have made use of non facial features such as the whole body, the hair or clothing [4,13]. However, those approaches including non facial features, have rarely considered uncontrolled large datasets, i.e. the gender recognition in the wild. In this context, the evaluation must tackle more variability in terms of 1) identities, aging and ethnicity, 2) pose and illumination control, and 3) low resolution images.

The contributions of this paper rely firstly on the addition to the information provided by the face, of features extracted from the head local context. Those features are studied at different resolutions, and their possibilities analyzed as additional features for the problem. Another main element of this paper is the use of large databases that are closer to real gender classification scenarios than those small databases obtained in controlled environments.

* This work was partially funded by the Institute of Intelligent Systems and Numerical Applications in Engineering (SIANI) and the Computer Science Department at ULPGC.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 270–277, 2013.

### 1.1 Previous Work in Large Datasets

**Table 1.** Gender recognition accuracy in the previous literature. Refer to each reference for experimental setup details.

| Reference | Dataset | Protocol | Accuracy |
|---|---|---|---|
| [19] | LFW | Subset 7443/13233 | 94.81% |
| [20] | LFW | Subset 7443/13233 | 98.01% |
| [7] | LFW | BEFIT protocol | 97.23% |
| [7] | GROUPS | Subset 15579/28231 | 84.55 – 86.61% |
| [12] | GROUPS | Subset 22778/28231 | 86.4% |
| [5] | MORPH | Subset | 88% |
| [17] | MORPH | Subset | 97.1% |

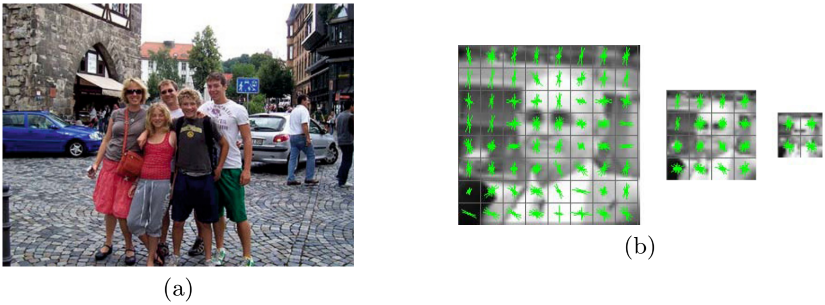
We argue that small databases are not representative for a real world scenario where a gender recognizer must cope with thousands of people, like for example a mall scenario. For that reason, we have reviewed the literature to detect state-of-the-art accuracies obtained for large public databases that contain many different identities acquired without controlled conditions. As far as we know, Table 1 presents the best accuracies reported on large datasets in the recent literature. The datasets studied are The Image of Groups (GROUPS) [10], Labeled Faces in the Wild (LFW) [11], and MORPH [18].

Observing in detail Table 1, there is not much space for improvement in datasets such as LFW and MORPH. Certainly, both datasets present a set of characteristics that might affect the impressive resulting performance. Indeed, in both datasets the same identity includes multiple samples. Additionally, sample images of both genders are not equally represented in the set, i.e. the number of samples corresponding to the male class is significantly larger. On the other side, the GROUPS dataset presents unrestricted imagery with balanced presence of both classes, reporting the lowest accuracy in the recent works. For all those reasons, we have selected to focus on the GROUPS dataset, that represents, in our opinion, the wildest available dataset for the problem, see Figure 1a.

## 2 Representation and Classification

Local descriptors have recently attracted the attention of researchers involved in the facial analysis community [21]. We will focus particularly on Local Binary Patterns [16] (LBPs) and Histograms of Oriented Gradients [8]. Both descriptors have already been used successfully for facial analysis [9,15].

Facial analysis with LBP is currently adopted considering a concatenation of histograms of a predefined grid. This approach was adopted for LBP by Ahonen et al. [1]. According to that work the face is divided into regions where the LBP operator is computed and later their corresponding histograms concatenated, following a Bag of Words scheme [6], into a single histogram. On the other side, HOG encloses a histogram in its definition. The pattern is scaled to a normalized resolution, and later a grid is defined.



(a)

(b)

**Fig. 1.** a) A GROUPS sample. b) Relative size of the different patterns used including the local context: respectively 64×64, 32×32 and 16×16 faces with head and shoulders context. Their respective HOG grid computed is also depicted.

For classification purposes, we will compare two different approaches. The first one will study the addition of features to an initial feature vector filled exclusively with features extracted from the face. In this scenario, two well known classifiers are compared: SVM with linear kernel, and bagging making use of SVM classifiers based on linear kernels.

In the second scenario, instead of combining features of different nature in the feature vector, we focus on the combination of the outputs provided by the different classifiers in a first stage. Their respective scores are combined in a second classification stage. This combination is compared based on different known classification techniques such as: SVM (linear kernel), bagging, naive Bayes, Nearest Neighbor (NN) and C4.5.

## 3 Results

In the experimental setup, we have adopted a k-fold cross-validation, partitioning the dataset into $k$ subsets, repeating $k$ times the experiment using a subset to test the model with the other $k-1$ subsets. In order to be comparable to previous works, we made use of the 5-folds defined in the work by Dago et al. [7].

The Uniform LBP descriptor is used only for the face area, at two different resolutions: 59 × 65 and 100 × 110, defining a 5 × 5 grid. When using HOG as descriptor, the face area is used just with the 59 × 65 resolution, but the head and shoulders pattern was tested at different resolutions: 16 × 16, 32 × 32 and 64 × 64, see Figure 1b. On each resolution the cell contains 8 × 8 pixels, each block 2 × 2 cells, the histogram contains 9 bins, and L2-hys as norm for the normalization stage [8].

### 3.1 Extending the Feature Vector

Firstly, we performed a comparison using just the facial information, i.e. the inner face details (Face), and its local context defined by the head and shoulders area (HS). Tables 2 and 3 present respectively those results. The face pattern resolution used in Table 2 was 59 × 65 pixels, with an inter-eye distance of 26 pixels. For comparison with a baseline, we have also included the results achieved with a classifier trained with the first 100 PCA components.

**Table 2.** Gender recognition accuracy (in brackets results per class: female/male) achieved using PCA, HOG or Uniform LBP features extracted from the face pattern. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

| Pattern and features | Test set GROUPS-Dago 5-folds subset | | | |
|---|---|---|---|---|
| | SVM linear | | Bagging | |
| | Acc. | AUC | Acc. | AUC |
| Face PCA | 0.773<br>(0.773/0.774) | 0.773 | 0.7749<br>(0.779/0.770) | 0.801 |
| Face HOG | 0.801<br>(0.797/0.805) | 0.801 | 0.822<br>(0.84/0.800) | 0.898 |
| Face LBP | **0.838**<br>(0.842/0.834) | 0.838 | **0.838**<br>(0.863/0.814) | 0.910 |

Table 3 presents the results using information extracted from the face and its local context. The head and shoulders were analyzed at different resolutions: 16 × 16, 32 × 32 and 64 × 64, with their respective inter-eye distances of 2.5, 5 and 10 pixels. Observe, that the facial resolution contained in the head and shoulders pattern is lower up to ten times compared to the results reported in Table 2. Even though, the best accuracy is rather similar, almost 84% using the 64 × 64 head and shoulders pattern, than exclusively the facial pattern at larger resolution. Even considering the smallest pattern, with an inter-eye distance under 3 pixels, the accuracy reaches 66%. That is not a bad result for low resolution images.

**Table 3.** Gender recognition accuracy (in brackets results per class: female/male) achieved using HOG features extracted from the head and shoulders (HS) pattern using different image dimensions. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

| Pattern and features | Test set GROUPS-Dago 5-folds subset | | | |
|---|---|---|---|---|
| | SVM linear | | Bagging | |
| | Acc. | AUC | Acc. | AUC |
| $HS_{16\times16}$ HOG | 0.6608<br>(0.6538/0.6684) | 0.661 | 0.659<br>(0.6616/0.6564) | 0.687 |
| $HS_{32\times32}$ HOG | 0.812<br>(0.8024/0.8216) | 0.812 | 0.8099<br>(0.8122/0.8076) | 0.865 |
| $HS_{64\times64}$ HOG | 0.8298<br>(0.829/0.83) | 0.829 | **0.8397**<br>(0.8562/0.8232) | 0.909 |

On a second step, we have considered to fuse in a single feature vector, features extracted from different cues. Table 4 presents results combining Uniform LBP or HOG features extracted from the face pattern, with HOG features extracted from the head and shoulders pattern at different resolutions. Bagging reports better accuracy for the experimental setup, while the use of Uniform LBP features seems to work slightly better than HOG. The notorious increase in the face pattern resolution, does not suggest a large improvement in accuracy. The best reported accuracy reaches 88.1%, four points better than our previous results, and 2% better than the literature for the same dataset, see Table 1. These results suggest the importance of the information contained in the facial local context.

**Table 4.** Gender recognition accuracy (in brackets results per class: female/male) achieved using different representation alternatives. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al. SVM linear and Bagging are used for classification.

| Pattern and features | Test set GROUPS-Dago 5-folds subset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SVM linear | | | | Bagging | | | |
| | face 59 × 65 | | face 100 × 110 | | face 59 × 65 | | face 100 × 110 | |
| | Acc. | AUC | Acc. | AUC | Acc. | AUC | Acc. | AUC |
| Face HOG | 0.801<br>(0.797/0.805) | 0.801 | - | - | 0.822<br>(0.84/0.800) | 0.898 | - | - |
| Face LBP | 0.827<br>(0.835/0.814) | 0.827 | 0.836<br>(0.836/0.836) | 0.836 | 0.823<br>(0.856/0.804) | 0.905 | 0.84<br>(0.862/0.817) | 0.909 |
| Face HOG and $HS_{16\times16}$ HOG | 0.827<br>(0.828/0.827) | 0.827 | - | - | 0.829<br>(0.848/0.809) | 0.904 | - | - |
| Face HOG and $HS_{32\times32}$ HOG | 0.852<br>(0.855/0.85) | 0.852 | - | - | 0.862<br>(0.881/0.843) | 0.93 | - | - |
| Face HOG and $HS_{64\times64}$ HOG | 0.845<br>(0.851/0.84) | 0.845 | - | - | 0.875<br>(0.893/0.858) | 0.941 | - | - |
| Face LBP and $HS_{16\times16}$ HOG | 0.838<br>(0.842/0.834) | 0.838 | 0.844<br>(0.843/0.846) | 0.844 | 0.838<br>(0.863/0.814) | 0.910 | 0.845<br>(0.864/0.826) | 0.915 |
| Face LBP and $HS_{32\times32}$ HOG | 0.859<br>(0.86/0.857) | 0.859 | 0.862<br>(0.861/0.863) | 0862 | 0.867<br>(0.889/0.845) | 0.933 | 0.869<br>(0.864/0.826) | 0.937 |
| Face LBP and $HS_{64\times64}$ HOG | 0.851<br>(0.851/0.85) | 0.851 | 0.861<br>(0.859/0.864) | 0.861 | 0.879<br>(0.897/0.862) | 0.944 | **0.881**<br>(0.897/0.866) | **0.946** |

## 3.2 Stacking Classifiers

We went further, and considered an alternative to the inclusion of more features in the feature vector. Instead, we considered a stacking of classifiers in two stages. The first stage is composed by the individual 11 feature vectors described in Tables 3 and 4, and summarized in the following list:

- HOG of the 64 × 64 head and shoulders pattern (HSHOG64).
- HOG of the 32 × 32 head and shoulders pattern (HSHOG32).
- HOG of the 16 × 16 head and shoulders pattern (HSHOG16).
- HOG of the 59 × 65 facial pattern (FHOG).
- Concatenated LBP histogram extracted from the 59 × 65 facial pattern (FLBP).

- HOG of the 64 × 64 head and shoulders pattern, and HOG of the 59 × 65 face pattern (HSHOG64-FHOG).
- HOG of the 32 × 32 head and shoulders pattern, and HOG of the 59 × 65 face pattern (HSHOG32-FHOG).
- HOG of the 16 × 16 head and shoulders pattern, and HOG of the 59 × 65 face pattern (HSHOG16-FHOG).
- HOG of the 64 × 64 head and shoulders pattern, and concatenated LBP histogram of the 59 × 65 face pattern (HSHOG64-FLBP).
- HOG of the 32 × 32 head and shoulders pattern, and concatenated LBP histogram of the 59 × 65 face pattern (HSHOG32-FLBP).
- HOG of the 16 × 16 head and shoulders pattern, and concatenated LBP histogram of the 59 × 65 face pattern (HSHOG16-FLBP).

Each of the first stage classifier is trained using a SVM with a liner kernel. In the second stage of the stacking classifier, their respective scores are feed into a classifier that is in charge of taking the final decision. For this second stage, we have analyzed the accuracy reported for SVM (linear kernel), Bagging, Naive Bayes, Nearest Neighbor (NN) and C4.5. The results achieved are reported in Table 5. They suggest an improvement, reaching with Naive Bayes almost 90%. The reader may observe, that this accuracy was achieved without using the classifiers based on the largest face pattern, i.e. an inter-eye distance of 26 pixels. Compared to Table 4 for similar facial resolution the improvement is almost 2%. Compared to the literature, see Table 1, the improvement is close to 4% even using a facial pattern that is twice smaller. The benefits introduced by the descriptors and the face local context are evident.

We have additionally performed a feature selection to reduce the system complexity avoiding the computation of all the classifiers present in the stacking first stage. After sorting attending to the information gain, the resulting accuracy considering as variable the number of classifiers included in the stacking is presented in Figure 2. With just 4 classifiers in the first stage, the system performance achieves an accuracy of 89% beating those results reported in the previous section and the literature. Those classifiers are: HSHOG64-FLBP, HSHOG32-FLBP, HSHOG16-FLBP and HSHOG32-FHOG.

**Table 5.** Gender recognition accuracy achieved using classifiers stacking. The table reports the results achieved using the 5-fold cross correlation experiment defined by Dago et al.

| Classifier | Accuracy |
|---|---|
| Naive Bayes | **0.8978** |
| SVM | 0.8736 |
| C4.5 | 0.8336 |
| NN | 0.8662 |

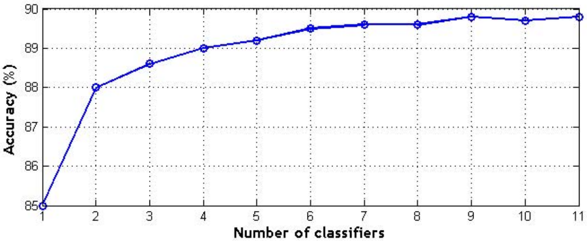


**Fig. 2.** Accuracy achieved adding more classifiers to the stacking

## 4 Conclusions

In this paper, we have studied gender recognition in large uncontrolled datasets. For that purpose, we have made use of facial and non facial features, in the large database that is currently reporting the lowest accuracy in the literature: The Images of Groups.

The addition of external facial features seem to bring benefits at lower resolution, and the combination with facial features reported better accuracies that the previous literature.

We have used features based on the Uniform LBP and HOG operators, both used widely in similar problems. For classification we have considered the used combination of features in a large dataset, and the stacking of classifiers, each one focused in a particular family of features. The stacking results are particularly better than those obtained when the feature vector is increased, reaching almost 90%. This accuracy is notoriously better than those previously reported in the literature, even if the face pattern considered makes use of a facial resolution at least twice smaller.

In summary, the performance exhibited at lower resolution, is best suited for real scenarios. However, the achieved at high resolution beats state of the art results.

## References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12) (December 2006)
2. Alexandre, L.A.: Gender recognition: A multiscale decision fusion approach. Pattern Recognition Letters 31(11), 1422–1427 (2010)
3. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Revisiting linear discriminant techniques in gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 858–864 (2011)
4. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: International Conference on Computer Vision (2011)

5. Chu, W.-S., Huang, C.-R., Chen, C.-S.: Identifying gender from unaligned facial images by set classification. In: International Conference on Pattern Recognition (ICPR), Istanbul, Turkey (2010)
6. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
7. Dago-Casas, P., González-Jiménez, D., Long-Yu, L., Alba-Castro, J.L.: Single- and cross- database benchmarks for gender classification under unconstrained settings. In: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C. (eds.) International Conference on Computer Vision & Pattern Recognition, vol. 2, pp. 886–893. INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334 (June 2005)
9. Déniz, O., Bueno, G., Salido, J., De La Torre, F.: Face recognition using histograms of oriented gradients. Pattern Recognition Letters 32(12), 1598–1603 (2011)
10. Gallagher, A., Chen, T.: Understanding images of groups of people. In: Proc. CVPR (2009)
11. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Univ, of Massachusetts, Amherst (October 2007)
12. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (October 2011)
13. Li, B., Lian, X.-C., Lu, B.-L.: Gender classification by combining clothing, hair and facial component classifiers. Neurocomputing 76(1), 18–27 (2012)
14. 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)
15. Marcel, S., Rodríguez, Y., Heusch, G.: On the recent use of local binary patterns for face authentication. International Journal of Image and Video Preprocessing, Special Issue on Facial Image Processing (2007)
16. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
17. Ramón-Balmaseda, E., Lorenzo-Navarro, J., Castrillón-Santana, M.: Gender classification in large databases. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 74–81. Springer, Heidelberg (2012)
18. Ricanek Jr., K., Tesafaye, T.: MORPH: A longitudinal image database of normal adult age-progression. In: IEEE 7th International Conference on Automatic Face and Gesture Recognition, Southampton, UK, pp. 341–345 (April 2006)
19. Shan, C.: Learning local binary patterns for gender classification on realworld face images. Pattern Recognition Letters 33, 431–437 (2012)
20. Tapia, J.E., Perez, C.A.: Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity and shape. IEEE Transactions on Information Forensics and Security 8(3), 488–499 (2013)
21. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods on the wild. In: Faces in Real-Life Images Workshop in ECCV (2008)

# Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System

Pierre-Michel Bousquet, Jean-François Bonastre, and Driss Matrouf

University of Avignon - LIA, France
{pierre-michel.bousquet,jean-francois.bonastre,
driss.matrouf}@univ-avignon.fr

**Abstract.** This paper focuses on the analysis of the i-vector paradigm, a compact representation of spoken utterances that is used by most of the state of the art speaker verification systems. This work was mainly motivated by the need to quantify the impact of their steps on the final performance, especially their ability to model data according to a theoretical Gaussian framework. These investigations allow to highlight the key points of the approach, in particular a core conditioning procedure, that lead to the success of the i-vector paradigm.

## 1 Introduction

Recent advances in speaker verification have revealed the discriminant power of a new representation of spoken utterances, referred as i-vector[1]. Easy to work with and bringing back the speaker recognition problem to a more traditional biometric pattern recognition problem, i-vectors are now largely used in the most recent speaker verification systems. A classical i-vector system can be briefly decomposed in three stages. First, the acoustic space is structured using the GMM-UBM approach [2] and each speech utterance is represented by a high-dimensional representation denoted "'supervector"'. Then, a low-dimensional representation of this supervector is extracted thanks to a factor decomposition approach. Lastly, a scoring module obtains the final score for a given test, taking advantages of the compact speech utterance representation. Quite often, an additional data conditioning procedure is applied before the scoring step.

The goal of this paper is to assess the impact of each of these stages in terms of global performance. This is important as i-vector approach allows in the past years a drastic progress in terms of performance. A better understanding of the origins of these progresses should allow further improvements and/or some simplification in the quite complex chain of processing. More precisely, we wish to quantify the role of the optional conditioning procedure as we suspect that this module plays a more important role than expected in the performance of i-vector systems.

At all three stages, data modelings have been designed to meet the constraints of a parametric approach, based on Gaussian probabilistic assumptions. The conditioning procedure is also known to help achieve these modeling goals.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 278–285, 2013.

To examine independently each of the stages, we proceed by replacing one by one these modules by methods based on deterministic or non-parametric approaches. The gaps of performance are compared with that involved by the conditioning procedure, then summarized in order to assess the impact of the different approaches. Moreover, replacing methods by others measures the robustness of concepts on which they rely. Results of these investigations can thus highlight the key points in the chain of processing that lead to the success of the i-vector paradigm.

The paper is organized as follows: Sections 2, 3, 4 describe the i-vector based speaker verification system on which we focus. Section 5 presents the alternative methods used at each stage of the system. The experimental results are presented and commented in Sections 6, 7 and conclusions are drawn in Section 8.

## 2 GMM Framework and i-Vector Extraction

Speaker information is modeled by using the Gaussian Mixture Model/Universal Background model (GMM/UBM) paradigm [2] where a weighted sum of Gaussian distributions performs a direct acoustic modeling of the acoustic space. A model of a given speech segment is represented by the Baum-Welch zero and first order statistics of its feature vectors, according to UBM prior distribution. This model is denoted "'supervector"'. The i–vector model [3] constrains the supervector $\mathbf{s}$ of a given speech segment to live in a single subspace following the linear model of a Factor Analysis:

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w} \tag{1}$$

where $\mathbf{m}$ is the supervector corresponding to the UBM, $\mathbf{T}$ is a low-rank rectangular matrix with $G \times F$ rows and $r$ columns, $G$ and $F$ are the number of GMM components and feature dimension, respectively. The $r$ columns of $\mathbf{T}$ are vectors spanning the "total variability" space, and $\mathbf{w}$ is a random vector of size $r$ having a standard normal prior distribution. Determination of $\mathbf{T}$ by using EM-ML procedure and explicit formula of the extracted i-vector $\mathbf{w}$ can be found in [1].

## 3 I-Vector Models and Scorings

The first i-vector based speaker verification systems were based on the LDA–WCCN approach [1], which performs intersession compensation thanks to Linear Discriminant Analysis (LDA) [1], where all the i-vectors belonging to the same speaker are associated with the same class. This technique projects the input data into a much lower dimensional space with minimal loss of discriminative ability, as the ratio of between-speaker and within-speaker variations is maximized. These speaker features are finally normalized by a Within Class Covariance Normalization (WCCN) [4]. The final scores are then computed using a cosine distance scoring [3].

A key evolution of i-vector approach was introduced in [5], using the Probabilistic Linear Discriminant Analysis (PLDA) [6]. Two assumptions on the prior probability distributions of the PLDA variables (speaker, session and residual factors of eq. 7 in [7]) have been proposed:

- Gaussian PLDA (G-PLDA) assumes that all latent variables are statistically independent. Standard normal priors are assumed for speaker and session factors. The residual term is assumed to be Gaussian with zero mean and diagonal covariance matrix.
- Student's t-distribution is proposed in [5] as an alternative to the Gaussian to model the speaker and channel subspaces in the i-vector space. Heavy-tailed PLDA (HT-PLDA) assumes that all the factors follow an heavy-tailed distribution, scaled by gamma distribution scalars.

The ML point estimates of the model parameters are obtained from a large collection of development data using an EM algorithm as in [6].

## 4 Pre-conditioning

A pre-processing before any i-vector modeling has been introduced in [8][9]. I-vectors are whitened and length-normalized, in order to make them more Gaussian. The most commonly used whitening technique is a standardization, and the transformation applied to an i-vector $\mathbf{w}$ can be resumed as follows:

$$\mathbf{w} \leftarrow \frac{\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)}{\left\|\mathbf{A}^{-\frac{1}{2}}\left(\mathbf{w}-\mu\right)\right\|} \tag{2}$$

where $\mu$ and $\mathbf{A}$ are the mean and a variability matrix of a training corpus. Data are standardized according to a variability matrix $\mathbf{A}$ then length-normalized, confining the i-vectors to the hypersphere of unit radius. Parameters are computed for the i-vectors present in the training corpus and applied to the test i-vectors. The matrix $\mathbf{A}$ can be the total covariance matrix or, as we proposed in [7], the within-class covariance matrix $\mathbf{W}$ defined in eq. 4 of [7].

In [9], it is shown that this technique improves the gaussianity of the i-vectors. It reduces the gap between the underlying assumptions on the data distribution and the real distribution and also reduces the dataset shift between development and trial i-vectors. Moreover, it is shown in [9] that performance of a G-PLDA system with this pre-conditioning is competitive versus the HT-PLDA, when the latter is much more complicated. As proposed in [8][7], these two-steps can be iterated. As a result, i-vectors tend to be simultaneously $\mathbf{A}$-standardized and length-normalized (magnitude 1), involving a number of properties related to intersession compensation. Some of them are detailed in [8][7]. Also in [7], we propose, after $\mathbf{W}$-standardization, a deterministic initialization of PLDA matricial metaparameters $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ of eq. 7 in [7]. It allows a faster convergence of the PLDA EM-ML procedure.

## 5 Alternative Methods

The state of the art i-vector-based system described below is composed of three stages: representation of segments by Baum-Welch zero and first order UBM-statistics, i-vector extraction using Factor Analysis total-variability (*FA-total-var*), Gaussian-PLDA modeling and scoring with an optional pre-conditioning. We present here the alternative methods that we have implemented for each of these three stages.

### 5.1 Models and Scorings

To analyze the efficiency of Gaussian-PLDA, we compare this probabilistic modeling with two simplified and deterministic versions. First, the LDA-two-covariance model [10] reduces the dimensionality by using LDA, then full rank matrices $\mathbf{\Phi}$ and $\mathbf{\Gamma}$ of eq.7 in [7] are deterministically estimated (no EM-ML procedure is performed) by $\mathbf{\Phi} = \mathbf{B}^{\frac{1}{2}}$ and $\mathbf{\Gamma} = \mathbf{W}^{\frac{1}{2}}$ where $\mathbf{B}$ and $\mathbf{W}$ are the between- and within-class covariance matrices defined in eq. 3, 4 of [7]. Comparing Gaussian-PLDA and LDA-two-covariance model measures the gain of the probabilistic ML-approach in a generative i-vector modeling. Second, the LDA-Mahalanobis model, introduced in [8] is a particular case of the previous two-covariance model which makes no assumption about the speaker factor distribution (speaker precision matrix $\mathbf{B}^{-1}$ is null). The deterministic Mahalanobis model is useful to estimate the relevance of a between-speaker modeling.

### 5.2 I-Vector Extraction

Factor analysis total variability (FA-total-var) is the state of the art factor decomposition technique used to extract i-vectors. To assess the pertinence of its probabilistic approach, we compare it with the well-known deterministic principal component analysis (PCA). But FA-total-var is based on zero and first order statistics and applying PCA to extract low dimensional vectors (that we will also call *i-vectors*) needs to determine the unique high-dimensional vectorial representation to compress. Some solutions have been suggested [11]. In order to fairly compare probabilistic FA-total-var and deterministic PCA, we introduce an adapted version $\widehat{\mathbf{s}}$ of a supervector $\mathbf{s}$, equal to:

$$\widehat{\mathbf{s}} = N_{\mathcal{X}} \left(\mathbf{\Sigma} + N_{\mathcal{X}}\right)^{-1} \left(\mathbf{s} - \mu\right) \tag{3}$$

$N_{\mathcal{X}}$ is the $GF \times GF$ diagonal matrix composed of $F$ blocks of $N_{\mathcal{X}}^{(g)}\mathbf{I}$ ($g = 1, ..., G$) where $N_{\mathcal{X}}^{(g)}$ are the zero–order statistics estimated on the $g$-th Gaussian component of the UBM observing the set of feature vectors in the sequence $\mathcal{X}$, and $\mu$ and $\mathbf{\Sigma}$ are the UBM mean and diagonal covariance matrix.

In the extreme case of a square and full rank identity matrix $\mathbf{T}$ (no dimensionality reduction applied), eq. 6 of [1] shows that FA-extraction provides an i-vector $\mathbf{w}$ equal to $\widehat{\mathbf{s}}$.

The supervector $\widehat{\mathbf{s}}$ is an adapted version of $\mathbf{s}$, centered and weighted by the amount of informations per Gaussian-component and by the variance per dimension.

### 5.3 UBM-Based Representation

In [12][13] a new approach for speaker recognition, denoted "Speaker Binary Key", was presented. Contrary to classical speaker recognition based on statistical modeling of the speaker information, this approach proposes to handle directly each piece of speaker specific information in a binary space. Each coefficient of this binary space corresponds to a targeted piece of speaker-specific information which could be present (the coefficient is equal to 1) or non present (the coefficient is equal to 0) in a given acoustic frame or acoustic segment. This new approach allows to exploit temporal or sequential information as a binary vector is extracted for each acoustic frame. It also focuses on speaker specific information in a non-parametric way as each coefficient of the binary space models speaker-specific information. As the binary key representation first ties each input frames with one or several GMM-UBM components (before non-parametric transformation to a binary space), it constitutes a GMM-UBM-based alternative to the zero and first order statistics. High-dimensional binary keys provided by this model are projected onto a PCA subspace (by the lack of a specific Factor Analysis), and handled as i-vectors for modeling and scoring.

## 6 Experimental Setup

The feature extraction and the 512-components GMM-UBM functionalities used in our experiments are described in [8]. For i-vector extraction, the total variability matrix $\mathbf{T}$ is trained using 15660 speech utterances from 1147 speakers (NIST 2004-05-06, Switchboard II part 1, 2 & 3; Switchboard cellular part 1 & 2, about 14 sessions per speaker). The results are reported with 400-dimensional i-vectors. The same database is used to estimate the parameters of the i-vector models and scorings. In PLDA, channel factor is kept full and speaker factor is varied, as proposed in [5]. Evaluation was performed on the NIST SRE 2008 DET conditions 6 and 7, male only, corresponding to telephone-telephone (all and English-only respectively) enrollment-verification trials, and on the NIST SRE 2010 DET extended condition 5, male only, corresponding to telephone-telephone. A global measurement of performance of a system is given by the average of the three Equal Error Rates (EER). These three conditions are the most currently used in the domain and their average EER is a robust performance measure of a system.

## 7 Results

Table 1 shows comparison result of systems applying the different representations, extractors, models and scorings listed above. The first eight systems use

**Table 1.** Comparison of performance, in terms of EER (%), between systems based on different representations, extractors, models and scorings (without and with pre-conditioning)

| | repr. | extract. | conditioning | model and scoring | det 7 | det 6 | det 5 ext | **average** |
|---|---|---|---|---|---|---|---|---|
| 1 | sv | FA | no | LDA-Maha | 5.70 | 9.5 | 9.73 | **8.31** |
| 2 | sv | FA | no | LDA-two-cov | 3.23 | 6.83 | 5.97 | **5.34** |
| 3 | sv | FA | no | G-PLDA | 3.39 | 6.37 | 6.38 | **5.38** |
| 4 | sv | FA | WCCN-cosine | LDA-WCCN-cosine | 3.26 | 6.29 | 3.69 | **4.41** |
| 5 | sv | FA | L$\boldsymbol{\Sigma}$ | LDA-Maha | 1.86 | 5.06 | 2.62 | **3.18** |
| 6 | sv | FA | L$\boldsymbol{\Sigma}$ | LDA-two-cov | 1.53 | 4.93 | 2.36 | **2.94** |
| 7 | sv | FA | L$\boldsymbol{\Sigma}$ | G-PLDA | 1.63 | 4.80 | 2.45 | **2.96** |
| 8 | sv | FA | L**W** | G-PLDA | 1.58 | 4.80 | 2.28 | **2.89** |
| 9 | BK | PCA | no | G-PLDA | 2.84 | 5.82 | 4.42 | **4.36** |
| 10 | sv | PCA | no | G-PLDA | 3.17 | 6.59 | 5.80 | **5.19** |
| 11 | BK | PCA | L**W** | G-PLDA | 2.16 | 5.26 | 2.87 | **3.43** |
| 12 | sv | PCA | L**W** | G-PLDA | 1.99 | 5.24 | 2.47 | **3.23** |

high-dimensional representation by zero and first order UBM statistics (**sv** for supervector) and Factor Analysis on total variability (**FA**) as i-vector extractor. Performance are given without (**no**) and with pre-conditioning: L$\boldsymbol{\Sigma}$, L**W** for standardization according to total $\boldsymbol{\Sigma}$ or within-class **W** covariance matrix, or **WCCN-cosine** as implicit normalization of LDA-WCCN-cosine scoring. HT-PLDA scoring has not been carried out, as pre-conditioning and Gaussian-PLDA are able to match its performance.

The state of the art system (line 8) yields the best result: average EER of 2.89 and best EERs for all the individual conditions. But, first, the gap between ML (lines 7 and 8) and deterministic approach (line 6) for i-vector modeling is slight or null (average EER of 2.89 and 2.96 vs 2.94). This observation is strengthened by the fact that the best system (line 8) deterministically initializes PLDA metaparameters then requires only 10 EM-ML iterations to converge, against 100 using the randomly initialized system (line 7). Second, comparison of systems without and with pre-conditioning shows that the quality of the modeling is, in a major proportion, the consequence of the conditioning: 5.34 to 2.94 for the best deterministic approach, 5.38 to 2.89 for the probabilistic approach. It is worth noting that the gap between the less efficient system (LDA-Mahalanobis) and the others is particularly significant in the absence of pre-conditioning (8.31 vs 5.34 without, 3.18 vs 2.94 with). This shows that the initial lack of gaussianity in the extracted i-vectors is mainly due to the within-speaker distribution.

The four last lines give comparison result between systems using representation by speaker binary key (**BK**) or by zero and first order UBM statistics, all using i-vector extraction techniques by PCA (**PCA**), each time without and with pre-conditioning (L**W** only, since it gives the better performance in the previous experiments). Comparing the extraction techniques (lines 8 and 12), FA brings a relative improvement of 10.5% of average EER: 2.89 vs 3.23 with PCA.

This slight gain recalls that i-vector extraction falls into the family of compression techniques rather than factor decompositions. Comparing representations for PCA-based systems (lines 11 and 12), the binary key representation yields a performance close to that of zero and first order UBM statistics (3.43 vs 3.23) with, which must be taken into account, a 32 times lower amount of information[1]. But once again, the improvement of performance is mainly due to the conditioning step. Systems based on different representations and dimensionality reductions are able to provide interesting performance but only if they include a pre-conditioning procedure.

## 8 Conclusion

The aim of this work was to assess the benefits of the different steps in a classical i-vector based speaker verification system. In particular, we quantify the role of the optional conditioning procedure in the good probabilistic modeling of data. As all stages of the system try to take into account the constraints of a Gaussian framework, we replace one by one these modules by a deterministic or nonparametric method and compare the gap of performance with that involved by the conditioning procedure. These comparisons also allow to measure the robustness of concepts involved in the i-vector approach. The results of this analysis can be summarized by the following key points:

- All the systems presented here rely on the GMM-UBM. Their good performance, following however various ways, show the robustness of the GMM-UBM to structure the acoustic feature space.
- High-dimensional UBM-based representations are stacking a fixed-length set of vectors from the feature space. The low gaps between systems with various representations and extractors show that any dimensionality reduction of stacked vectors built by using UBM, according to the total variability, is able to capture and summarize correlated behaviors between UBM-components. As remarked in the introduction of [14], the i-vector random variables can be viewed as principal components of utterances. The coordinates represent physical quantities, which are constant for a given utterance but which differ from one utterance to another.
- Resulting low-dimensional vectors do not match the assumptions of an usual probabilistic framework. More than FA-total-var or PLDA decompositions, the conditioning procedure mainly contributes to make vectors compatible with a linear-Gaussian modeling and scoring. WCCN-cosine-scoring can be decomposed into an inner-product applied to standardized and length-normalized vectors, as done in eq. 2. A core procedure, composed of standardization according to a target variability, followed by length-normalization

[1] In our configuration of 512-components GMM-UBM, 50-dimensional feature space and, for binary modeling, 100 specificities per component, the size of a binary key is 6.4 KB and the size of double precision zero and first order UBM-statistics is 208.9 KB.

(which ignores the magnitude to focus on the directional information), turns out to be decisive in the final performance.

Works about the properties of the conditioning and dimensionality reduction procedures are presented in [1][9][8][7]. But we are now continuing a thorough study of their properties, in order to better explain their impact in the performance and improve further i-vector based speaker verification systems.

## References

1. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19, 788–798 (2011)
2. Reynolds, D.A.: A Gaussian mixture modeling approach to text-independent speaker identification. PhD thesis, Georgia Institute of Technology (1992)
3. Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P., Dumouchel, P.: Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In: International Conference on Speech Communication and Technology, pp. 1559–1562 (2009)
4. Hatch, A.O., Kajarekar, S., Stolcke, A.: Within-Class Covariance Normalization for SVM-based Speaker Recognition. In: International Conference on Speech Communication and Technology, pp. 1471–1474 (2006)
5. Kenny, P.: Bayesian speaker verification with heavy-tailed priors. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2010)
6. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: IEEE International Conference on Computer Vision, pp. 1–8 (2007)
7. Bousquet, P.M., Larcher, A., Matrouf, D., Bonastre, J.F., Plchot, O.: Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)
8. Bousquet, P.M., Matrouf, D., Bonastre, J.F.: Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In: International Conference on Speech Communication and Technology, pp. 485–488 (2011)
9. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of i-vector length normalization in speaker recognition systems. In: International Conference on Speech Communication and Technology, pp. 249–252 (2011)
10. Brummer, N., de Villiers, E.: The speaker partitioning problem. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2010)
11. Campbell, W.M., Sturim, D., Borgstrom, B.J., Dunn, R., McCree, A., Quatieri, T.F., Reynolds, D.A.: Exploring the impact of advanced front-end processing on nist speaker recognition microphone tasks. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)
12. Bonastre, J.F., Bousquet, P.M., Matrouf, D., Anguera, X.: Discriminant binary data representation for speaker recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, pp. 5284–5287 (2011)
13. Bonastre, J.F., Anguera, X., Sierra, G.H., Bousquet, P.M.: Speaker modeling using local binary decisions. In: International Conference on Speech Communication and Technology, pp. 485–488 (2011)
14. Kenny, P.: A small footprint i-vector extractor. In: Speaker and Language Recognition Workshop, IEEE Odyssey (2012)

# Revisiting LBP-Based Texture Models for Human Action Recognition*

Thanh Phuong Nguyen[1], Antoine Manzanera[1],
Ngoc-Son Vu[2], and Matthieu Garrigues[1]

[1] ENSTA-ParisTech, 828, Boulevard des Maréchaux, 91762 Palaiseau, France
[2] LIRIS, INSA Lyon, 20, Avenue Albert Einstein, 69621 Villeurbanne, France

**Abstract.** A new method for action recognition is proposed by revisiting LBP-based dynamic texture operators. It captures the similarity of motion around keypoints tracked by a realtime semi-dense point tracking method. The use of self-similarity operator allows to highlight the geometric shape of rigid parts of foreground object in a video sequence. Inheriting from the efficient representation of LBP-based methods and the appearance invariance of patch matching method, the method is well designed for capturing action primitives in unconstrained videos. Action recognition experiments, made on several academic action datasets validate the interest of our approach.

**Keywords:** action recognition, local binary pattern, dynamic texture,...

## 1 Introduction

Human activity recognition has been an active research topic in recent years due to its interesting application domains such as video surveillance, human computer interaction, video analysis, and so on. Many approaches have been introduced using different video features for action representation, we refer to [1] for a comprehensive survey. However a robust and real time method for action recognition with unconstrained videos is still a difficult challenge.

An interesting approach is to consider the action as a texture pattern, and to apply dynamic or static texture based methods to action modelling and recognition. Thanks to the effective properties of Local Binary Patterns (LBP) for texture representation, several LBP-based methods have also been proposed in the past for action recognition. Kellokumpu et al. [2] used dynamic texture operator (LBP-TOP) to represent human movement. They also presented another approach [3] using classical LBP on temporal templates (MEI and MHI images [4]) that were introduced to describe motion information from images. All extracted features in the two methods are then modelled using HMM (Hidden Markov Model). Mattivi and Shao [5] presented a different method using LBP-TOP to describe cuboids detected by Dollar's feature detector. Recently, Yeffet

* This work is part of an ITEA2 project, and is supported by french Ministry of Economy (DGCIS).

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 286–293, 2013.

and Wolf proposed LTP (Local Trinary Patterns) [6] that combines the effective description of LBP with the adaptivity and appearance invariance of patch matching methods. They capture the motion effect on the local structure of self-similarities considering 3 neighbourhood circles at a spatial position and different instants. Kliper-Gross et al. developed this idea by capturing local changes in motion directions with Motion Interchange Patterns (MIP) [7]. Nanni et al. [8] improved LBP-TOP using ternary units in the encoding step.

In this paper, we revisit dynamic texture based methods for action recognition. We are inspired by 2 popular LBP based representation: uniform LBP for texture coding and LTP for motion coding. We propose a new self-similarity operator to capture spatial relations in a trajectory beam, by representing the similarity of motion between the tracked point along its trajectory, and its neighbourhood. The semi-dense point tracker computes the displacement of many points in real time, then we apply self-similarity operator on appearance information to represent the motion information of a larger zone surrounding the trajectory. Our method can be seen as a hybrid solution between optical flow methods and dynamic texture based approaches. The rest is organised as follows. Section 2 briefly presents the basic material. The next section proposes our approach for action representation. The last sections are experiments and conclusions.

# 2 Basic Materials

## 2.1 LBP Based Operators

**Uniform LBP.** Local Binary Patterns [9] were introduced by Ojala et al. Their idea is to capture the local structures of texture images using binary patterns obtained by comparing a pixel value with its surrounding neighbours. LBP operator has two important properties: it is invariant to monotonic gray scale changes, and its complexity is very low. As a consequence, LBP-based approaches are suitable for many applications, aside from texture recognition. A LBP is called uniform if the number of binary transitions (from 0 to 1, from 1 to 0) while scanning the circle clockwise is at most 2. The uniform pattern coding ($LBP^{u2}_{n,r}$, corresponding to ignoring the non uniform patterns) is widely used in real applications because it reduces significantly the length of feature vectors while capturing important texture primitives (see Fig. 1).

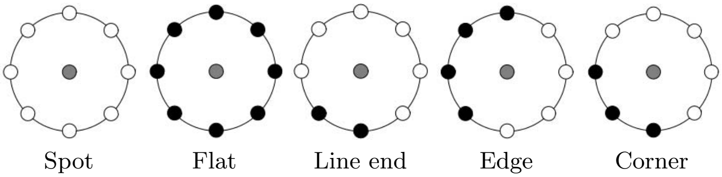
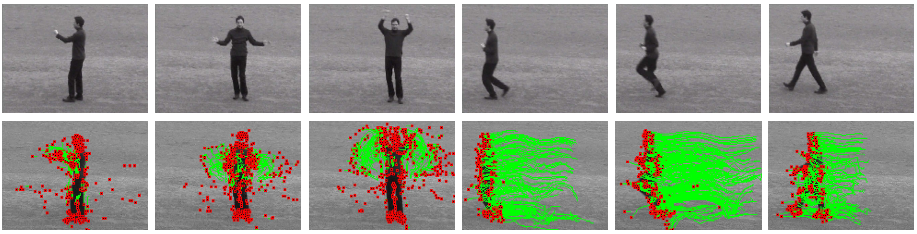


**Fig. 1.** Texture primitives corresponding to Uniform LBPs [9]

**LTP.** Local Trinary Patterns [6] use sum of squared differences (SSD) between patches centred at different space and time locations. Let $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{\Delta_t}$ be the SSD between the patch centred at pixel $\mathbf{x}$ at frame $t$ and the patch centred at pixel $\mathbf{x} + \Delta_{\mathbf{x}}$ at frame $t + \Delta_t$. One ternary code $\{-1, 0, 1\}$ is obtained for each shift direction $\Delta_{\mathbf{x}}$, by comparing $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{-\Delta_t}$ and $\mathrm{SSD}^{\Delta_{\mathbf{x}}}_{+\Delta_t}$.

### 2.2 Motion Representation Using a Beam of Dense Trajectories

Trajectories are compact and rich information source to represent motion in videos, and have been used already for action recognition [10]. Generally, to obtain reliable trajectories, the spatial information is dramatically reduced to a small number of keypoints, and then it may be hazardous to compute statistics on the set of trajectories. In this work we use the semi dense point tracking method [11] (see also Fig. 2) which is a trade-off between long term tracking and dense optical flow, and allows the tracking of a high number of weak keypoints in a video in real time, thanks to its high level of parallelism. Using GPU implementation, this method can handle 10 000 points per frame at 55 frames/s on 640 × 480 videos. In addition, it is robust to sudden camera motion changes.



Boxing Hand clapping Hand waving Jogging Running Walking

**Fig. 2.** Several actions of KTH dataset and their corresponding beam of trajectories. Red points represent tracked particles, green curves describe their trajectories.

## 3 Action Descriptor Using Spatial Motion Patterns

We present now our descriptor for action representation. The input data is the semi-dense trajectory beam described in Section 2. The classic approach to build motion information from optical flow is to consider histogram of optical flow (HOOF). This approach is simple to compute but neglects the spatio-temporal relation between moving points. One popular but limited solution is to consider the extracted histograms in different sub-volumes defined by a spatio-temporal grid. In this section, we introduce a descriptor that addresses more finely this problem. Briefly, the motion information is exploited at different context levels: (1) *Point level*; (2) *Local spatio-temporal level*; (3) *Regional to global spatio-temporal level.* This is detailed hereafter.

### 3.1 Point Level

Let $\overrightarrow{p_t}$ be the 2d displacement of the point between frames $t$ and $t + \delta$. The first part of the encoding is simply a dartboard quantisation of vector $\overrightarrow{p_t}$ (see Fig. 3). In our implementation, we used intervals of $\pi/6$ for the angles and 2 pixels for the norm (the last interval being $[6, +\infty[$), resulting in 12 bins for direction angle, 4 bins for norm.

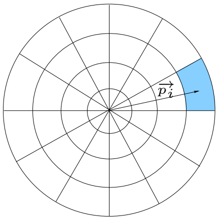


**Fig. 3.** Dartboard quantisation of the motion vector

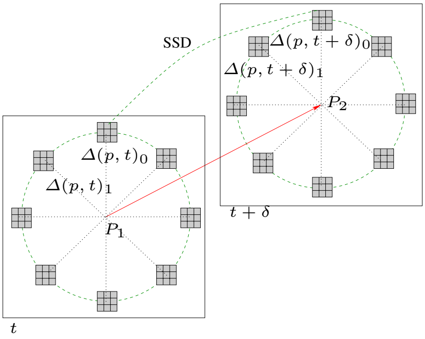


**Fig. 4.** The SMP descriptor is calculated at each tracked keypoint, along its trajectory. The consistency of motion in every direction is checked by computing the SSD between the corresponding image patches.

### 3.2 Local Spatio-temporal Level

At the local spatio-temporal level, we use an LBP-based dynamic texture to capture the relations between a point and its neighbours. Our idea is to capture the inter-trajectory relations among a beam of trajectories. We propose to combine the LBP-based self-similarity operator [9] and the appearance invariance of patch matching method inspired by [6]. This operator, called Spatial Motion Pattern (SMP), is presented below.

**Spatial Motion Patterns**
Consider a point $p$ that moves from position $P_1$ at frame $t$ to position $P_2$ at frame $t + \delta$, provided by the semi dense tracker [11]. The similarity of motion between this point and its neighbours is obtained by considering the $2 \times n$ patches sampled from circles centred at $P_1$ and $P_2$ in their corresponding frames (see Fig. 4). Every index $i \in \{0, n - 1\}$ represents a direction, which is encoded by 0 if the motion in this direction is similar to the motion of the centre point, and by 1 otherwise. Following [6], SSD (sum of square difference) score is used as similarity measure to check the consistency of motion.

Let $\{\Delta(p,t)_i\}_{i=0}^{n-1}$ be the set of $n$ patches surrounding particle $p$ at frame $t$. The corresponding SMP codeword $(b_0, b_1, \ldots, b_{n-1})$ is constructed as follows:

$$b_i = \begin{cases} 1 & \text{If } SSD\big(\Delta(p,t)_i, \Delta(p,t+\delta)_i\big) \geq \tau \\ 0 & \text{otherwise} \end{cases}$$

where $\delta$ is the time interval between two frames, $\tau$ is the SSD threshold. Our local descriptor differs significantly from LTP in several aspects:

- *Encoding process.* Unlike [6], our descriptor uses only 2 bits. The encoding of LTP is done by comparing SSD scores between neighbouring patches of past and future frames, and the centre patch of the middle frame. Our method estimates the SSD scores between two corresponding patches in two consecutive frames.
- *Neighbouring configuration.* LTP used three circles centred at the same position in 2D space. In our approach, the two neighbouring circles are centred at the tracked position of each keypoint.
- *Interpretation.* LTP aims to represent motion information at a given position, whereas in our case, the motion information is already known, the SMP is interpreted as a local disparity map of velocities around each trajectory.

**Properties of Spatial Motion Patterns**

Inheriting from [6, 9], Spatial Motion Patterns have attractive properties:

- *Simple computation.* They use SSD scores on small image patches. In addition, the calculation is only applied on tracked keypoints, not on the whole image, avoiding many irrelevant calculations.
- *Appearance invariance.* This property is due to: (1) the LBP based encoding and (2) the basic information which only relates to the trajectory, not to the appearance.

*SMP uniform patterns ($SMP^{u2}$) captures local primitives action* in a similar way as LBP uniform patterns ($LBP^{u2}$). They detect the motions between foreground objects and the background in videos, and more generally, between two rigid parts of a moving object. We can point out the relation between $SMP^{u2}$ and action primitives as follows (see also Fig. 1).

- *Spot:* A small foreground object move on the background.
- *Flat:* A big rigid part of a moving object.
- *Line end:* End of a thin foreground object.
- *Edge:* Border between two parts of a moving object, or between a foreground object and the background.
- *Corner:* A corner of a rigid part of a moving object.

Fig. 5 illustrates the interpretation of SMP uniform patterns ($SMP^{u2}$).

It is also worth mentioning that, unlike many other methods, the more complex the background, the more efficiently should the SMP describe the rigid parts of the moving object.

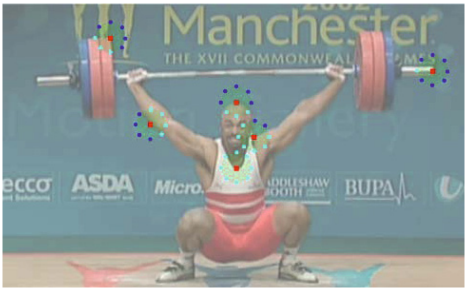


**Fig. 5.** $SMP^{u2}$ configurations allow to determine the shape of the rigid parts of the moving object around the keypoints (in red points). In the neighbouring circles, image patches in green (resp. blue) indicates that they belong to the same rigid part of the moving object as the keypoint (resp. another rigid part or the background).

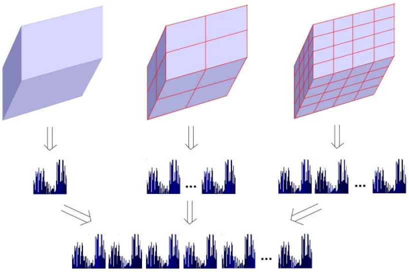


**Fig. 6.** Action modelling by SMP histogram concatenation

### 3.3 Regional to Global Spatio-temporal Level

In this context, a pyramidal bag of feature (BoF) [12] is used to represent action by histograms of codewords made of the two previous primitives (motion code and spatial motion patterns) on spatio-temporal volumes. All histograms are concatenated into one vector that is then normalised for action representation. Fig. 6 shows how to construct the action description using three different grids.

## 4 Experimentation on Human Action Classification

### 4.1 Classification

To perform action classification, we choose the SVM classifier of Vedaldi et al. [13] which approximates a large scale support vector machines using an explicit feature map for the additive class of kernels. Generally, it is much faster than non linear SVMs and it can be used in large scale problems.

### 4.2 Experimentation

We evaluate our descriptor on two well-known datasets. The first one (KTH) [14] is a classic dataset, used to evaluate many action recognition methods. The second one (UCF Youtube) [15] is a more realistic and challenging dataset.

**Parameter Settings.** There are several parameters concerning the construction of SMP. Like [6], we compute SSD score on image patch of size $3 \times 3$ with threshold $\tau = 1000$ that represents 0.17% maximal value of SSD. The time interval $\delta$ is set to 1. Because every tracked keypoint already represents a certain spatial structure, the radius of SMP must be sufficiently large to better capture the geometric shape of rigid parts of moving object surrounding the keypoints. In our implementation, we consider 16 neighbours sampled on a circle of radius 9. In addition, only uniform patterns ($\text{SMP}^{u2}_{16,9}$) are considered. To construct the histograms of codewords, we used 3 spatiotemporal grids: $1 \times 1 \times 1$, $2 \times 2 \times 2$ and $3 \times 3 \times 3$.

**Experimentation on KTH Dataset.** The dataset contains 25 people for 6 actions (running, walking, jogging, boxing, hand clapping and hand waving) in 4 different scenarios (indoors, outdoors, outdoors with scale change and outdoors with different clothes). It contains 599 [1] videos, of which 399 are used for training, and the rest for testing. As designed by [14], the test set contains the actions of 9 people, and the training set corresponds to the 16 remaining persons. Table 1 shows the confusion matrix obtained by our method on the KTH dataset. The ground truth is read by row. The average recognition rate is 93.33 % which is comparable to the state-of-the-art of LBP-based approaches (see Table 2). We remark that unlike [2, 3] that work on segmented box, our results are obtained directly on unsegmented videos. Applying the same pre-processing step would probably improve our result.

**Table 1.** Confusion matrix on KTH dataset

| | Box. | Clap. | Wave | Jog. | Run. | Walk. |
|---|---|---|---|---|---|---|
| Boxing | 97.5 | 2.5 | 0 | 0 | 0 | 0 |
| Clapping | 2.5 | 97.5 | 0 | 0 | 0 | 0 |
| Waving | 2.5 | 0 | 97.5 | 0 | 0 | 0 |
| Jogging | 0 | 0 | 0 | 95.0 | 0 | 5.0 |
| Running | 0 | 0 | 0 | 12.5 | 80.0 | 7.5 |
| Walking | 0 | 0 | 0 | 10.0 | 0 | 90.0 |

**Table 2.** Comparison on KTH dataset

| Method | Result | Method | Result |
|---|---|---|---|
| **Ours** | 93.33 | [6] | 90.17 |
| [3] | 90.8 | [7] | 93.0 |
| [5] | 88.38 | [2] | 93.8 |

**Table 3.** Comparison on UCF Youtube

| **Our method** | [16] | [17] | [15] |
|---|---|---|---|
| 72.07 | 64 | 64 | 71.2 |

**Experimentation on UCF Youtube Dataset.** The UCF Youtube dataset records 11 categories (basketball shooting, cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog), and contains 1 600 video sequences. Each category is divided into 25 groups sharing common appearance properties (actors, background, or other). It is much more challenging than KTH because of its large variability in terms of viewpoints, backgrounds and camera motions. Following the experimental protocol proposed by the authors [15], we used 9 groups out of the 25 as test and the 16 remaining groups as training data. Our mean recognition rate on UCF Youtube dataset is 72.07 % (see Table 3), which outperforms recent methods.

[1] It should contain 600 videos but one is missing.

## 5 Conclusions

We have presented a new method for action recognition based on semi-dense trajectory beam and the LBP philosophy. Its main idea is to capture spatial relation of moving parts around the tracked keypoints, along their trajectories. Our descriptor is designed to capture geometric shape of the rigid parts of moving object in unconstrained videos with complex background. In the future, we are interested in several perspectives related to this method such as multi-scale SMPs, and extension to moving backgrounds.

## References

1. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Comput. Surv. 16, 16:1–16:43 (2011)
2. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. In: BMVC (2008)
3. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Texture based description of movements for activity analysis. In: VISAPP (2), pp. 206–213 (2008)
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. PAMI 23, 257–267 (2001)
5. Mattivi, R., Shao, L.: Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 740–747. Springer, Heidelberg (2009)
6. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: ICCV, pp. 492–497 (2009)
7. Kliper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 256–269. Springer, Heidelberg (2012)
8. Nanni, L., Brahnam, S., Lumini, A.: Local ternary patterns from three orthogonal planes for human action classification. Expert Syst. Appl. 38, 5125–5128 (2011)
9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24, 971–987 (2002)
10. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR, pp. 3169–3176 (2011)
11. Garrigues, M., Manzanera, A.: Real time semi-dense point tracking. In: Campilho, A., Kamel, M. (eds.) ICIAR 2012, Part I. LNCS, vol. 7324, pp. 245–252. Springer, Heidelberg (2012)
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2), pp. 2169–2178 (2006)
13. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. PAMI 34, 480–492 (2012)
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: ICPR, pp. 32–36 (2004)
15. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from video "in the wild". In: CVPR, pp. 1996–2003 (2009)
16. Lu, Z., Peng, Y., Ip, H.H.S.: Spectral learning of latent semantics for action recognition. In: ICCV, pp. 1503–1510 (2011)
17. Bregonzio, M., Li, J., Gong, S., Xiang, T.: Discriminative topics modelling for action feature selection and recognition. In: BMVC, pp. 1–11 (2010)

# A New Triangular Matching Approach for Latent Palmprint Identification

José Hernández-Palancar, Alfredo Muñoz-Briseño, and Andrés Gago-Alonso

Advanced Technologies Application Center (CENATAV), Cuba
{jpalancar,amunoz,agago}@cenatav.co.cu

**Abstract.** Palmprint identification is still considered as a challenging research line in Biometrics. Nowadays, the performance of this techniques highly depends on the quality of the involved palmprints, specially if the identification is performed in latent palmprints. In this paper, we propose a new feature model for representing palmprints and dealing with the problems of missing and spurious minutiae. Moreover, we propose a novel verification algorithm based in this feature model, which uses a strategy for finding adaptable local matches between substructures obtained from images. In experimentation, we show that our proposal achieves highest scores in latent palmprint matching, improving some of the best results reported in the literature.

## 1 Introduction

In the last years, the interest in recognition of persons by their palmprints has grown. There are some scientific studies that ensure the uniqueness of the palmprint of a person and the much fancied stability over time or age [1]. In this sense such technique is much better than others, specially in forensic cases where other biometric information is not available.

Palmprints are marks produced by the contact of the palm of the hand with a surface. These marks reflect the different patterns formed by the ridges that are visible in the epidermis. Most of the verification approaches use minutiae as basis for representing palmprints and checking mutual matches. However, the features extraction is still a challenging problem since the possibility of finding false minutiae always exists [6].

There are not many published articles about the topic, specially in the latent case since palmprint identification is considered as a challenging problem. Until today, few works on latent-to-full palmprint matching have been done. One of the first relevant proposed methods was based on a feature called MinutiaCode [6]. However, this proposal is time consuming and not robust to distortions. Another recent works use radial triangulations in order to extract features [9,10]. Even when the use of radial triangulations increase the accuracy, the features extracted from them are still affected by stretching in the skin. Finally, Jain *et al.* [7] proposed a method based on minutiae clustering and minutiae match propagation.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 294–301, 2013.

One of the most relevant contribution of this paper is to present the results of our palmprint matching algorithm dealing with low quality or distorted palmprint images. It is important to note that unlike other approaches [6,7,9,10], our proposal do not uses any enhancement method in the minutiae extraction process. Our novel matching algorithm uses a representation proposed in literature [5], called expanded triangle set, which is based on minutia triplets obtained from Delaunay triangulation and other redundant ones for reducing the negative effect of structural distortions. Expanded triangle set was previously used for fingerprint indexing and retrieving tasks [5], whereas it is currently used for palmprint matching, in our research. In our matching step, we propose a new strategy to find local matches between substructures formed in the palmprints.

This work is organized as follows. In Section 2 some concepts and definitions necessary to understand our proposal, are described. The Section 3 is dedicated to define our palmprint representation and to describe the process of features extraction. In Section 4, is defined a matching algorithm that uses the features extracted. In Section 5, some experimental results that validate the accuracy of our proposal, are shown. Finally, Section 6 contains the conclusions.

## 2 Background

In this section, we present some basic concepts and a general scheme of palmprint matching algorithms. Thus, we declare the necessary background for understanding our proposal and the rest of the paper. Finally, we describe the Delaunay triangulation and its properties, considering that this kind triangulation is used in many contexts for representing ridges patterns, including our approach.

### 2.1 The Expanded Triangle Set

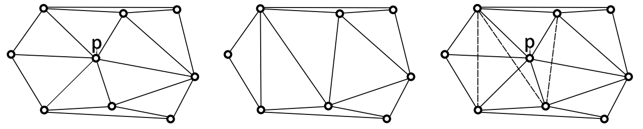
In general, a triangulation of a set of points, $P = \{p_1, p_2, \ldots, p_N\}$, in the plane is the set of triangles that conforms a maximal planar subdivision whose vertex set is $P$. A maximal planar subdivision is a subdivision $S$ such that no edge connecting two vertices can be added to $S$ without destroying its planarity [2]. Especially, a triangulation of $P$ is a Delaunay triangulation if and only if every triangle $\triangle P_i P_j P_k$ that belongs to $T$ satisfies that its circumcircle contains no other point of $P$ [2]. The Delaunay graph of a Delaunay triangulation $T$ is defined as a tuple $G = \langle P, E \rangle$ where $P$ is the set of planar points that originated $T$, and $E$ is the set of edges that conforms the triangles of $T$; each edge has a single occurrence in $E$.

Delaunay triangulations have some theoretical properties, which are very useful for palmprint matching. However, it must be highly affected, when the extraction method fails to find a minutia [5]. For example, in Fig. 1(a), we can see a Delaunay triangulation of a set of points. In Fig. 1(b), we can appreciate major structural changes in the same triangulation when removing the vertex $p$. In literature there is a proposal that introduces an interesting criterion for selecting minutia triplets called expanded triangle set [5], which is defined as follows.

**Definition 1 (Triangular hull).** *Let $p_i$ be a point of $P$. The set $N_i = \{p_j | \{p_i, p_j\} \in E\}$ denoted the point set formed by all the adjacent vertices of $p_i$ in the Delaunay graph $G$. The triangular hull of $p_i$ is defined as the Delaunay triangulation of the planar point set $N_i$, and it is denoted by $H_i$.*

**Definition 2 (Expanded triangle set).** *The expanded triangle set of $P$ is defined as $R = T \cup H_1 \cup H_2 \cup \ldots \cup H_N$.*

The set $R$ includes the triangles in the Delaunay triangulation of $P$ and any triangle in the triangular hulls of the points in $P$. Despite the fact that $|R|$ is greater than $|T|$, the number of triangles of $|R|$ is still linear with respect to $N$ [5]. This is very desirable if we consider that the sets $R$ will be used as a representation for palmprints in verification tasks.



(a) Delaunay triangulation, $T$ (b) Delaunay triangulation without $p_i$, $H_i$ (c) Expanded triangle set, $R = T \cup H_i$

**Fig. 1.** Triangle set examples

The advantage of the set $R$ is that it contains all of the Delaunay triangles that are formed when each minutia is eliminated individually. In this way, we ensure that even when the extraction method fails to find a minutia, some of the matchings will be found. For example, Fig. 1(c) shows the expanded set of the points including $p_i$. As we can see, Fig. 1(c) has corresponding triangles with both, Fig. 1(a) and Fig. 1(b) due to the use of the expanded triangle set. In this paper, the expanded triangle set of minutiae is used for representing palmprints in verification tasks.

### 2.2 Palmprint Matching

In general, we can say that a palmprint matching algorithm compares two palmprints and returns either a degree of similarity or a binary decision. Until today, matching palmprints is still a topic of interest due to the noise and distortions in palm images that can be produced by scars, creases and cuts.

In our case, the palmprints are described as vectors of minutiae were each one can have some attributes. The most commonly used attributes are the coordinates, direction and type of minutiae.

More formally, let $T_1 = \{m_1, m_2, \ldots, m_n\}$ and $T_2 = \{m_1, m_2, \ldots, m_m\}$ be minutia vectors that describe two palmprints, where $m_i = (x_i, y_i)$. In order to obtain a similarity score between $T_1$ and $T_2$, the matching algorithms try to establish similarities between their minutiae. A later step of consolidation consists on computing a global score based on the matches found among minutiae.

## 3 Feature Extraction Step

In this section, we propose a new feature model for representing palmprints, using the expanded triangle set obtained from minutiae, see section 2.1. We are considering that the minutiae extraction process for full palmprints is carried out by any algorithm reported in the state-of-the-art and marked manually for latent palmprints.

Let $P = \{p_1, p_2, \dots, p_N\}$ be the set containing all the planar points representing the minutiae in a palmprint $F$. Let $R$ be the expanded triangle set of $P$, and let $t \in R$ be a triangle, which represents a minutia triplet. Let $m_1 = (x_1, y_1)$, $m_2 = (x_2, y_2)$, and $m_3 = (x_3, y_3)$ be the three points of $t$, with their corresponding planar coordinates, which are sorted in ascending order regarding the length of the opposite side.

The feature vector associated to $t$ in the palmprint $F$ is denoted by $f(t)$, and it is defined as follows

$$f(t) = (s_t, \beta_1, \beta_2, \beta_3, r_1, r_2, r_3, d_1, d_2, d_3), \tag{1}$$

where $s_t$ is the triangle sign, $\beta_i$ are the relative directions of $m_i$ with respect to his opposite side in $t$, $r_i$ are the ridge counter between minutiae and $d_i$ represent the length of the sides of the triangle. These components are formally defined as follows. The twice signed area of $t$ is calculated using the following mathematical expression

$$A_t = x_1(y_2 - y_3) + x_2(y_3 - y_1) + x_3(y_1 - y_2). \tag{2}$$

Using $A_t$, we define the triangle sign of $t$ as $s_t = 0$ if $A_t < 0$; otherwise $s_t = 1$. This feature is invariant to rotation and is included in order to discard possible correspondences between similar palmprints of different hands (left and right).

We define $d_i$ as the Euclidean distance between the corresponding minutiae $m_j$ and $m_k$. Finally, the ridge counter $r_i$ is defined as the number of ridges crossed by the segment joining the pair of minutiae. We verify the statistical behavior presented in [5]; therefore, we also remove from $R$ those triangles with at least one value outside the interval, $0 \leq r_i < 16$.

The feature vectors presented in this section can be represented as a function $f : R \to \Phi$ called feature function, where the set $\Phi = K_1 \times K_4^3 \times K_4^3 \times \mathbb{R}^3$, assuming $K_n = \{0, 1, \dots, 2^n - 1\}$ represents the feature space. Thus, we are able to define the formal representation of a palmprint $F$, which is used in this paper.

**Definition 3 (The feature model).** *Let $F$ be a palmprint. The model of $F$ is defined as a triplet $M = \langle P, R, f \rangle$, where $P$ is the planar point set representing the minutiae of $F$, $R$ is the expanded triangle set of $P$, and $f$ is the function $f : R \to \Phi$, see (1).*

This feature model is used during the matching step for representing the involved palmprints. The described features, combined with the mechanism defined in section 4 to reduce the negative effects of noise, show a good performance when they are used in identification tasks, see section 5.

## 4 Matching Step

In this step, we obtain a similarity value between two models $M_p = \langle P_p, R_p, f_p \rangle$ and $M_q = \langle P_q, R_q, f_q \rangle$. In order to do this, we present the following.

Let

$$f(t_l) = (s_{tl}, \beta_{1l}, \beta_{2l}, \beta_{3l}, r_{1l}, r_{2l}, r_{3l}, d_{1l}, d_{2l}, d_{3l})$$

with $l \in \{q, p\}$, be the two feature vectors of two triangles $t_p \in R_p$ and $t_q \in R_q$, we say that $t_p$ and $t_q$ are corresponding if the following geometric constraints are fulfilled:

$$\begin{aligned} s_{tp} &= s_{tq}, \\ |\beta_{ip} - \beta_{iq}| &\leq \delta_\beta, \\ |r_{ip} - r_{iq}| &\leq \delta_r, \\ |d_{ip} - d_{iq}| &\leq \delta_d, \end{aligned} \tag{3}$$

for all $i \in \{1, 2, 3\}$, where $\delta_\beta$, $\delta_r$, and $\delta_d$ are predefined thresholds empirically set to 3.

Let $t_p(m_{1p}, m_{2p}, m_{3p})$ and $t_q(m_{1q}, m_{2q}, m_{3q})$ be two corresponding triangles. We define their correlation tuples as $ct_i = (\alpha_i, \overline{m_{ip}m_{jp}}, \overline{m_{iq}m_{jq}})$ with $j = 1$ if $i = 3$ and $j = i + 1$ otherwise; were $\alpha_i$ represents the normalized difference between the i-th interior angles of $t_p$ and $t_q$, and $\overline{m_{ip}m_{jp}}, \overline{m_{iq}m_{jq}}$ are segments of the triangles. Interior angle is defined as the angle inside two adjacent sides of a triangle.

The process followed to obtain the value of $\alpha_i$, is very similar to that presented by Chikkerur *et al.* [3], to obtain the similarity between an edge that connects two minutiae of an impression and one edge joining two minutiae of other fingerprint.

Let $R_p$ and $R_q$ be two triangles sets, we define the set $T(R_p, R_q) = \{ct_1, ct_2, \ldots, ct_n\}$ as the union of all the correlation tuples of every corresponding triangle between $R_p$ and $R_q$.

In our matching step, we use a reduced set $Tr(R_p, R_q) = \{ct_1, ct_2, \ldots, ct_r\}$ that contains only the correlation tuples whose value of $\alpha_i$ are equal to the statistic mode in $T(R_p, R_q) = \{ct_1, ct_2, \ldots, ct_n\}$, for the values of $\alpha_i$ of every $ct_i$. The main goal of this process is finding the most probable value of relative rotation between the matched models and using only the correlation tuples that are consequent with this.

With the reduced set $Tr(R_p, R_q)$ we construct a similarity graph $G_s = \langle V, E, L, s, l \rangle$ where $s : E \to \mathbb{R}$ is a similarity function that assign a value to every edge, $l : P_i \times P_j \to L$ is a labeling function given two vertices and $L$ is a set of vertex labels. $s$ is a similarity function that represents in fuzzy terms the grade of closeness between the two segments $\overline{m_{ip}m_{jp}}$ and $\overline{m_{iq}m_{jq}}$ that originated a edge in $G_s$. Similar functions had been used in other fingerprint recognition approaches [3,4].

In algorithm 1, the generation of $G_s$ is described. For each $ct_i = (\alpha_i, \overline{m_{ip}m_{jp}}, \overline{m_{iq}m_{jq}}) \in Tr$ two vertices that represent the mutual match between $m_{ip} \Longleftrightarrow m_{iq}$ and $m_{jp} \Longleftrightarrow m_{jq}$, are generated. If these vertices are not in $V$, they are added. Also, a new edge that represents the mutual match between segments $\overline{m_{ip}m_{jp}}$ and $\overline{m_{iq}m_{jq}}$ is added to $E$. In this way, we have a graph that represents matches between points of the models $M_p$ and $M_q$, weighted with a similarity function. The graph $G_s$ may be not connected.

**Algorithm 1.** Generating similarity graph

```
Input: $Tr(R_i, R_j) = \{ct_1, ct_2, \ldots, ct_r\}, l$
Output: $(V, E)$ - $V$ and $E$ of similarity graph
foreach $ct_i = (\alpha_i, \overline{m_{ip}m_{jp}}, \overline{m_{iq}m_{jq}}) \in Tr$ do
    $u \leftarrow l(m_{ip}, m_{iq})$; - Creating two new vertices
    $v \leftarrow l(m_{jp}, m_{jq})$;
    if $u \notin V$ then
        $V = V \cup \{u\}$; - Adding $u$ if is not included yet
    end
    if $v \notin V$ then
        $V = V \cup \{v\}$; - Adding $v$ if is not included yet
    end
    $e \leftarrow (u, v)$; - Creating new edge
    $E = E \cup \{e\}$; - Adding new edge
end
return $(V, E)$;
```

In order to find the spanning tree of every connected components of $G_s$ with the higher value of similarity in their edges, we applied the Kruskal algorithm to $G_s$. This is a well known method to find a minimum (or maximum in our case) spanning forest of disconnected graphs. Unlike the proposal presented by Zhu *et al.* [11] based on the Prim algorithm, our solution is superior and it has not been reported in previous works.

Let $\{F_1, F_2, \ldots, F_n\}$ be the set of spanning trees returned by the Kruskal algorithm, sorted in descending order by the amount of edges. We implement a strategy to merge $F_1$ and $F_2$ by trying to add a virtual edge $e_v$ between then. This virtual edge must complain with some geometric restrictions. If this process is successful then $F_1 = F_1 \cup F_2 \cup \{e_v\}$, $F_2$ is eliminated and $F_{i-1} \leftarrow F_i, \forall\, i, 3 < i < n$, $n \leftarrow n - 1$. This process is repeated while $F_1$ and $F_2$ can be merged.

Finally, the similarity value between between models $M_p = \langle P_p, R_p, f_p \rangle$ and $M_q = \langle P_q, R_q, f_q \rangle$ is given by the following expression:

$$similarity(M_p, M_q) = \frac{sim \times |V|}{min(|P_p|, |P_q|)} \tag{4}$$

where $|V|$ is the number of vertices in the similarity graph $G_s$, $|P_p|$ and $|P_q|$ are the cardinalities of $P_p$ and $P_q$ respectively, and $sim$ is the sum of the weights of every edge of $F_1$.

## 5 Experimental Results

In our experiments, a dataset in which some other approaches were tested [6,9,10], was used. This dataset is conformed by 22 latent palmprints from real forensic cases and 8680 full palmprints from criminal investigation field, captured by Beijing Institute of Criminal Technology in China. All the impressions have a resolution of 500 ppi. In the case of latent palmprints, the minutiae were manually extracted by forensic chinese experts. On the other hand, the minutiae of the 8680 full palmprints were extracted automatically using the VeriFinger 4.2 [8], and we did not used any palmprint enhancement process.

**Table 1.** Comparison results of identification rate

| Algorithm | Identification Rate | | |
|---|---|---|---|
| | Rank-1 | Rank-10 | Rank-20 |
| Jain and Feng [6] | 67 % | 73 % | 80 % |
| Wang *et al.* [9] | 63 % | 68 % | 72 % |
| Wang *et al.* [10] | 69 % | 78 % | **82 %** |
| Our proposal | **77 %** | **82 %** | **82 %** |

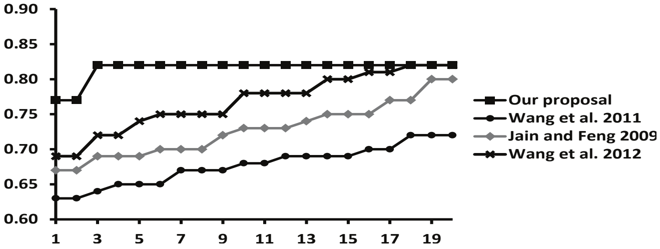


**Fig. 2.** Comparison of methods using CMC curves

Using the described dataset we compared our latent-to-full palmprint verification algorithm with three other proposals found in the state-of-the-art.

Comparison results of identification rate are shown in Table 1. For each identification rate, the higher reached value is highlighted in bold. As we can see, in almost every case our method outperforms the other proposals. In Figure 2, CMC curves of our algorithm and other state-of-the-art methods are shown. In this graphic, the higher accuracy of our proposal is evidenced,especially for rank-1, where we have eight percentage points over the best second algorithm. This same algorithm achieves the 82.2% of identification rate at rank-18, while our algorithm obtains the same value at rank-3.

## 6 Conclusions

Many of the reported palmprint matching algorithms are highly affected when they are used to compare low-quality and distorted images or latent palmprints captured at uncontrolled context. The feature model and the novel matching algorithm, proposed in this paper, can be considered as a promising approach for palmprint identification in such context, dealing with the problems of missing and spurious minutiae, and other noises. Experimental results show that our proposal achieves high accuracy in latent palmprint matching tasks, outperforming other state-of-the-art proposals. We did not use any preprocessing or enhancement method to reduce the number of false minutiae that an automatic extraction process usually has.

## References

1. Ashbaugh, D.R.: Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology. Practical Aspects of Criminal and Forensic Investigation Series. CRC Press Inc. (1999)
2. Berg, M., Krevelt, M., Overmars, M., Scharzkopf, O.: Computational Geometry (Algorithms and Applications). Springer, Heidelberg (1997)
3. Chikkerur, S., Cartwright, A.N., Govindaraju, V.: K-plet and coupled bfs: A graph based fingerprint representation and matching algorithm. In: International Conference on Biometrics (ICB 2006), pp. 309–315 (2006)
4. Fu, X., Liu, C., Bian, J., Feng, J., Wang, H., Mao, Z.: Extended Clique Models: A New Matching Strategy for Fingerprint Recognition. In: International Conference on Biometrics, ICB 2013 (2013)
5. Gago-Alonso, A., Hernández-Palancar, J., Rodríguez-Reina, E., Muñoz-Briseño, A.: Indexing and retrieving in fingerprint databases under structural distortions (2013), doi:10.1016/j.eswa.2012.12.004
6. Jain, A.K., Feng, J.: Latent Palmprint Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(6), 1032–1047 (2009)
7. Liu, E., Jain, A.K., Tian, J.: A Coarse to Fine Minutiae-Based Latent Palmprint Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013), doi:10.1109/TPAMI.2013.39
8. Neurotechnologija Verifinger 4.2 SDK (2004), http://www.neurotechnologija.com/vfsdk.html
9. Wang, R., Ramos, D., Fiérrez, J.: Latent-to-full palmprint comparison based on radial triangulation under forensic conditions. In: International Joint Conference on Biometrics (IJCB 2011), pp. 1–6. IEEE (2011)
10. Wang, R., Ramos, D., Fiérrez, J.: Improving Radial Triangulation-based Forensic Palmprint Recognition According to Point Pattern Comparison by Relaxation. In: International Conference on Biometrics (ICB 2012), pp. 427–432 (2012)
11. Zhu, E., Hancock, E., Ren, P., Yin, J., Zhang, J.: Associating minutiae between distorted fingerprints using minimal spanning tree. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010, Part II. LNCS, vol. 6112, pp. 235–245. Springer, Heidelberg (2010)

# Fusion of Multi-biometric Recognition Results by Representing Score and Reliability as a Complex Number

Maria De Marsico[1], Michele Nappi[2], and Daniel Riccio[3]

[1] Sapienza University of Rome, Rome – Italy
demarsico@di.uniroma1.it
[2] University of Salerno, Fisciano (SA) – Italy
mnappi@unisa.it
[3] University of Naples Federico II, Napoli – Italy
daniel.riccio@unina.it

**Abstract.** A critical element in multi-biometrics systems, is the rule to fuse the information from the different sources. The component sub-systems are often designed to further produce indices of input image quality and/or of system reliability. These indices can be used as weights assigned to scores (weighted fusion) or as a selection criterion to identify the subset of systems that actually take part in a single fusion operation. Many solutions rely on the estimation of the joint distributions of conditional probabilities of the scores from the single subsystems. The negative counterpart is that such very effective solutions require training and a high number of training samples, and also assume that score distributions are stable over time. In this paper we propose a unified representation of the score and of the quality/reliability index that simplifies the process of fusion, provides performance comparable to those currently offered by top performing schemes, yet does not require a prior estimation of score distributions. This is an interesting feature in highly dynamic systems, where the set of relevant subjects may undergo significant variations across time.

**Keywords:** Reliability, unified value score-reliability, complex numbers.

## 1 Introduction

Multi-biometric systems [16] are considered as one of the best viable solutions to overcome limitations of classical single biometrics, since flaws of one sub-system may be balanced by strengths of a companion one. Among the most relevant issues raised by the combined approach, we mention the need for an effective fusion strategy of the results. Information fusion in a biometric system can be performed at feature, score, or decision level [6], but most schemes in literature opt for score level fusion [5]. Score normalization is one of the important aspects to consider during fusion. Fusion schemes may also rely on treating scores as a unified feature vector, which requires a further classifier, or on transforming the scores in a posteriori probabilities [10]. A further issue is represented by the introduction of quality measures computed for the input samples [7][8] and of confidence margins [10]. The former (e.g. sharpness, lighting) allows to

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 302–309, 2013.

possibly discard problematic samples, but can also be exploited after classification, as a weight on the final obtained score. The latter can be used after classification to decide whether to trust in the system response. Two trends are currently developing, to take them into account. In the first one, all subsystems always participate in the fusion, and the quality is used to weight their responses. In the second one, only a subset of subsystems takes part from time to time in the fusion, which are selected according to reliability of their responses. In both cases, reliability measure is an additional information, and mostly handled as a separate value.

Among the many simple score fusion rules (e.g. sum, weighted sum, product, min, max) [10], a number of authors claim that simple sum is the best compromise between simplicity and performance. On the other hand, significantly better results can be obtained through more complex techniques [1]. Likelihood Ratio (LR) is one of the most interesting ones. The authors of [17] discuss how product of Likelihood Ratios represents an optimal rule to get the highest Genuine Accept Rate (GAR) for a fixed False Accept Rate (FAR) in a multi-biometric system. The main disadvantage of this rule of fusion is that it assumes an accurate estimate of the joint distribution (across all the subsystems) of the conditional probabilities of the scores achieved by genuine and impostors users. This requires a complex modeling phase (in [12] finite Gaussian Mixture Model - GMM is used to model the genuine and impostor score densities), and a significant number of training samples. Despite such complexity, performance of systems whose operational parameters are based on a preliminary estimation of score distributions, may degrade if these significantly change along time. Nevertheless, given the optimality of LR, it can be considered as an asymptotic limit for which to strive when devising a new rule of fusion, while trying to overcome its limitations.

This work proposes a novel way of assembling the recognition score and the response reliability measure into a single complex number, facilitating the fusion in identification operations. The technique used in [18] maps the feature vectors from two biometric systems into the real and imaginary part of a complex vector. We rather use the score and the reliability, associated with an identification result by a single subsystem, to derive the module and the anomaly in the exponential representation of a complex number. The fusion of results related to the same identity relies on a modified operation of complex product among the responses from the single subsystems returning such identity. Further processing detailed in Section 2 allows to obtain a single real value as the final score assigned to a given identity by the global system. We use the *System Response Reliability* (SRR) measure [4], which does not require training, and is able to provide reliability information for each single recognition operation, differently from aggregate values like Recognition Rate.

## 2 Merging Scores and Reliability Values by Complex Numbers

There is a major difference between a quality measure for an input sample and a reliability measure for the response of a biometric system. The former is generally bound to a specific biometrics and to a specific classifier: for instance, a measure based on the quality of minutiae only applies to fingerprint recognition, which specifically uses minutiae for classification. Reliability measures devised without any reference to a specific biometric trait and/or algorithm can be generally used for any recognition system. The biometrics-independent reliability measure that we exploit takes into account the composition of the gallery of the recognition system. From now on, we will use the *System Response Reliability* (SRR) [4] as a measure of reliability. The SRR relies on different versions of function $\varphi$ defined in [4], which respectively exploit the relative distance and the density ratio, as well as a combination of them. All three functions measure the amount of "confusion" among possible candidates. We assume that the result of an identification operation is the whole gallery ordered by distance from the probe. Given a probe *p* and a system *A* with gallery *G*, the first function is:

$$\varphi_1(p) = \frac{d(p, g_{i_2}) - d(p, g_{i_1})}{d(p, g_{i_{|G|}})}, \tag{1}$$

where *d* is a distance function with codomain [0, 1], $g_{i_k}$ is the k-th identity in the returned gallery ordering, and |*G*| is the size of the gallery; distance values falling in a different codomain can be suitably normalized. Here we use the *Quasi Linear Sigmoidal (QLS)* [4]. It better preserves the original distribution of data, and is robust to a missing reliable evaluation for the maximum value. With relative distance if a person is genuine, there is a great difference between the distance from the first retrieved identity and the immediately closest one. Density ratio is instead defined as:

$$\begin{gathered} \varphi_2(p) = 1 - |N_b| / |G|, \\ \text{where} \quad N_b = \{ g_{i_k} \in G \mid d(p, g_{i_k}) < 2 \cdot d(p, g_{i_1}) \}. \end{gathered} \tag{2}$$

The formula considers the distinct identities returned during identification as a cloud centred in *p*; the higher the density of this cloud, the more unreliable is the answer, as there are many individuals as potential candidates. In this paper we also adopt a variation of the density ratio. As one can observe in the definition of $N_b$ in (2), the radius of the considered cloud depends on the distance from the probe of the first returned identity and from a constant. This function is less sensible to outliers, than $\varphi_1$, but it considers narrower clouds when the first retrieved identity is closer to the probe. On the contrary, a large distance takes to a wider cloud, which can be expected to be more crowded anyway. To further improve $\varphi$, we define here the term $N_c$ such that the cloud radius depends on the difference between the first two distances:

$$\begin{gathered} \varphi_3(p) = 1 - |N_c| / |G|, \\ \text{where} \quad N_c = \left\{ g_{i_k} \in G \mid d(p, g_{i_k}) < \frac{(1 + d(p, g_{i_2}))(1 + d(p, g_{i_2}) - d(p, g_{i_1}))}{4} \right\}. \end{gathered} \tag{3}$$

The new radius increases with the second distance, and with the difference between the first and the second ones. In practice, the farthest the second returned subject from the probe, also with respect to the first one, the wider the cloud we inspect. However, being all distances in [0,1], we add 1 to both terms to maintain direct proportionality. We also use the appropriate normalization factor since the value of *d* is in [0,1], and the maximum value for the numerator in (3) is *4*.

Once chosen the function $\varphi$ to use, some more steps are required to compute the value of SRR for the probe at hand. For each $\varphi(p)$, we identify a value $\overline{\varphi}$ fostering a correct separation between genuine and impostor subjects. We also define $S(\varphi(p),\overline{\varphi})$ as the width of the subinterval from $\overline{\varphi}$ to the proper extreme of the overall [0,1) interval of possible values, depending on the comparison between the current $\varphi(p)$ and $\overline{\varphi}$:

$$S\left(\varphi(p),\overline{\varphi}\right)=\begin{cases}1-\overline{\varphi} & if \quad \varphi(p)>\overline{\varphi} \\ \overline{\varphi} & otherwise\end{cases} . \tag{4}$$

SRR index can finally be defined as:

$$SRR=(\varphi(p)-\overline{\varphi})/S(\overline{\varphi}) \cdot \tag{5}$$

In detail, we measure the distance between $\varphi(p)$ and the "critical" point $\overline{\varphi}$, which gets higher values for $\varphi(p)$ much higher than $\overline{\varphi}$ (genuine), or for $\varphi(p)$ much lower than $\overline{\varphi}$ (impostors). However, it is also important to take into account how much it is significant with respect to the subinterval over which it is measured. SRR gets values in [-1, 1]. More details on computation and its motivations can be found in [4].

Numbers in the complex field can be represented as $a+ib$ ,or by the exponential representation $z=\rho\cdot e^{i\theta}$, where $\rho$ is the modulus and $\theta$ is the anomaly. In our fusion, the score and the reliability measure are used to derive the $\rho$ and the $\theta$ of this latter representation, respectively. We chose this representation because it better adapts to the kind of processing for fusion. In fact, the product operation with the real/imaginary form, would suffer from misleading cross-influence between heterogeneous parts. Given a score $s$ and a reliability value $srr$, $\rho=(1+s)\cdot$and $\theta=srr$. Since $s$ is in the interval [0,1], and $srr$ ranges between -1 and 1, then $\rho$ is in the interval [1, 2] and $\theta$ is in [-1,1]. We take the set of the complex numbers obtained in this way from the values returned by the different subsystems voting for the same identity in a multibiometric identification. We define a new operation over them that we denote with ⊗, such that:

$$z=\bigotimes_{j=1}^{k} z_j=\frac{\prod_{j=1}^{k}\rho_j}{k}e^{i\frac{\sum_{j=1}^{k}\theta_j}{k}} \tag{6}$$

Thanks to the denominators, the final $\rho_{\otimes}$and $\theta_{\otimes}$ are still in the same intervals as the initial values. The final composed score will be $s_{\otimes}=(\rho_{\otimes}/2)$ and the final reliability will be $srr_{\otimes}=\theta_{\otimes}$, and will be respectively in the interval [0,1] and [-1,1]. In the absence of a reliability measure, its value can be set to 1 for any response. The two values after fusion can again be used to obtain the exponential form of a complex number. This can be done for each group of subsystems voting for a same identity, so that at the end we will have a complex numbers for each candidate identity. However, we have to choose a winning identity, so we would prefer single and easier to compare values. To this aim, we first pass to the representation of the complex numbers in real and imaginary part $z=a+ib$, with $a,b\in$ R. The ($a$, $b$) pair can be interpreted as a couple of coordinates in a two-dimensional space, and as such can be represented in the Argand-Gauss plane (especially devised to represent complex numbers in this form).

**Fig. 1** presents an example of the effects of the approach using *SRR* defined above with $\varphi_I$. The values $a+ib$ in the plots refer to the half-plane with positive x-axis (real part a, derived from scores). The first three plots represent pairs of points for the same classifier on three different biometrics (face, ear and iris). Section 4 reports details on datasets and classifiers. We see that genuine scores (red/light circles) are mainly concentrated in the first quadrant, while impostor scores (blue/dark squares) mainly lie in the fourth quadrant, with some overlap. The last plot is the result of the introduced operations over these values. Notice that the values for genuine users are distributed in the positive quadrant, while those for the impostors are concentrated in the negative one, but the interesting feature to notice is that values are much more sharply divided.

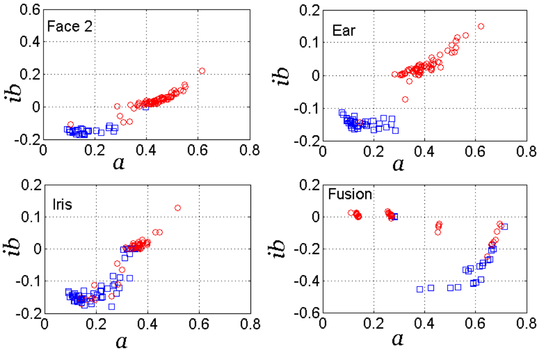


**Fig. 1.** First three plots: distribution of pairs real/imaginary parts obtained from the responses of a correlation based classifier (see below) over face, ear and iris datasets (left-right and top-bottom); last plot: the distribution after the product. Red/light circles are genuine scores, blue/dark squares are impostor scores.

As coordinates in a 2D space, (*a, b*) pairs can be further transformed in single values using Peano keys. Peano rule maps a 2D onto a 1D space such that two close points in the starting space, tend to be close also in the final one. However, the rule requires integer values, so that it is necessary to consistently map *a* and *b* onto integers with a finite number *n* of bits. In our implementation the new integers $a_P$ and $b_P$ have $n = 16$ bits. The associated Peano key $K_P$ is obtained by interleaving bits from $a_P$ and $b_P$, from the least significant to the most significant one, so to obtain a final value of 32 bits. Values for different identities can be straightforwardly compared.

## 3 Experimental Framework

The presented framework was tested in a multi-biometric setting (face, ear and iris) and compared with the LR discussed in [12], using the same implementation for the estimation of the GMM model. The multi-biometric database consists of Chimeric users whose biometric traits were taken from three different datasets. It is worth noticing that  it is presently accepted that results obtained in this way are worthy of full reliability [9]. The number of subjects in the database is constrained by the size of the database of ears, namely 100 subjects in the Notre Dame Ear Database [13]. In order to consider an open set identification setting, i.e. a situation where not all users are enrolled and impostors can also occur, the gallery consists of 75 enrolled subjects, for which there is a single image, while the probe is made up of 100 subjects, each accompanied by a single image. The faces are from a subset of AR-Faces database [14] (50 males and 50 females), for which 4 different datasets were considered: gallery (normal), Face-2 (smile), Face-5 (left-light) and Face-11 (scarf). The irises were from the first 100 subjects in UBIRISv1s1 database [15]. Performance were measured in terms of Recognition Rate (RR) and Equal Error Rate (EER) [2].

In order to understand the relation between the behavior of the presented framework and the classifier used, we tested it with Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and the local correlation-based classifier which is part of FACE system [3], indicated from now on as FACE for short. **Table 1** shows the performance on each dataset, which appear quite heterogeneous, as expected. This is interesting to understand later how the fusion technique works, not only when all classifiers provide optimal results, but also when one or more of them fail.

**Table 1.** Performance of single classifiers on each dataset, in terms of RR and EER

| **Dataset** | **PCA** | | **LDA** | | **FACE** | |
|---|---|---|---|---|---|---|
| | **RR** | **EER** | **RR** | **EER** | **RR** | **EER** |
| Face-2 | 0.97 | 0.039 | 0.94 | 0.027 | 0.97 | 0.052 |
| Face-5 | 0.21 | 0.144 | 0.61 | 0.124 | 0.98 | 0.013 |
| Face-11 | 0.04 | 0.441 | 0.05 | 0.354 | 0.93 | 0.053 |
| Ear | 0.65 | 0.207 | 0.76 | 0.091 | 0.85 | 0.120 |
| Iris | 0.69 | 0.092 | 0.74 | 0.093 | 0.62 | 0.185 |

Results in **Table 1** show that PCA and LDA are much more sensible to local variations within a face image. In particular on the Face 11 set, where the lower part is completely occluded by a scarf. In combinating with other biometrics, this condition may be particularly stressing for the fusion process, making this case very interesting. In the first experiment, we tested the best function $\varphi$. The same classifier was applied to the different biometrics and the reliability was measured from time to time by a different $\varphi$. Given the score $s_j$ (as an inverse of distance from the probe) from biometrics $j$ (F=face, where F2, F5 and F11 indicate the datasets from AR-Faces, E=ear and I=iris), and given $srr_j$ its reliability value, according to the chosen $\varphi$, Complex Fusion (CF) computes the presented operation for the three $(1+sj)ei^{\cdot srrj}$. *Simple sum* rule was also tested, and results were comparable to those of complex values with no reliability, i.e. with the imaginary part set to 1 (CF none). For sake of space, **Table 2** only reports the results of FACE, which resulted better than PCA and LDA classifiers.

**Table 2.** RR and EER, when different $\varphi$ functions are used in fusion of FACE results

| **Method** | **F2/E/I** | | **F5/E/I** | | **F11/E/I** | |
|---|---|---|---|---|---|---|
| | **RR** | **EER** | **RR** | **EER** | **RR** | **EER** |
| **Simp. sum** | 1.00 | 0.026 | 1.00 | **0.001** | 1.00 | 0.067 |
| **Simp. prod** | 1.00 | 0.026 | 1.00 | 0.001 | 1.00 | 0.060 |
| **CF none** | 1.00 | 0.046 | 1.00 | 0.033 | 0.97 | 0.039 |
| **CF $\varphi_1$** | **1.00** | **0.020** | **1.00** | 0.006 | **1.00** | **0.033** |
| **CF $\varphi_2$** | 1.00 | 0.246 | 1.00 | 0.153 | 0.99 | 0.342 |
| **CF $\varphi_3$** | 0.98 | 0.039 | 1.00 | 0.033 | 0.94 | 0.061 |

**Table 2** shows that $\varphi_1$ and $\varphi_3$ give the best results and will be used to compare performance of Simple product, Complex fusion e Likelihood ratio. Values in **Table 2** highlight that the simple sum provides acceptable results, when the classifier offers good performance for every fused biometrics. However, it was observed that, having all biometrics the same weight, if one of them provides poor results this significantly influences the overall system performance, as confirmed by the results in **Table 3**, where each fusion technique is evaluated with all classifiers.

**Table 3.** Performance when different techniques are used for fusion, in terms of RR and EER

| **PCA** | **F2/E/I** | | **F5/E/I** | | **F11/E/I** | |
|---|---|---|---|---|---|---|
| | **RR** | **EER** | **RR** | **EER** | **RR** | **EER** |
| **Simple sum** | 1.00 | 0.073 | 1.00 | 0.112 | 1.00 | 0.278 |
| **Complex Fusion ($\varphi_1$)** | 0.99 | 0.420 | 0.72 | 0.329 | 0.65 | 0.560 |
| **Complex Fusion ($\varphi_3$)** | 1.00 | 0.326 | 0.74 | 0.333 | 0.64 | 0.470 |
| **Likelihood Ratio** | 1.00 | 0.033 | 0.95 | 0.140 | 0.85 | 0.214 |
| **LDA** | **F2/E/I** | | **F5/E/I** | | **F11/E/I** | |
| | **RR** | **EER** | **RR** | **EER** | **RR** | **EER** |
| **Simple sum** | 1.00 | 0.040 | 0.96 | 0.120 | 0.84 | 0.171 |
| **Complex Fusion ($\varphi_1$)** | 0.99 | 0.427 | 0.86 | 0.170 | 0.73 | 0.359 |
| **Complex Fusion ($\varphi_3$)** | 0.99 | 0.118 | 0.84 | 0.160 | 0.77 | 0.181 |
| **Likelihood Ratio** | 1.00 | 0.040 | 0.99 | 0.112 | 0.95 | 0.171 |
| **FACE** | **F2/E/I** | | **F5/E/I** | | **F11/E/I** | |
| | **RR** | **EER** | **RR** | **EER** | **RR** | **EER** |
| **Simple sum** | 1.00 | 0.026 | 1.00 | 0.001 | 1.00 | 0.067 |
| **Complex Fusion ($\varphi_1$)** | 1.00 | 0.020 | 1.00 | 0.006 | 1.00 | 0.033 |
| **Complex Fusion ($\varphi_3$)** | 0.99 | 0.039 | 1.00 | 0.033 | 0.93 | 0.061 |
| **Likelihood Ratio** | 1.00 | 0.010 | 1.00 | 0.000 | 1.00 | 0.013 |

In **Table 2** and **Table 3** $\varphi_1$ provides the best results with a classifier robust to variations, like FACE. On the contrary, e.g., with PCA and partly with LDA, $\varphi_3$ sometimes provides better results. **Table 3** shows that, in many cases for PCA and LDA, complex fusion performance is below simple sum. This is because these two algorithms are both poorly robust to distortions, and provide poorly reliable responses. In fact, we would observe a wide overlap between genuine and impostor distributions. With FACE classifier we achieve both higher robustness, and higher reliability. The latter makes the fusion results with complex numbers better than those with simple sum, especially with function $\varphi_1$. The overall interesting aspect is that, using a robust classifier aligned with the state of the art, the proposed fusion technique is able to provide better results that simple sum and only slightly lower that the optimum LR. This is very important if we consider that it is simple like the sum, yet does not require any preliminary estimation of genuine and impostor score distributions. In other words, at the expense of a slightly lower performance, we are able to adopt a strategy which is stable over time and delivers results which are congruous for each single probe, we avoid an expensive training phase, and save computation even in operational phases.

## 4 Conclusions

This paper has presented a multi-biometric fusion framework based on the joint representation in the complex field of score values and reliability measures. The experimental results show that in the case of robust classifiers the performance of the proposed framework are comparable to those of LR, which proves to be the best criterion for fusion. The product of complex values, however, has the further advantage of not needing an accurate approximation of the distributions of the scores. Future studies will focus on even better criteria to use the complex representation.

## References

1. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D Face Recognition: A Survey. Pattern Recognition Letters 28(14), 1885–1906 (2007)
2. Bolle, R.M., Connell, J.H., Pananti, S., Ratha, N.K., Senior, A.W.: The Relation Between the ROC Curve and the CMC. In: Proc. of 4th IEEE Work. on Automatic Identification Adv. Technologies, pp. 15–20 (2005)
3. De Marsico, M., Nappi, M., Riccio, D.: FACE: Face analysis for commercial entities. In: Proc. of Int. Conference on Image Processing (ICIP), Honk Kong, pp. 1597–1600 (2010)
4. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: NABS: Novel Approaches for Biometric Systems. IEEE Trans. on Systems, Man, and Cybernetics–Part C: Applications and Reviews 40(6) (2011)
5. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recognition Letters 24(13), 2115–2125 (2003)
6. Jain, A.K., Nandakumar, K., Ross, A.: Score Normalization in multimodal biometric systems. Pattern Recognition 38(12), 2270–2285 (2005)
7. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J.: Discriminative Multimodal Biometric Authentication Based on Quality Measures. Pattern Recognition 38(5), 777–779 (2005)
8. Nandakumar, K., Chen, Y., Jain, A.K., Dass, S.: Quality-Based Score Level Fusion in Multibiometric Systems. In: Proc. Int'l Conf. Pattern Recognition, pp. 473–476 (August 2006)
9. Garcia-Salicetti, S., Mellakh, M.A., Allano, L., Dorizzi, B.: A Generic Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 494–502. Springer, Heidelberg (2005)
10. Kittler, J., Hatef, M., Duin, R.P., Matas, J.G.: On Combining Classifiers. IEEE Trans. PAMI 20(3), 226–239 (1998)
11. Poh, N., Bengio, S.: Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In: Proc. Fifth Int'l Conf. Audio Video-Based Biometric Person Authentication, pp. 474–483 (July 2005)
12. Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K.: Likelihood Ratio-Based Biometric Score Fusion. IEEE Transactions on PAMI 30(2), 342–347 (2008)
13. Notre Dame Ear Database, `http://www.nd.edu/~cvrl/UNDBiometricsDatabase.html`
14. Martinez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. PAMI 24(6), 748–763 (2002)
15. Proença, H., Alexandre, L.A.: UBIRIS: A noisy iris image database. In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 970–977. Springer, Heidelberg (2005)
16. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer (2006)
17. Ulery, B., Hicklin, A.R., Watson, C., Fellner, W., Hallinan, P.: Studies of Biometric Fusion. Technical Report IR 7346, NIST (September 2006)
18. Yang, J., Yang, J., Zhang, D., Lua, J.: Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition 36(6), 1369–1381 (2003)

# Fusion of Iris Segmentation Results

Andreas Uhl and Peter Wild

Multimedia Signal Processing and Security Lab.
Department of Computer Sciences, University of Salzburg, Austria
{uhl,pwild}@cosy.sbg.ac.at

**Abstract.** While combining more than one biometric sample, recognition algorithm, modality or sensor, commonly referred to as multibiometrics, is common practice to improve accuracy of biometric systems, fusion at segmentation level has so far been neglected in literature. This paper introduces the concept of multi-segmentation fusion for combining independent iris segmentation results. Fusion at segmentation level is useful to (1) obtain more robust recognition rates compared to single segmentation; (2) avoid additional storage requirements compared to feature-level fusion, and (3) save processing time compared to employing parallel chains of feature-extractor dependent segmentation. As proof of concept, manually labeled segmentation results are combined using the proposed technique and shown to increase recognition accuracy for representative algorithms on the well-known CASIA-V4-Interval dataset.

## 1 Introduction

Aiming to bridge the performance gap of image-based biometric systems between highly accurate standardized cooperative applications and less constrained scenarios has attracted many researchers to propose algorithms improving preprocessing and segmentation techniques, which are reported to play an important role due to susceptibility to poor image quality [10]. The human iris is one of the most unique biometric identifiers, and also selected to be one of two modalities to be employed in the world's largest biometric deployment, Aadhaar, targeting biometric identification of each Indian citizen. It is clear, that such large-scale applications demand high accuracy to avoid misclassification. Furthermore, the discrepancy between users aware of the acquisition and the observed decreased rate when applied in unconstrained scenarios with reported VR (verification rate) as low as 44.6% [14] versus >99% VR at 0.1% FAR (false acceptance rate) for a series of iris biometric systems in constrained environments [1] support the claimed need for higher accuracy in less constrained applications.

A combination of multiple biometric information can increase accuracy at the cost of additional resources and is traditionally employed at the score or decision-level [15]. Such fusion rules unfortunately exhibit limitations: (1) many algorithms conduct the same or similar costly processing steps; (2) segmentation errors propagate along the biometric processing chain, and; (3) contradicting information may derogate system performance. This leads to the question: *Can*

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 310–317, 2013.

*fusion at lower levels (segmentation) lead to more accurate (and faster) biometric systems?* This paper is dedicated to providing a positive answer to the feasibility of fusion at segmentation stage, i.e. whether the combination of independent segmentation results lead to better system performance in terms of recognition accuracy, independent of the chosen feature extraction and comparison algorithm. Note, that the choice of methods may impact processing time.

The remainder of this paper is organized as follows: Section 2 reviews related work with respect to multi-biometric fusion. Section 3 formalizes the referred segmentation model and introduces the concept of fusion at segmentation stage. Section 4 introduces experimental setup and Section 5 presents an experimental evaluation of the proposed technique. Finally, Section 6 concludes this work.

## 2 Multibiometric Fusion

Multibiometric fusion refers to the "use of multiple pieces of evidence in order to deduce or verify human identity" [18] and can be applied at different stages in the biometric processing chain [15]:

1. *Data/Feature Level*: consolidating information from the raw biometric signal or after feature extraction from individual classifiers into a single high-dimensional template.
2. *Score Level*: consolidating comparison scores with density-based (using the likelihood ratio after modeling genuine and imposter score distributions), transformation-based and classifier-fusion-based (learning boundaries from observed data) solutions, this is probably the most-intensively studied type of fusion leaving other processing modules unaffected.
3. *Rank/Decision Level*: depending on whether biometric authentication is performed in identification mode (1-to-N comparison with all subjects registered with the system to determine an identity from a biometric sample) or verification mode (1-to-1 comparison to justify the authenticity of an identification claim as *genuine* or *imposter*), this fusion type consolidates the outcome of individual decision processes, i.e. ranking lists or class decisions.

Due to the development of embedded solutions and with the rise of new biometric modalities focusing on specific parts and/or scales, the original classification of fusion scenarios in [15] into (1) multiple sensors, (2) multiple biometrics, (3) multiple units, (4) multiple snapshots, (5) multiple matchers is less strict and new scenarios emerge [16]. While an integration at early level is claimed to be more effective [15], it is more complex to design. The majority of proposed multi-biometric techniques targeting biometric surveillance (e.g., [19,14,5]) are score-level fusion approaches. Only few data/feature level fusion techniques exist: [4] is the first signal-level fusion approach in iris recognition creating a single high-resolution image from multiple frames in video outperforming score-level fusion techniques. Their proposed technique is essentially an image fusion of iris images at the pixel level. Our approach is different in targeting not multiple snapshots but a single-snapshot only and combining the result of multiple segmentation

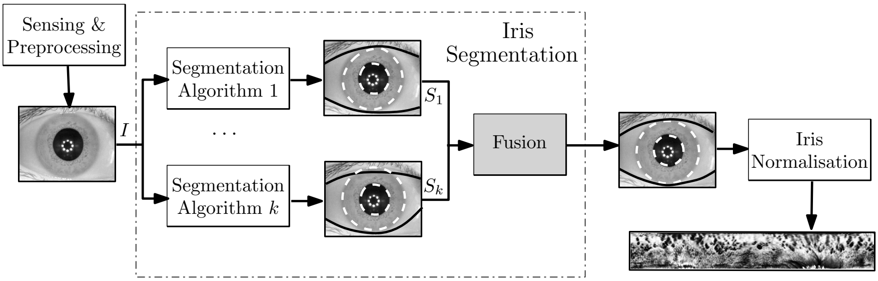


**Fig. 1.** Basic operation mode of novel proposed iris segmentation fusion

results in order to improve recognition accuracy. This is the first approach on combining segmentation results for improved iris recognition.

## 3 Iris Segmentation Fusion

This paper proposes to allow for a combination of multiple segmentation results $S_1, S_2, \ldots S_k$ of the same input iris image $I$ using multiple segmentation algorithms, see Fig 1 for an illustration on how iris segmentation fusion can be integrated into iris processing chains between sensing and normalization. Since not all iris feature extraction techniques require the same preprocessing tasks, the proposed fusion technique uses segmentation results by employing Daugman's normalization [2], which serves as the basis for most commercial applications [12]. A good reference work for practices on image segmentation classifier combination is [6].

### 3.1 Daugman's Iris Normalization Model

In Daugman's algorithm [2], binary features are extracted after mapping iris texture between inner pupillary and outer limbic boundary into a representation called "Faberge coordinates" applying a rubbersheet transform, see Fig. 2. This process involves essentially two tasks [12]: (1) iris segmentation detecting the two (originally circular, but extensible to arbitrarily shaped) boundaries, pupillary and limbic polar curves $P, L : [0, 2\pi) \to [0, m] \times [0, n]$, for the eye instance in the $m \times n$ input image (we assume, that eye detection and quality checks indicate exactly one such instance is present and of sufficient quality); and (2) iris normalization, which creates a normalized representation of the iris texture, invariant under pupillary dilation and facilitating for rotational alignment via simple pixel-shifts using angular $\theta$ and pupil-to-limbic radial $r$ coordinates:

$$R : [0, 2\pi) \times [0, 1] \to [0, m] \times [0, n]. \quad R(\theta, r) := (1 - r) \cdot P(\theta) + r \cdot L(\theta). \quad (1)$$

Besides the mapping in doubly dimensionless coordinates using $R$, due to eyelids and reflections, the resulting rectangular area does not only contain iris texture, but also areas, which should be masked out during feature extraction

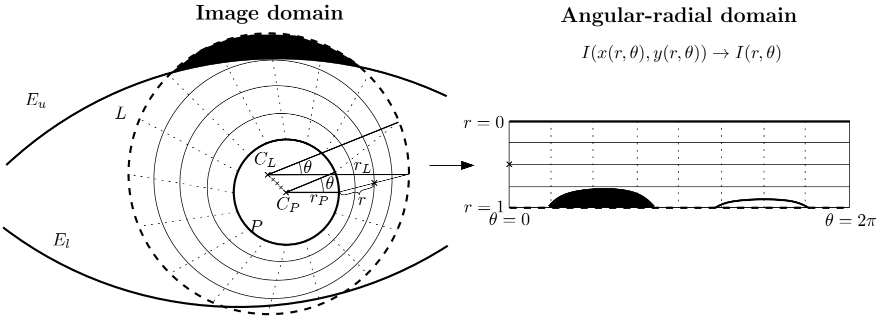


**Fig. 2.** Iris rubbersheet transform model with circular $P, L$ and paraboloid $E_u, E_l$

and comparison. While [13] shows, that indeed, a reordering of pixels based on reliability has almost a similar effect like noise masks (and render their use less effective), traditional processing also creates a binary noise mask as part of the normalization task, $N : [0, 2\pi) \times [0, 1] \to \{0, 1\}$, marking areas occluded by eyelids, eyelashes or reflections. Usually, in order to build this noise mask, upper and lower eyelids are fitted by paraboloid or polynomial curves $E_u, E_l : [0, 1] \to [0, m] \times [0, n]$ to mask out occluded areas in the noise mask.

### 3.2 Combination of Segmentation Results

Motivated by the observation, that more generic alignment using Levenshtein instead of Hamming distance (HD) is able to increase recognition [17], the goal of the fusion module is to obtain a better pupillary and limbic boundary representation for minimizing the effect of mapping deformations due to inaccurately localized boundaries in the rubbersheet transform. While there are several different possibilities to accomplish this task (e.g., for practices on image segmentation classifier combination see [6]), we exemplary introduce two different techniques:

- **Sum-Rule Interpolation**: A very natural choice of a fusion rule combining multiple boundaries $B_1, B_2, \ldots B_k : [0, 2\pi) \to [0, m] \times [0, n]$ into a single boundary $B$ is, in analogy to the sum rule in score-level-fusion, the arithmetic mean of sampled boundaries:

$$\mathit{Sum\ Rule} : B(\theta) := \frac{1}{k} \sum_{i=1}^{k} B_i(\theta) \tag{2}$$

  This interpolation is executed for $B = P$ and $B = L$ separately. The same method can be applied to interpolate between upper and lower eyelid approximations $E_u, E_l$ to derive a common noisemask.
- **Augmented-Model Interpolation**: in case boundaries are rather different and/or the curves' sampling interval $[0, 2\pi]$ is not "equally spaced", i.e. for discretized equidistant samples of arguments $x_1, \ldots x_s \in [0, 2\pi]$ the boundary polygon $B(x_1), B(x_2), \ldots B(x_s)$ has large variation in the length of boundary line segments, sum rule interpolation may lead to inaccurate results. While in

this case a re-parametrization of boundary curves may be useful or necessary for sum-rule interpolation, an alternative approach to the fusion of boundary curves is fitting a model to the union of sampled edge points:

$$Aug\ Rule : B(\theta) := ModelFit\Big(\bigcup_{1\leq i\leq k}\ \bigcup_{1\leq j\leq s} B_i(x_j)\Big)(\theta) \tag{3}$$

where *ModelFit* is a fitting routine taking a set of points and providing a suitable shape (closed boundary curve) minimizing a model-error, e.g. Fitzgibbon's ellipse fitting method [3] in case of $B = P$ or $B = L$. For upper and lower eyelid curves $B = E_u, B = E_l$, input points can be used to fit a polygon of second order to the input points.

## 4 Experimental Setup

In order to estimate the usability of the proposed new fusion framework, we assess its performance on manually segmented iris images. This test is useful, since (1) any dependencies between segmentation algorithms can be avoided in this case enabling a fair test of the fusion rule, (2) a positive outcome justifies its application in building high-confidence fused ground truth for evaluating segmentation algorithms, (3) manual segmentations are state-of-the-art (e.g. in the Noisy Iris Challenge Evaluation [11]) to evaluate segmentation techniques (i.e. considered superior to automated evaluations), therefore if segmentation fusion is able to improve manual segmentation, it is a positive result for also automated segmentation techniques, which are continuously improved to achieve close-to-manual performance.

For experiments we employ the entire *CASIA-V4-Interval*[1] dataset of high quality NIR illuminated indoor images with 320 × 280 pixel resolution (2639 images, 395 classes). For manual segmentations, a male (*Manual 1*) and female (*Manual 2*) expert manually labeled boundary points until the fitted elliptic inner pupillary and outer limbic boundaries sufficiently (according to the opinion of the expert) approximated the true possibly occluded iris boundary. The same procedure was also executed for upper and lower boundaries using a polynomial of order two as the curves' model. During manual segmentation, the expert could zoom in/out and see the original and fitted (segmented) image.

As feature extraction algorithms operating on normalized iris textures, three representative implementations available in USIT[2] were employed: *Masek* [8] is a feature extraction algorithm extracting phase angles from the row-wise convolution of the 1D intensity signals with scaled and oriented Log-Gabor kernels encoding each phase angle with 2 bits leading to a 10240 bits code. Fractional HD is employed for comparison. *Ma* [7] is a feature extraction algorithm tracking the positions of minima and maxima (switching bit sequences) after executing 1D

[1] The Center of Biometrics and Security Research, CASIA Iris Image Database, `http://biometrics.idealtest.org`

[2] University of Salzburg Iris Toolbox, `http://wavelab.at/sources/USIT/`

wavelet transform on the 10 one-dimensional signals, each one averaged from the pixels of 5 adjacent rows for each of two subbands. Again, fractional HD is the comparison criterion. Finally, *Monro* [9] employs a 1D discrete cosine transform (DCT) on diamond-shaped image patches using a Hanning window approach to locally summarize data. The final 2x128 bytes code tracks zero crossings of the differences between the DCT coefficients of adjacent patch vectors using first three DCT coefficients for a total of 7 shift positions $(0; \pm 4; \pm 8; \pm 12)$. Also for the other codes (*Masek, Ma*) the comparison routine employed 7 bit shifts in either direction for optimal alignment.

## 5 Results

We evaluate segmentation accuracy by assessing the impact on verification recognition accuracy, i.e. ROC curves plotting false acceptance rate (FAR) versus genuine acceptance rate (GAR), given in Figs. 3, 4 and 5. GAR at fixed FAR ($\leq 0.01\%$) for each of the two manual segmentations as well as fused results are reported in Table 1.

First, we can see that independent of the employed feature extraction algorithm, both manual segmentations exhibit the same order in performance over the entire operational ROC range: manual segmentation 2 delivers more accurate results with 97.64% GAR at FAR $\leq$ 0.01% for Masek, 98.34% for Ma and 95.72% for DCT-based Monro versus 97.46% for Masek, 98.19% for Ma and 93.94% Monro in case of the first manual segmentation. This suggests, that manual segmentation 2 is more accurate/consistent. Both segmentations needed approximately 9 working days to segment the dataset.

The second important observation is an algorithm-dependent impact of segmentation on recognition accuracy. While typically, algorithms are compared by using their own segmentation technique, we can see that the sensitivity against segmentation among algorithms is quite different and should be considered when comparing algorithms. While for Masek performance differences are almost invisible (1.17% EER Manual 1 vs. 1.15% EER Manual 2, but still better performance for segmentation fused Sum Rule with 1.13% EER and Aug-Rule with 1.12% EER), differences for Monro are clearly present (1.84% EER Manual 1 vs. 1.62% EER Manual 2, vs. Sum Rule with 1.52% EER and Aug-Rule with 1.48% EER).

Third, with respect to the targeted feasibility study of segmentation fusion we can report, that fusion algorithms were able to increase accuracy of both segmentation results, independent of the chosen feature extraction algorithm - a result which is not self-evident and justifies its future investigation with existing segmentation algorithms. Sum Rule Interpolation, which has the advantage of being fast in computing an averaged segmentation representation, could increase GAR from 97.46% to 97.84% for Masek, from 98.19 to 98.57% for Ma, and from 93.94% to 96.74% for the Monro implementation, which did not consider noise masks. Relative performance differences to the Augmented Model Interpolation were insignificant (97.84% GAR for Masek, 98.51% Ma and 96.8% for Monro), i.e. both fusion rules performed almost equally well.

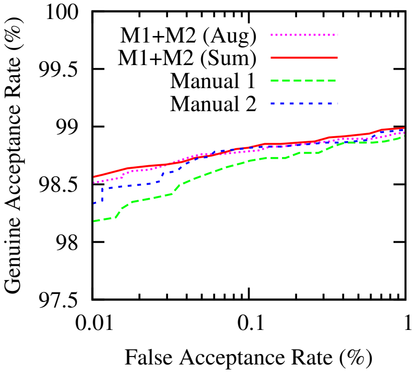


**Fig. 3.** ROC for the *Ma* feature

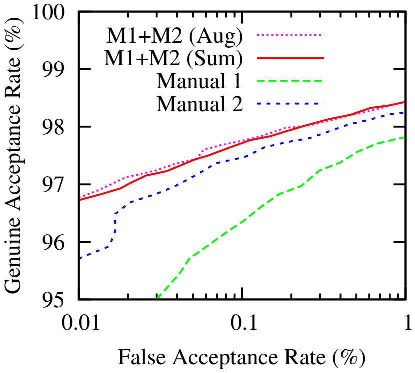


**Fig. 4.** ROC for the *Monro* feature

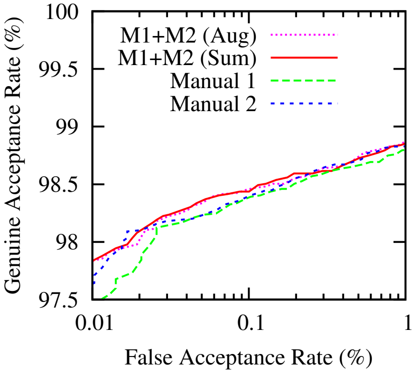


**Fig. 5.** ROC for the *Masek* feature

**Table 1.** Recognition accuracy in Genuine Acceptance Rate at ≤ 0.01% False Acceptance Rate.

| Algorithm | GAR at FAR ≤ 0.01%<br>Masek | <br>Ma | <br>Monro |
|---|---|---|---|
| Manual 1 | 97.46 | 98.19 | 93.94 |
| Manual 2 | 97.64 | 98.34 | 95.72 |
| Sum-Rule | **97.84** | **98.57** | **96.74** |
| Aug-Model | **97.84** | **98.51** | **96.80** |

## 6 Conclusion

Recent challenges like the Noisy Iris Challenge Evaluation (NICE) or Multiple Biometrics Grand Challenge (MBGC) have put a strong focus on the segmentation problem of challenging iris images. But so far, there has been no systematic framework of combining segmentation results from different algorithms. We showed, that besides combining outcomes of biometric feature extraction or comparison algorithms, it may be useful to combine segmentation and normalization information from multiple sources. Evaluations using manual segmentation on CASIA-V4-Interval revealed improvement by segmentation fusion for each of the employed feature extraction algorithms and fusion rules. Segmentation-fused recognition was as high as 96.8% GAR at ≤ 0.01% FAR (Aug-Rule) vs. 93.94% and 95.72% for individual segmentations in case of Monro's feature. Future work will focus on automatic segmentation algorithms, more challenging datasets, and quality-related information assisting fusion rule selection.

## References

1. Bowyer, K.W., Hollingsworth, K.P., Flynn, P.J.: A survey of iris biometrics research: 2008-2010. In: Burge, M., Bowyer, K. (eds.) Handbook of Iris Recognition, pp. 15–54. Springer, New York (2012)
2. Daugman, J.: How iris recognition works. IEEE Trans. on Circiuts and Systems for Video Technology 14(1), 21–30 (2004)
3. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. IEEE Trans. Pattern Anal. Mach. Intell. 21(5), 476–480 (1999)
4. Hollingsworth, K., Peters, T., Bowyer, K., Flynn, P.: Iris recognition using signal-level fusion of frames from video. IEEE Trans. Inf. Forensics and Sec. 4(4), 837–848 (2009)
5. Jillela, R., Ross, A.: Mitigating effects of plastic surgery: Fusing face and ocular biometrics. In: Proc. 5th IEEE Int'l Conf. on Biometrics: Theory, App. and Syst., pp. 402–411 (2012)
6. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, Hoboken (2004)
7. Ma, L., Tan, T., Wang, Y., Zhang, D.: Efficient iris recognition by characterizing key local variations. IEEE Transactions on Image Processing 13(6), 739–750 (2004)
8. Masek, L., Kovesi, P.: MATLAB Source Code for a Biometric Identification System Based on Iris Patterns (2003), `http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html`
9. Monro, D.M., Rakshit, S., Zhang, D.: Dct-based iris recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 586–595 (2007)
10. Proença, H., Alexandre, L.A.: Iris recognition: Analysis of the error rates regarding the accuracy of the segmentation stage. Image and Vision Comp. 28(1), 202–206 (2010)
11. Proença, H., Alexandre, L.: Toward covert iris biometric recognition: Experimental results from the nice contests. IEEE Trans. Inf. Forensics and Sec. 7(2), 798–808 (2012)
12. Rathgeb, C., Uhl, A., Wild, P.: Iris Recognition: From Segmentation to Template Security. Advances in Information Security, vol. 59. Springer, New York (2012)
13. Rathgeb, C., Uhl, A., Wild, P.: Incremental iris recognition: A single-algorithm serial fusion strategy to optimize time complexity. In: Proc. Int'l Conf. on Biometrics: Theory App. and Syst., pp. 1–6. IEEE (2010)
14. Ross, A., Jillela, R., Smereka, J., Boddeti, V., Kumar, B., Barnard, R., Hu, X., Pauca, P., Plemmons, R.: Matching highly non-ideal ocular images: An information fusion approach. In: Proc. 5th Int'l Conf. on Biometrics, pp. 446–453 (2012)
15. Ross, A.A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, Secaucus (2006)
16. Uhl, A., Wild, P.: Single-sensor multi-instance fingerprint and eigenfinger recognition using (weighted) score combination methods. Int'l Journal on Biometrics 1(4), 442–462 (2009)
17. Uhl, A., Wild, P.: Enhancing iris matching using Levenshtein distance with alignment constraints. In: Bebis, G., et al. (eds.) ISVC 2010, Part I. LNCS, vol. 6453, pp. 469–478. Springer, Heidelberg (2010)
18. Vatsa, M., Singh, R., Noore, A., Ross, A.: On the dynamic selection of biometric fusion algorithms. IEEE Trans. on Inf. Forensics and Security 5(3), 470–479 (2010)
19. Woodard, D., Pundlik, S., Miller, P., Jillela, R., Ross, A.: On the fusion of periocular and iris biometrics in non-ideal imagery. In: Proc. 20th Int'l Conf. on Pattern Recognition, pp. 201–204 (2010)

# A Non-temporal Approach for Gesture Recognition Using Microsoft Kinect

Mallinali Ramírez-Corona, Miguel Osorio-Ramos, and Eduardo F. Morales

Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, Puebla, 72840, México
{mallinali.ramirez,mjoramos,emorales}@inaoep.mx

**Abstract.** Gesture recognition has become a very active research area with the advent of the Kinect sensor. The most common approaches for gesture recognition use temporal information and are based on methods such as Hidden Markov Models (HMM) and Dynamic Time Warping (DTW). In this paper, we present a novel non-temporal alternative for gesture recognition using the Microsoft Kinect device. The proposed approach, Recognition by Characteristic Window (RCW), identifies, using clustering techniques and a sliding window, distinctive portions of individual gestures which have low overlapping information with other gestures. Once a distinctive portion has been identified for each gesture, all these sub-sequences are used to recognize a new instance. The proposed method was compared against HMM and DTW on a benchmark gesture's dataset showing very competitive performance.

**Keywords:** Machine Learning, Gesture Recognition, Kinect.

## 1 Introduction

Advances in computer vision technology provide us with a large number of tools that give us different types of information, making the data manipulation and extraction easier and more precise. A trending device is the Kinect sensor, a technology developed by Microsoft mainly for movement recognition and tracking. It integrates an RGB camera, a depth sensor consisting of an infrared laser projector, and a multi-array of microphones. The Kinect sensor has triggered an increased interest in gesture recognition.

Most gesture recognition systems use temporal information for building their models and for classifying new gestures. Common techniques include Hidden Markov Models (HMM) and Dynamic Time Warping (DTW). The rationale is that taking into account the temporal information from the gesture a better classifier can be build.

In this paper, we take an alternative approach where we train a classifier using "static" information. The advantage is that there is a large number of off-the-shelf robust algorithms that can be directly applied.

Our approach, Recognition by Characteristic Window (RCW), is based on the idea that for each gesture there is a sub-sequence of frames (window) distinct

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 318–325, 2013.

from all other gestures. In this paper, we implement a novel approach that scans a gesture with a sliding window to find, using clustering, a distinctive sub-sequence of that gesture. The generated windows for each gesture are used as input to a classifier to recognize new instances.

We trained different classifiers and compared different classification policies. Our proposed method was also compared against Dynamic Time Warping (DTW) and Hidden Markov Models (HMM) on a benchmark dataset. It is shown that RCW obtained very competitive results when compared against DTW and HMM models.

The remainder of the paper is organized as follows. Section 2 summarizes the most relevant related work for this research. Section 3 describes how the data is pre-processed to obtain new attributes which are robust to translations and rotations. In Section 4 our method is described, Section 4.1 describes the clustering phase of the method where the best windows are found for each gesture and Section 4.2 explains the way the classifier is trained and how the classification is produced. Section 5 describes the performed experiments and results and Section 6 provides conclusions and future research directions.

## 2 Related Work

Several approaches have been recently proposed for gesture recognition using the Kinect sensor. Kurakin et al. [5] propose a real-time system for hand-gesture recognition using an action graph which shares similar robust properties with standard HMM.

Raptis et al. [6] propose a real-time dance gesture recognition system based on an angular skeleton representation, and a cascaded correlation-based max-likelihood multivariate classifier that takes into account that dancing adheres to a canonical time-base to simplify the template matching process. It uses a space-time contract-expand distance metric to compare the input with an oracle (the ideal movement).

Biswas and Basu [1] propose a method to recognize human gestures using the Kinect® depth camera. First they isolate the human figure from the background and create a region of interest (ROI) by placing a grid on the extracted foreground, the gesture is parametrized using depth variation and motion information content of each cell of the grid.

Wu et al. [4] propose an actionlet ensemble model to represent each action and to capture the intra-class variance. An actionlet is a particular conjunction of the features for a subset of the joints that are important for each gesture. They also add new features called local occupancy pattern (LOP), these features are robust to noise, invariant to translational and temporal misalignment, and capable of characterizing both the human motion and the human-object interactions

Yang et al. [7] choose 3-dimensional feature vector for 3D gesture recognition from consecutive hand coordinates in a spherical coordinate. They propose a hand tracking algorithm that detects a moving object, if it moves like a wave motion the algorithm decides the object is a hand. Gestures are recognized by a HMM using Baum-Welch algorithm to estimate the parameters.

Carmona and Climent [2] discussed about the best technique for hand gesture recognition: HMM or DTW using Kinect® skeleton. The first step in gesture recognition is the selection of the features; usually, these features are location, orientation and velocity. For HMM they used Baum-Welch algorithm to find the model that best describes the spatio-temporal dynamics of each gesture, the probability of the gesture produced by each HMM is evaluated using Viterbi algorithm. DTW calculates the distance between two signals, thus they used a k-NN classifier to determine which is the most likely class. They obtained best results in their dataset using DTW.

Unlike previous approaches, we employ traditional classifiers using a distinctive part of each gesture.

## 3 Preprocessing

The performance of gestures by a user can be done at different distances and from different orientation angles. In this paper, we transformed the raw data produced by the joints of the "skeleton" generated by the Kinect, into a scheme invariant to translation and rotation. In particular, we simplified the method presented in [6], that transforms the data from joint points to angles. Our approach computes the angles between three consecutive joints (e.g. wrist-elbow-shoulder), using the cosine formula (1). This formula gets the angle between two vectors, in this case represented by the joints coordinates.

$$\cos\theta = \frac{a \cdot b}{\|a\| \cdot \|b\|} \tag{1}$$

From the twenty joint coordinates produced by the "skeleton" from the Kinect, only nine were selected as the most descriptive joints. These selected joints were used to obtain the relative angles between consecutive joints, reducing then the attributes from 20 × 3 (points x, y and z of each joint) to 9, producing a representation invariant to translation and rotation. The attributes are shown in Figure 1.

## 4 Recognition by Characteristic Window

Our method, RCW, is divided in two phases, the first (Section 4.1) finds the most representative section for each gesture and the second (Section 4.2) classifies the frames and returns a prediction based on the information obtained in the first phase.

### 4.1 Clustering

Given a set of gestures $G = \{g_1, g_2, ..., g_k\}$, our hypothesis is that for each gesture there exists a sub-sequence of frames that is different from any subsequence of all the other gestures. We implemented a method to find that subsequence through clustering. The algorithm proceeds as follows: we take a sliding

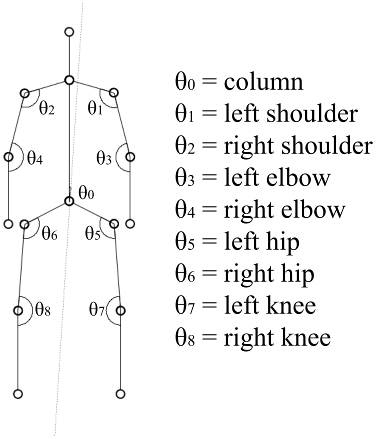


**Fig. 1.** Skeleton joints showing the most descriptive angles

window with a predefined size relative to the number of frames (percentage) of a gesture. Given a particular window (set of instances) of one gesture ($g_i$) and the complete sequences of all the other gestures, we run k-means with k equal to the number of classes (different gestures) we want to recognize. If the clustering method generates a cluster whose elements are mostly samples from the selected window, this is returned as a sub-sequence that is distinctive enough from the other gestures. For each sliding window we use the f-score (see Equation (2)) to evaluate how distinctive is this window with respect to the other gestures.

$$F1 = 2 \cdot \left( \frac{precision + recall}{precision \cdot recall} \right) \tag{2}$$

### 4.2 Classification

We trained a classifier using either the complete sequences of the gestures or using only the distinctive identified windows for all the gestures, with the nine angles as attributes (Section 3). The trained classifier is used to assign one of the possible gestures to each frame of a testing gesture.

For testing, we implemented two decision policies:

1. We classify each frame from the testing gesture and return the class of the longest set of consecutive frames classified equally. We call this policy, *longest sequence* (LS).
2. The second policy takes advantage of the positions of the identified windows in the clustering process. The testing gesture is evaluated only in the windows that were selected during the clustering phase. For each window we obtain a percentage of coincidence and return the class belonging to the window with the highest value. We call this policy *window verification* (WV).

**Algorithm 1.** RCW clustering algorithm, where *windowSizes* are a pre-defined set of percentages of windows to tried, *step* is the percentage of how much a window is slid each time and *currentScore* is a temporal variable that stores the accuracy of the clusterization for an specific window.

```
Require: $G, windowSizes, step \geq 0$
Ensure: $bestWindowSize, bestWindowPosition : \forall g_i \in G$
for all $g_i \in G$ do
    $maximumScore \leftarrow -1$
    for all $size$ such that $size \in windowSizes$ do
        for $position = 0$ to $position \leq (100-\text{step})$ do
            $datasetToCluster \leftarrow (\forall frame|frame \in window(g_i, position, size) \bigcup (\forall frame \in g_k|g_k \neq g_i)$
            $currentScore \leftarrow eval(kNN(datasetToCluster))$
            if $currentScore > maximumScore$ then
                $maximumScore \leftarrow currentScore$
                $bestWindowSize \leftarrow size$
                $bestWindowPosition \leftarrow position$
            end if
            $position \leftarrow \text{position} + step$
        end for
    end for
end for
```

Since the windows are selected as percentage of the gesture, its use still works with longer or shorter gesture instances.

## 5 Results

We tested RCW on the dataset *Microsoft Research Cambridge-12* (MSRC-12) which consists of 594 sequences of movements of an skeleton characterizing the human body. These sequences were collected from 30 persons doing 12 gestures having a total of 6244 instances. The set of files contains the tracking of 20 joints presented as points in the space $< x, y, z >$; each of these files contains around ten instances per gesture performed one after the other. The gestures can be categorized into two abstract categories: Iconic gestures, those that imbue a correspondence between the gesture and the reference, and Metaphoric gestures, those that represent an abstract concept. For the experiments we used the subset of iconic gestures.

- Gesture 2: *Crouch or hide* [500 instances]
- Gesture 4: *Put on night vision goggles* [508 instances]
- Gesture 6: *Shoot a pistol* [511 instances]
- Gesture 8: *Throw an object* [515 instances]
- Gesture 10: *Change weapon* [498 instances]
- Gesture 12: *Kick* [502 instances]

The accuracy from the clustering and the classification phases were measured using the f-score (see Equation (2)).

In the training phase, different window sizes and positions were tested for each gesture. We slid each window 2% of the total gesture each time. For the evaluation phase we used 10-cross fold validation.

Table 1 shows the different values in terms of window size of the precision results and the window position for the six gestures. The best results are marked in bold face. Figure 2 depicts the best windows found for the training data.

**Table 1.** Best window starting and window length for each gesture, where AC is Accuracy (%) and WP is Best Window position in percentage

| Gesture | Window size | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 10% | | 15% | | 20% | | 25% | |
| | AC | WP | AC | WP | AC | WP | AC | WP |
| Duck | **96.96** | **90** | 93.44 | 84 | 89.77 | 80 | 82.03 | 74 |
| Googles | 29.29 | 6 | 42.58 | 4 | 49.83 | 2 | **60.41** | **0** |
| Shoot | 33.22 | 90 | 41.80 | 84 | 47.65 | 80 | **49.91** | **74** |
| Throw | 15.05 | 76 | 20.58 | 74 | 24.78 | 70 | **27.87** | **68** |
| Change Weapon | 14.93 | 90 | 20.34 | 0 | **32.08** | **78** | 27.68 | 0 |
| Kick | 8.04 | 84 | 9.75 | 48 | **30.12** | **80** | 15.64 | 74 |

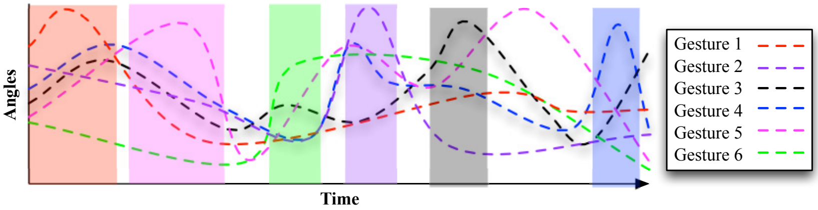


**Fig. 2.** Graphical representation of the sections found by the clustering phase, the colored columns represent an example of the most representative part of each gesture

Once the distinctive windows were identified for the gestures, we tried four different classifiers from Weka: Naïve Bayes, SVM, C4.5 and Random Forest. After training the classifier, the classification of new gestures was carried out using *longest sequence* (LS) and *window verification* (WV) policies.

We performed tests with two training datasets:

1. Pre-processed dataset (PP-MSRC), which uses the whole transformed sequence of frames to train a classifier.
2. Pre-processed dataset which uses only the frames that belong to the window for each example of gesture (W-MSRC).

The obtained results are shown in Table 2 using a 10-fold cross-validation; the best results are marked in bold face. The overall best is marked with an asterisk.

As can be seen from the results, considering only the distinctive window for evaluation (the WV policy) increases the accuracy in all cases.

**Table 2.** Obtained results with different classification schemes for each dataset using Longest Sequence (LS) and Window Verification (WV) policies

| Classifier | PP-MSRC | | | | | | W-MSRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LS | | | WV | | | LS | | | WV | | |
| | Prec | Rec | Acc | Prec | Rec | Acc | Prec | Rec | Acc | Prec | Rec | Acc |
| C4.5 | 80.13 | 80.34 | 80.23 | 89.35 | 89.35 | **89.35** | 67.48 | 69.70 | 69.90 | 85.51 | 86.70 | 86.10 |
| SVM | 62.09 | 63.45 | 62.76 | 83.71 | 83.86 | 83.78 | 39.73 | 61.12 | 48.15 | 90.61 | 91.11 | **90.86** |
| Naïve Bayes | 48.26 | 58.05 | 52.74 | 80.78 | 82.13 | 81.45 | 41.15 | 62.15 | 49.52 | 90.65 | 91.24 | **90.94** |
| Rand. Forest | 85.79 | 86.18 | 85.99 | 91.82 | 91.84 | **91.82*** | 75.98 | 77.39 | 76.67 | 91.10 | 91.85 | 91.47 |

RCW (WV policy, PP-MSRC dataset and Random Forest classifier) was compared against two typical methods of gesture recognition: DTW and HMM. As in the previous experiment the accuracy was measured with f-score. The experiment was evaluated using 10-fold cross-validation.

A HMM for each gesture was learned using the Baum-Welch algorithm, then the probability for the frames sequence is computed using Viterbi algorithm, the returned prediction is the one with the best predicted probability. We tried with different number of hidden nodes and report only the best results, that were obtained using three nodes.

To calculate the most probable gesture using DTW, the distance to a subset of examples of each of the gestures (50 examples for this experiment) was computed using the mean of the calculations, the predicted gesture was the one where the distance was smaller.

The results of these experiments are shown in Table 3. A paired t-test was carried out to find statistical significance in the results (marked with an arrow). As can be seen RCW is very competitive against temporal-based approaches and it is statistically better (with 95% of confidence value) against DTW. Apart from that, the small difference between HMM and RCW shown in the results suggests that RCW is a suitable substitute of HMM for this specific problem.

**Table 3.** Comparing accuracy of RCW against DTW and HMM (percentage)

| | Overall | Duck | Googles | Shoot | Throw | Ch. Weapon | Kick |
|---|---|---|---|---|---|---|---|
| DTW | 82.74 ↓ | 97.11±1.26 | 71.83±0.89 | 97.14±0.79 | 76.89±1.53 | 75.55±2.17 | 55.74±2.56 |
| HMM | 91.81 | 97.73±1.42 | 88.06±1.30 | 87.45±2.66 | 90.14±2.41 | 90.82±0.75 | 93.95±2.39 |
| RCW | **91.82** | 95.49±1.89 | 85.25±1.88 | 93.71±1.43 | 95.43±1.24 | 82.07±3.2 | 97.71±0.75 |

↓ Statistically inferior result with respect to RCW.

## 6 Conclusions

This article described a novel non-temporal approach to classify gestures from information obtained by a Kinect sensor. RCW identifies distinctive portions of each gesture using a sliding window and a clustering technique. Each window is given as input to a classifier and a new gesture is classified using also a window-based approach. It is shown that our non-temporal approach is very competitive against standard temporal approaches normally used for gesture recognition. As future work we would like to perform more tests involving a larger set of gestures. We would also like to combine more than one discriminatory window for each gesture to improve performance.

## References

1. Biswas, K.K., Basu, S.K.: Gesture recognition using microsoft kinect. In: 2011 5th International Conference on Automation, Robotics and Applications (ICARA), pp. 100–103 (2011)
2. Carmona, J.M., Climent, J.: A performance evaluation of hmm and dtw for gesture recognition. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 236–243. Springer, Heidelberg (2012)
3. Fothergill, S., Mentis, H., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 1737–1746. ACM, New York (2012)
4. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, pp. 1290–1297. IEEE Computer Society, Washington, DC (2012)
5. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 1975–1979 (2012)
6. Raptis, M., Kirovski, D., Hoppe, H.: Real-time classification of dance gestures from skeleton animation. In: Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2011, pp. 147–156. ACM, New York (2011)
7. Yang, C., Jang, Y., Beh, J., Han, D., Ko, H.: Gesture recognition using depth-based hand tracking for contactless controller application. In: 2012 IEEE International Conference on Consumer Electronics (ICCE), pp. 297–298 (2012)
8. Zhang, H., Du, W., Li, H.: Kinect gesture recognition for interactive system (2012)

# Automatic Verification of Parent-Child Pairs from Face Images

Tiago F. Vieira*, Andrea Bottino, and Ihtesham Ul Islam

Politecnico di Torino,
Corso Duca degli Abruzzi, 24 - 10129 Torino, Italy
{tiago.figueiredo,andrea.bottino,ihtesham.ulislam}@polito.it
http://www.polito.it/cgvg

**Abstract.** The automatic identification of kinship relations from pairs of facial images is an emerging research area in pattern analysis with possible applications in image retrieval and annotation, forensics and historical studies. This work explores the computer identification of pairs of kins using different facial features, based on geometric and textural data, and state-of-the-art classifiers. We first analyzed different facial attributes individually, selecting the most effective feature variables with a two stage feature selection algorithm. Then, these features were combined together, selecting again the most relevant ones. Experiments shows that the proposed approach provides a valuable solution to the kinship verification problem, as suggested by the comparison with a different method on the same data and on the same experimental protocol.

**Keywords:** Kinship verification, SVM, Random Forests, mRMR, SFS.

## 1 Introduction

The analysis of 2D or 3D facial images is a main research topic in pattern analysis and computer vision. Automatic Kinship Verification (KV) has recently received attention from the research community. KV aims at recognizing the degree of kinship of two individuals from their facial images and has possible applications in historic and genealogic research, automatic management and labeling of image databases, forensics and finding missing family members. This is a challenging problem, which should deal with different degrees of kinship and variations in age and gender.

Automatic KV was first introduced by Fang et al. [1], who analysed a database of 150 pairs of parent-child images. Features were extracted for each face with a simplified Pictorial Structure Model and the best classification achieved 70.69%, outperforming the 67.19% obtained by a panel of human raters on the same data. In order to identify parent-child pairs considering the influence of age factor, Xia et al. [2] proposed an extended Transfer Subspace Learning (TSL), which

* The author is also with the Dept. of Electronics and Systems from the Federal University of Pernambuco, Brazil.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 326–333, 2013.

is meant to simplify the recognition task by transferring the knowledge learnt from the similar, but easier task, of recognizing the same parent-child pairs but using images of parents in youth. Classification based on geometric and textural features provided a 60% accuracy. Somanath et al. [3] analyzed the multi-class problem of identifying both parent-child and siblings using Metric Learning, providing 75% and 80% accuracies for, respectively, siblings and parent-child pairs. In [4], Lu et al. proposed a new neighborhood repulsed metric learning (NRML) method for kinship verification. Working on different degrees of kinship, they obtained an average class accuracy of 76.5%.

Recently, in [5] we presented a work on sibling identification. Different facial attributes, related to geometric, holistic and textural features, were first extracted and then combined together. Support Vector Machines (SVM) classification, with the contribution of a Feature Selection process, outperformed the recognition capabilities of a panel of human raters.

This paper extends the main ideas of our previous work in order to analyze the capabilities of different facial attributes to recognize other degrees of kinship and, specifically, of parent-child relationships. The contribution of this work is twofold. First, identifying the facial attributes more fit to tell parent-child pairs from unrelated individuals. Second, comparing on these attributes the accuracy of two state-of-the-art classifiers, namely, SVM and Random Decision Forests (RDF), with previous results in the literature. Our experiments show that some of the facial attributes, when considered individually, are indeed able to classify pairs of parent-child images with performances better than previous works, and, most of all, that the combination of attributes of different natures improves the performance of the final classifier.

The remaining of the paper is organized as follows, in Section 2, we describe the database of parent-child images we used in our experiments. Section 3 details the algorithm we used for tackling the kin verification problem. Results are presented and discussed in Section 4 and in Section 5 we draw the conclusions.

## 2 Image Database

The recent interest into the KV problem led to availability of some databases of facial images of individuals related by different kinship degrees, such as those used in the papers referenced in the Introduction. They are all composed by an heterogeneous set of images, mostly collected through the Internet, and are characterized by non-uniform illumination, background, pose, expression and different age range and ethnicity of the depicted individuals.

However, some of these databases were not found suitable for our approach. For instance, the parent-child dataset collected by Jiwen Lu et al. [6] is composed by faces cropped in a way that precludes the very first step of our method, *i.e.* the automatic detection of facial landmarks (see Section 3). The one used in [7] contains many grayscale images, thus hampering the use of color-based textural features which have proven to be effective to identify siblings [5]. The database we found more suited to our work was the one collected by Fang et al. [1]. This dataset consists of 288 individuals' images[1] (144 parent-child pairs). The colour photographs depict several public celebrities with different age, gender and race in slightly different poses (mostly frontal), illumination conditions and expressions (some neutral but often smiling). The individuals are 50% Caucasians, 40% Asians and 10% of other ethnicities; 40% of the samples are father-son pairs, 22% are father-daughter, 13% are mother-son, and 26% are mother-daughter.

## 3 Algorithm Outline

In the literature, different representations of the information conveyed by faces have been experimented for different tasks. In our previous work on the automatic identification of siblings [5], we found that the discriminative power of separate facial attributes, related to geometric, textural and holistic features, is substantially improved by that of the integration of information of different nature. Based on this consideration, in our work we first analysed individually the contribution of different attributes to the parent-child recognition problem, and then we evaluated different combinations of them.

The outline of the proposed classification algorithm is the following. For each individual, we normalized his/her image and we extracted different feature vectors, one for each feature extraction technique considered. When an image is characterized combining different attributes, their corresponding feature vectors are concatenated. Then we constructed a pair dataset containing all the positive (kin) and an equal number of randomly chosen negative (non-kin) pairs. For each attribute (or attribute group) and each pair, a representative vector is built. Finally, the most relevant pair feature variables were selected and used to train and test a classifier.

### 3.1 Image Normalization

Image normalization is aimed at reducing the influence of different illumination, background and orientation of the faces. First, 76 facial landmarks were automatically identified using the Active Shape Models (ASM) technique [8] (Fig. 1). Second, the ellipse best fitting the 15 landmarks around the chin was used to segment the face and discard the image background. Third, images were geometrically aligned by making the external corners of the eyes coincident with two reference position. Geometric normalization involved translation, rotation and isotropic scaling of the original images. The size of the final normalized images is 100 by 100 pixels.

### 3.2 Features Extraction and Characteristic Vectors

The choice of the facial feature used in this work takes into account the lessons learned in our previous experience on sibling verification. In particular, we found

[1] Although the authors reported the use of 150 pairs, 300 images, the online version contains only 144 pairs.

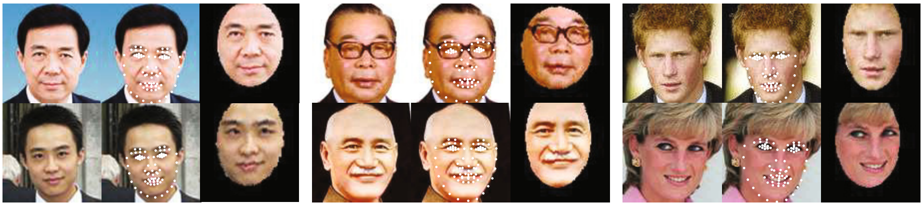


**Fig. 1.** Examples of parent ($1^{st}$ row) - child ($2^{nd}$ row) pairs. For each individual, we show the original image, the detected landmarks and the normalized image

that the contribution of holistic attributes and of some of the geometric and textural attributes experimented in [5] was negligible. Conversely, more discriminative attributes were found in our preliminary experiments on the parent-child dataset, whose detailed results are not reported here for the sake of brevity. In the following, we briefly summarize the characteristics of the chosen attributes.

In order to extract geometric attributes, we first created a dense reference net composed by 184 segments for each face, which was obtained from the Delaunay Triangulation (DT) of the average position of each ASM landmark over the database images. Then, we computed the following attributes: **SEGS**, the 184 lengths of the DT segments, **ANGLES**, the 342 angles of the triangles obtained from DT, and **RATIOS**, the set of 862 ratios of pairs of DT segments sharing the same vertex. Each pair is considered only once to compute a ratio, *i.e.* when a ratio is computed, its inverse is not considered.

Two image descriptors were used to characterize our samples. The first (**CLID**) is based on color local image descriptors. Their general idea is to encode, for a reference point, the Scale-Invariant Feature Transform (SIFT) [9] descriptors computed separately on each image channel. This allows to obtain a representation of the point neighbourhood which is invariant to several image variations (e.g. ligh intensity change and shift, light color change and shift; see [10] for details). Among the color descriptors surveyed in [10], we choose C-SIFT [11] since in our preliminary experiments it performed slightly and consistently better. To characterize a sample with this attribute, we computed a C-SIFT descriptor (a vector of 384 components) on each of the 76 facial landmarks.

The second textural attribute is the Weber local Descriptor (**WLD**) [12] which is based on the Weber's law. It states that a just-noticeable difference in a stimulus is proportional to the magnitude of the original stimulus. Translating this concept into image intensities, WLD first characterizes a pixel with (i) the *differential excitation*, computed from the sum of differences of intensity with its neighbors later divided by its intensity, and (ii) the orientation of the pixel gradient. Then, the WLD features, computed using a multi-scale analysis onto each image pixel of the intensity image, are encoded into a histogram containing 2.880 elements.

Each of the described attributes summarizes a facial image into a *characteristic vector*. When more attributes are considered, the *characteristic vectors* of the images are obtained by simply concatenating the different attributes. For each attribute, or attribute group, the *characteristic vector* $v^{(ab)}$ for a pair of individuals $a$ and $b$ is given by the vector of Euclidean distances, in their respective n-dimensional space, of the corresponding elements of the *characteristic vectors* of $a$ and $b$. Thus, the *characteristic vectors* of a pair are commutative, *i.e.* $v^{(ab)} = v^{(ba)}$.

### 3.3 Building the Classifiers

In order to assess the capabilities of the different attributes, or attribute groups, to tell kin from non-kin pairs, we compared two state-of-the-art classifiers, namely, SVM [13] and Random Decision Forests RDF [14], which are widely recognized for their classification performances by computer scientists and machine learning researchers.

Concerning SVM, we used a radial basis kernel, optimizing its parameters by means of a grid search as suggested in [15]. Before applying SVM, each feature variable was linearly scaled to the range [0,1]. This avoids the variables in larger scales to dominate those in smaller ranges and reduces numerical problems in the computation of the SVM kernels.

Since SVM classifiers are likely to be affected by overfitting, being in most experiments the number of features much greater than the number of samples, we applied the two-step feature selection (FS) process described in [5]. First, the features are ranked for relevance according to the min-Redundancy Max-Relevance (mRMR) method [16]. Then, the set including the top 50 mRMR features is further reduced to its optimal size (*i.e.*, that optimizing the SVM classification accuracy) with a Sequential Forward Selection (SFS) scheme. For feature vectors obtained as combination of different attributes, the FS selection was first performed separately on each attribute and then repeated on the aggregation of the selected feature variables.

As for RDF, we first optimized their parameters (*e.g.*, number of trees, tree depth and so on) with a grid search over the parameter space, choosing the set with the lowest out of bag error on the training set. Since at each split node a single feature variable is selected for decision, scaling the feature vectors is not relevant in this case.

Due to their working mechanism, RDF are relatively unaffected by overfitting. Nevertheless, selecting the most relevant features can improve their overall accuracy. To this end, once the optimal parameters were found, we sorted the features according to their Variable Importance [14] and we trained again the RDF with a subset of these variables. The size of this subset was chosen, heuristically, by iteratively increasing the number of candidates in the feature set until the global optimum was found.

## 4 Results and Discussion

In our experiments, we first computed the accuracies based on the classification of each individual attribute. Then, we evaluated the accuracies obtained by characterizing each facial image with three different groups of attributes: **GEOMETRIC**, grouping the geometric attributes, **TEXTURE**, combining textural information, and **ALL**, concatenating all the described attributes. The classification results obtained are summarized in Table 1 and organized by attribute, or attribute group, and by classification algorithm (SVM vs. RDF). Results were assessed using stratified five-fold cross-validation (CV), and, hence, we report the average classification rates of each classifier over the different CV rounds.

The following remarks can be drawn:

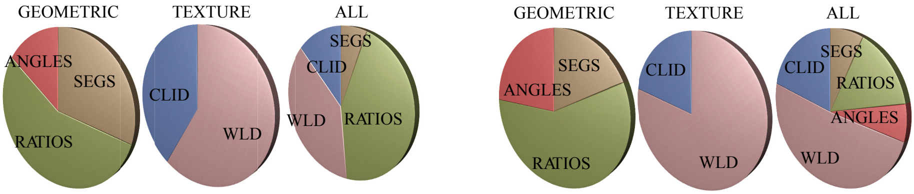
– concerning the individual attributes, textural features have a higher discriminative power than the geometric ones, with WLD obtaining the best performances (78.0% with SVM);
– the more heterogeneous the information, the better the accuracies. As a matter of facts, grouped attributes performed consistently better than their single components, and the best accuracies were obtained for both algorithms considering all attributes together achieving 81.8% and 77.5% for, respectively, SVM and RDF;
– as for the classification techniques, SVM, in combination with a proper selection of the most relevant features, provides, in this specific problem, consistently better performances than RDF.

One expected result, not shown in Table 1, is that Feature Selection (FS) always provides a significant classification improvement (between 6% and 14%, for SVM, and 1% and 12% for RDF). Concerning the selection process, it is also interesting to analyse the distribution of features surviving the FS pruning for composite attributes and different classifiers (see Figure 2), which could provide some insights into the more relevant facial characteristics to recognize kins.

We can observe the following: (i) RATIOS is the most relevant geometric attribute, suggesting its good descriptive capabilities; (ii) as for textural features, WLD is more relevant than CLID in the TEXTURE groups, but CLID features contribute reasonably when geometric features are added; (iii) when attributes are grouped, features from all attributes are chosen to compose the final vector;

**Table 1.** Accuracy results. For each attribute and each classification algorithm, we show the percentage of correct classifications and, in brackets, the optimal number of variables selected by the FS process.

| | SVM | RDF |
|---|---|---|
| SEGS | 68.2 (18) | 60.1 (40) |
| RATIOS | 73.1 (13) | 59.3 (175) |
| ANGLES | 68.9 (30) | 57.2 (100) |
| CLID | 74.1 (14) | 66.3 (62) |
| WLD | 78.0 (27) | 70.6 (250) |
| GEOMETRIC | 74.3 (8) | 65.4 (175) |
| TEXTURE | 80.1 (18) | 76.1 (150) |
| ALL | 81.8 (29) | 77.5 (150) |



(a) SVM: feature selection.

(b) RDF: feature selection.

**Fig. 2.** Feature selection applied to (a) SVM and (b) RDF: distribution of feature variables per type for different attribute groups

the only exception was the ALL group with SVM where ANGLES were discarded (which could be expected since they convey an information similar to SEGS); (iv) when geometric and textural features are combined (ALL group), the latter are preferably selected to compose the final dataset, in particular with RDF.

Finally, we can asses the quality of our results by comparing the classification accuracies of our experiments with that obtained on the same dataset by Fang et al. [1]. The performance of their approach (70.69%) and that of a panel of human raters on the same data (67.19%) are already improved by that obtained in our work with several individual attributes, and outperformed by our best result (81.8%), achieved with the integration of all attributes.

Concluding, the experimental results show that our approach, based on the integration of geometric and textural features, together with a proper selection of the feature variables, is indeed a valuable solution to the automatic KV process.

## 5 Conclusion and Future Work

We presented an approach for automatically identifying pairs of parent-child images through the extraction and selection of several features from face images. Different attributes, related to geometry and texture data, have been first analysed individually and then combined together to provide higher classification performances.

Simulation results using state-of-the-art classification algorithms show that our combination of features, together with a proper selection of the feature variables, is indeed a valuable solution to the automatic KV process, obtaining high classification accuracies (81.8%) and outperforming previous approaches on the same data.

As future work, we are planning to address the multi-class problem of identifying the specific degree of kinship (e.g., parent-child, parent-grandchild and so on), a problem which has not been yet thoroughly investigated. Another interesting point to be considered is how factors such as gender and age influence a kinship predictor, and possible approaches to alleviate such influences.

## References

1. Fang, R., Tang, K.D., Snavely, N., Chen, T.: Towards computational models of kinship verification. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 1577–1580. IEEE (September 2010)
2. Xia, S., Shao, M., Fu, Y.: Kinship verification through transfer learning. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI 2011, vol. 3, pp. 2539–2544. AAAI Press (2011)
3. Somanath, G., Kambhamettu, C.: Can faces verify blood-relations? In: IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS (2012)
4. Lu, J., Hu, J., Zhou, X., Shang, Y., Tan, Y.P., Wang, G.: Neighborhood repulsed metric learning for kinship verification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2594–2601 (2012)
5. Vieira, T.F., Bottino, A., Laurentini, A., DeSimone, M.: Detecting Siblings in Image Pairs. The Visual Computer (to appear, 2013)
6. Lu, J., Hu, J., Zhou, X., Shang, Y., Tan, Y.P., Wang, G.: Neighborhood repulsed metric learning for kinship verification (2012), Database available at `https://sites.google.com/site/elujiwen/download`
7. Xia, S., Shao, M., Luo, J., Fu, Y.: Understanding kin relationships in a photo. IEEE Transactions on Multimedia 14(4), 1046–1056 (2012), Database available at `http://www.ece.neu.edu/~yunfu/research/Kinface/Kinface.htm`
8. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008)
9. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
10. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
11. Abdel-Hakim, A.E., Farag, A.A.: CSIFT: A SIFT Descriptor with Color Invariant Characteristics. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, vol. 2, pp. 1978–1983 (2006)
12. Chen, J., Shan, S., He, C., Zhao, G., Pietikäinen, M., Chen, X., Gao, W.: Wld: A robust local image descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1705–1720 (2010)
13. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
14. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
15. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
16. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)

# Are Haar-Like Rectangular Features for Biometric Recognition Reducible?

Kamal Nasrollahi and Thomas B. Moeslund

Visual Analysis of People Laboratory, Aalborg University
Sofiendalsvej 11, 9200 Aalborg, Denmark

**Abstract.** Biometric recognition is still a very difficult task in real-world scenarios wherein unforeseen changes in degradations factors like noise, occlusion, blurriness and illumination can drastically affect the extracted features from the biometric signals. Very recently Haar-like rectangular features which have usually been used for object detection were introduced for biometric recognition resulting in systems that are robust against most of the mentioned degradations [9]. The problem with these features is that one can define many different such features for a given biometric signal and it is not clear whether all of these features are required for the actual recognition or not. This is exactly what we are dealing with in this paper: How can an initial set of Haar-like rectangular features, that have been used for biometric recognition, be reduced to a set of most influential features? This paper proposes total sensitivity analysis about the mean for this purpose for two different biometric traits, iris and face. Experimental results on multiple public databases show the superiority of the proposed system, using the found influential features, compared to state-of-the-art biometric recognition systems.

## 1 Introduction

Biometric recognition, the identification of people based on their biological and/or behavioral characteristics like face, ear, iris, fingerprint, finger vein patterns, hand vein pattern, hand geometry, and gait, is nowadays being used in many real-world applications from security and surveillance systems, to human-computer interaction systems, to gaming, to name a few. Biometric recognition is still a challenging task as the acquired biometric signals (visual signals in this paper) are usually affected by degradation factors like noise corruption, illumination, blurriness, and occlusion. Furthermore, for contactless biometrics (like face, iris, and ear) for which there is a distance between the sensor and subject of interest, the resolution of the acquired image is another important challenge.

Several biometric recognition systems have been developed for dealing with the aforementioned challenges. These methods can be generally divided into two groups: appearance based and feature based. The appearance based algorithms use the grayscale values of the input images directly, while the feature based systems extract some features from the grayscale values and then use these extracted features for the actual recognition. In this paper we focus on two

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 334–341, 2013.

biometric traits, face and iris, but the discussion can be extended to other traits easily. Several features based approaches can be found in the literature for the chosen biometrics. For example for face recognition in [17] local texture features, in [11] local directional number patterns, and in [10] local gradient information are used. For feature based iris recognition systems for example in [19] Gabor filters, in [16] Scale Invariant Feature Transform (SIFT), and in [13] wavelets have been used. Known appearance based methods include, but not limited to, Principal Component Analysis (PCA)-based methods, Independent Component Analysis (ICA) algorithms, Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), and Neural Networks (NN), to name a few. These classifiers have been well applied to the chosen biometrics of this paper. For example for face recognition PCA in [14], LDA in [3], SVM in [5], ICA in [2] and more recently Sparse Representation (SR) based methods in [8] have been used. For iris recognition a Probabilistic NN (PNN) in [19] and an LDA classifier in [21] have been used. The problem with the appearance based algorithms is that they usually need to register the input images to a fixed frame. This means that these methods mostly are sensitive to registration errors. The problem with the feature based methods is that their performance is directly depended on the effectiveness and robustness of the employed features. Furthermore, the performance of both groups of algorithms degrades when the input images are noisy, occluded by some obstacles, of low resolution, and not properly illuminated.

In our very recent work [9] a feature based approach for biometric recognition has been introduced which is shown to be robust against most degradations and poor imaging conditions. The employed features in this system [9] are Haar-like rectangular features that are extracted from integral images. These features are fed to a PNN classifier in [9] for the final recognition. The Haar-like rectangular features were first introduced for rapid and robust object detection using a boosted cascade of simple weak classifiers [15] and have usually been used for the same purpose in the literature (see [9] for more information). The problem with Haar-like rectangular features is that one can define many different such features for a given image, while only few of these features are useful for the actual recognition and the rest just impose extra computations to the system. This varies from a biometric trait to another one. Finding proper sets of Haar-like rectangular features from an initial set of such features is the exact concern of this paper. To do so, the proposed system introduces the Total Sensitivity Analysis (TSA) about the mean, which is further explained later in this paper.

The rest of the paper is organized as follows: biometric recognition using the Haar-like rectangular features of [9] is briefly revisited in the next section, then, TSA about the mean is explained in section 3, experimental results are discussed in section 4, and finally the paper is concluded in section 5.

## 2 Biometric Recognition Using Haar-like Features

The Haar-like rectangular features are obtained by filters composing of two types of regions: white and black regions (see Figure 1). The common way for generating these filters is to consecutively divide the entire area of the filter to 2, 3, ..., $N$ regions [1]. Then, paint these regions to black or white. This can result in many such filters, which some of them, for $N = 20$, are shown in Fig. 1.

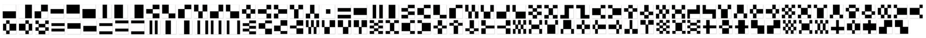


**Fig. 1.** The initial set of Haar-like rectangular filters. The index of the top-left filter is 1 and the one in the right-bottom is 115 (those in between change accordingly).

For calculating the value of a specific Haar-like rectangular feature from a given image, first, the filter is resized to the same size as the input image (without changing the relative size of its black and white regions). Then, the filter is lied on the input image such that the four corners of the filter lie on the four corners of the input image. Then, the summation of those pixel values of the input image that lie in the black region of the Haar-like rectangular filter is subtracted from the summation of those pixel values of the input image that lie in the white region of the filter. To reduce the computational time these features are usually calculated from the integral counterparts of the input image [15].

Having extracted the Haar-like rectangular features of Fig. 1 from the integral image of the input biometric, they are fed to a PNN classifier in [9]. PNN performs the recognition by finding the Probability Distribution Functions (PDF)s of the involved classes using a Parzen window like:

$$f_j(s) = \frac{1}{\sigma_j n_j} \sum_{k=1}^{n_j} W(\frac{||s - s_{kj}||^2}{\sigma_j}) \tag{1}$$

where $f_j$ is the PDF of the $j$th class, $n_j$ is the number of the samples of this class, $\sigma_j$ is a smoothing parameter, $s_{kj}$ contains the features of the $k$th training sample of the $j$th class, $s$ contains the features of the unknown sample, and $W$ is a weighting function. In PNN, $W$ is replaced by an exponential function to use PDFs of Gaussian form (see [12], and [9] for more information on PNN).

## 3 Total Sensitivity Analysis about the Mean

Having explained the Haar-like rectangular features and the employed classifier, in this section TSA is elaborated. Sensitivity analysis is a technique for finding the importance and the influence of the input features to the system [18]. Let's assume that we have a recognition system which takes $A$ features (here all the features shown in Fig. 1) as input to distinguish between $B$ different classes. Having trained the recognition system using the training samples (which are separated from the testing samples), the classical sensitivity analysis about the mean works as follows: first, all the $A$ features of system are extracted for all the testing samples. Then, for $i = 1...A$ fix the values of all the features except the

$i$th one to their mean values. Then, change the value of the $i$th feature between $\pm\sigma$ where $\sigma$ is the standard deviation of this feature. Each time that the value of the $i$th feature is changing, a new set of testing samples is generated which is used for testing the system. During the testing of the system the recognition rates of the system for each individual class are monitored. If changing the value of $i$th feature results in changing the recognition rate of the system for class $b \in B$ , the $i$th feature is considered as an influential feature for recognition of this class.

The classical sensitivity analysis measures the sensitivity of each input feature in recognizing each individual class in a given data. This however can not directly be used as a measure for monitoring the overall recognition rate of the system as improving the recognition rate of the system for one specific class may result in reducing the recognition rate of the system for another class. Therefore, the proposed system introduces the TSA as follows: for each input feature TSA is simply obtained by summing up the results of the classical sensitivity analysis for all the involved classes. It is obvious that TSA of a specific feature increases if changing the value of this feature results in improvement of the recognition rate of the system for a larger number of classes.

Having obtained the results of TSA for all the features, a threshold like $T$ can be found such that any feature with a TSA value larger than $T$ can be considered as an influential feature. The set of the influential features, $F$, is a set of features by which the recognition rate of the system is the same as the recognition rate of the system with the original set of features. It means the rest of the features that have TSA values below $T$ are actually non-contributive features. The exact value of $T$ depends on the employed traits and changes from one trait to another one and can be found experientially. The set of the sensitive features and $T$ change also from a trait's database to another database of the same trait. But there is a good similarity between the sensitive features of one trait from one database to another database of the same trait. It is shown in the experimental results that removing the non-contributive features not only gives higher recognition rates, but it results in a faster system.

## 4 Experimental Results

To show the efficiency of the employed TSA method in discarding the non-contributive features and hence finding the most influential features, multiple public databases of the two biometric traits, iris and face, have been employed. The iris database (ID) has been taken from [7]. This database contains 2240 iris images of 224 subjects each providing 10 grayscale iris images. These images are of size 320×240 pixels (Fig. 2). Four public databases have been employed for face recognition: ORL [22], UMIST [24], Faces94 [23], and Extended YaleB [4]. The number of the images in these databases are 400, 564, 3060, and 16128 images of 40, 20, 153, and 28 subjects, respectively. The sizes of the images are 92×112, 92×112, 105×120, and 168×192 pixels, respectively. These images contain variations in head-pose, expression, and illumination conditions (Fig. 2).

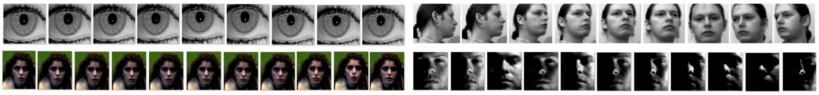


**Fig. 2.** Some samples of four of the employed databases, from top, clockwise: ID, UMIST, Extended YaleB, and faces94 databases

The reported results in this section are obtained when the available databases are divided randomly to three parts for training, cross-validation, and testing. The sizes of each of these portions are 60%, 15%, and 25% of the entire database, respectively. The degradation in the performance of the classifier when the sizes of these three parts change is studied in [9].

The proposed system has gone through three experiments. In the first experiment, for each individual database the recognition rate of the proposed system is compared against the state-of-the-art systems when the proposed system is trained using the entire set of the extracted Haar-like rectangular features (shown in Fig. 1). The results are shown in Fig. 3. In this figure, the results of the proposed system (PS) for iris are compared against S1-S4 which are decision tree-based, appearance based PNN, SVM, and fuzzy binary decision tree-based classifiers, respectively [6]. The results for face are compared against PCA [14], LDA [3], SVM [5], ICA [2], Local Binary Patterns (LBP), and some very recent Sparse Representation (SR) based methods, DDSR, FDDL, RPCA. The results of these methods on ORL, UMIST, and YaleB are reported in [8]. It should be mentioned that some of these methods like PCA, ICA and LDA are also based on feature reduction concept.

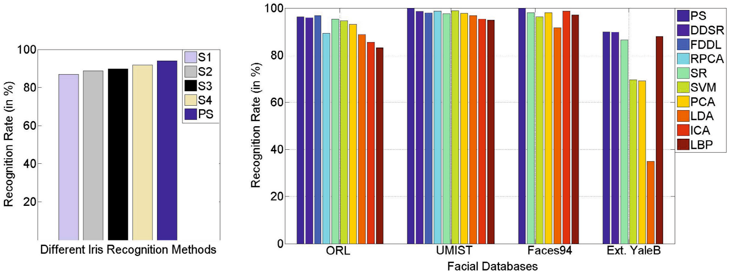


**Fig. 3.** The recognition rate of the proposed system against: (left) state-of-the-art iris recognition algorithms using ID database and (right) state-of-the-art face recognition algorithms using ORL, UMIST, Faces94, and Extended YaleB databases

In the second experiment the explained TSA method (section 3) is applied to the entire feature set to find the most influential features and discard the non-contributive ones. To do so, for each database we define a set of most influential features, $F$, which is initially empty. Having obtained the TSA values of all the

Haar-like rectangular features, we keep adding features to $F$ based on their TSA values in a descending order. Every time a new influential feature is added to $F$, the employed PNN is trained and tested (The training and testing samples are kept separate from each other). This process continues until the recognition rate of the proposed system using $F$ is the same (within $\pm$ 0.005) as the recognition rate of the proposed system using the entire set of the Haar-like rectangular features. The results of applying TSA to four of the employed databases using the initial set of features are shown in Fig. 4. The second experiment reduces these initial sets of features to the 44, 30, 39, 41, and 44 most influential features for ID, ORL, UMIST, Faces94, and Extended YaleB databases, respectively. It means that for each of these databases only these numbers of top influential features are enough for achieving the same recognition rate as the case where the entire Haar-like rectangular features are used.

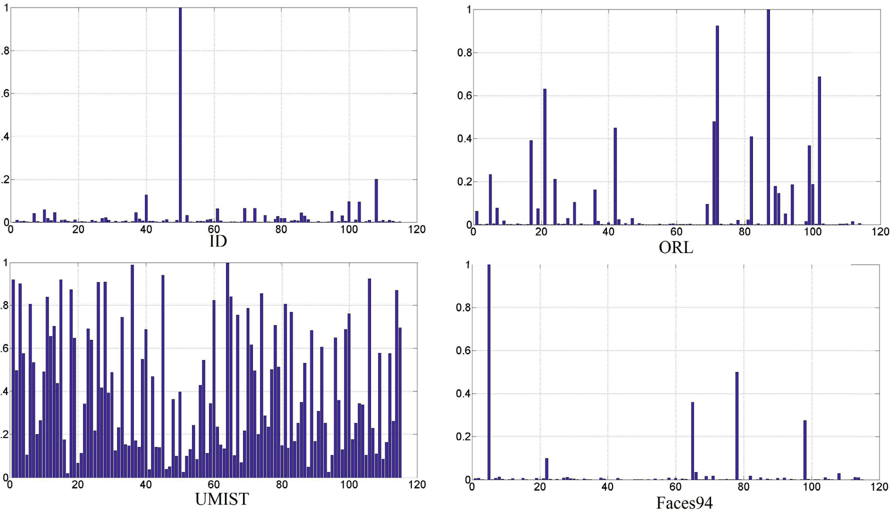


**Fig. 4.** The normalized results of the employed TSA method applied to the Haar-like rectangular features obtained from four of the employed databases. The $x$ axis in a) and b) represents the name of the Haar-like rectangular features from Fig. 1.

The most influential features of the facial databases (the features with highest TSA value in Fig. 4) change from one database to another one. However, it can be seen from Fig. 4 that the most influential features of one of the databases is usually among the top influential features of the other ones. It may seem as a drawback for the employed TSA method that the most influential features of these databases are not completely the same. But this actually makes sense as the images of these databases are captured under very different imaging conditions (Fig. 2. For example, ORL images are well focused, UMIST images have wide head poses, Faces94 images are not of good quality in terms of illumination, and Extended YaleB images mostly suffer from directional illumination). The interesting point is that regardless of the content of the database the set of possible Haar-like rectangular features that can be extracted from the database can be summarized to a set of influential features like $F$.

The third and last experiment compares the computational time of the proposed system against the systems of [9]. This timing information is shown in Fig. 5. It is obvious from this figure and Fig. 3 that beside achieving the same recognition rates of [9] the proposed system works faster.

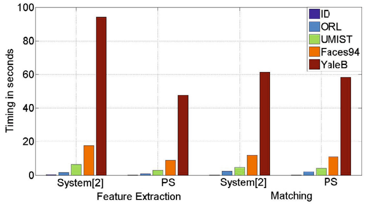


**Fig. 5.** Timing comparison of the proposed system against [9]

## 5 Conclusion

This paper proposes a biometric recognition system using Haar-like rectangular features which mostly have been used for object detection. The set of these features has proved to result in high recognition performance, but the problem is that this set may contain many different number of features while only few of them contribute to the actual recognition and the rest of the features are non-contributive. For finding and discarding the non-contributive features this paper uses total sensitivity analysis about the mean. Experimental results on two types of biometric traits, iris and face, show that total sensitivity analysis can find these most influential features which can result in a fast and reliable biometric recognition system.

## References

1. Chesnokov, Y.: Face Detection C++ Library with Skin and Motion Analysis. Biometrics AIA 2007 TTS (2007)
2. Choi, J., Juneho, Y., Turk, M.: Effective representation using ICA for face recognition robust to local distortion and partial occlusion. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1977–1981 (2005)
3. Etemad, K., Chellappa, R.: Discriminant analysis for recognition of human face images. Journal of the Optical Society of America A 14(8), 1724–1733 (1997)
4. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 643–660 (2001)

5. Hotta, K.: Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel. Image and Vision Computing 26(11), 1490–1498 (2008)
6. Kumar, A., Hanmandlu, M., Das, A., Gupta, H.M.: Biometric based personal authentication using fuzzy binary decision tree. In: Proceedings of 5th IAPR International Conference on Biometrics, pp. 396–401 (2012)
7. Kumar, A., Passi, A.: Comparison and combination of iris matchers for reliable personal authentication. Pattern Recognition 43(3), 1016–1026 (2010)
8. Ma, L., Wang, C., Xiao, B., Zhou, W.: Sparse representation for face recognition based on discriminative low-rank dictionary learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2586–2593 (2012)
9. Nasrollahi, K., Moeslund, T.B., Rashidi, M.: Haar-like Rectangular Features for Biometric Recognition. In: Proceedings of 6th IAPR International Conference on Biometrics (2013)
10. Ngoc-Son, V.: Exploring Patterns of Gradient Orientations and Magnitudes for Face Recognition. IEEE Transactions on Information Forensics and Security 8(2), 295–304 (2013)
11. Rivera, A.R., Castillo, J.R., Chae, O.: Local Directional Number Pattern for Face Analysis: Face and Expression Recognition. IEEE Transactions onImage Processing 22(5), 1740–1752 (2013)
12. Specht, D.F.: Probabilistic neural networks. Neural Networks 3, 109–118 (1990)
13. Szewczyk, R., Grabowski, K., Napieralska, M., Sankowski, W., Zubert, M., Napieralski, A.: A reliable iris recognition algorithm based on reverse biorthogonal wavelet transform. Pattern Recognition Letters 33(8), 1019–1026 (2012)
14. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
15. Viola, P., Jones, M.: Rapid object detection using a boosted classifier of simple features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1(1), pp. 511–518 (2001)
16. Xiaomin, L., Peihua, L.: Tensor decomposition of SIFT descriptors for person identification. In: IAPR International Conference on Biometrics, pp. 265–270 (2012)
17. Xiaoyang, T., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Transactions on Image Processing 19(6), 1635–1650 (2010)
18. Yooyoung, L., Filliben, J.J., Micheals, R.J., Phillipstitle, P.J.: Sensitivity analysis for biometric systems: A methodology based on orthogonal experiment designs. Computer Vision and Image Understanding 117(5), 532–550 (2013)
19. Yung-Hui, L., Savvides, M.: An Automatic Iris Occlusion Estimation Method Based on High-Dimensional Density Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(4), 784–796 (2013)
20. Zhu, Z., Morimoto, T., Adachi, H., Kiriyama, O., Koide, T., Mattausch, H.J.: Multi-view face detection and recognition using haar-like features. In: Proceedings of COE Workshop (2006)
21. Chengqiang, L., Mei, X.: Iris Recognition Based on DLDA. In: Proceedings of 18th International Conference on Pattern Recognition, vol. 4, pp. 489–492 (2006)
22. ATT Laboratories Cambridge, The ORL Database of Faces, `http://www.cl.cam.ac.uk/research/dtg/`
23. Collection of Facial Images: Faces94, `http://cswww.essex.ac.uk/mv/allfaces/faces94.html`
24. Image Engineering Laboratory, The Sheffield UMIST Face Database, `http://www.sheffield.ac.uk/eee/research/iel/research/face`

# A New Approach to Detect Splice-Sites Based on Support Vector Machines and a Genetic Algorithm

Jair Cervantes[1], De-Shuang Huang[2], Xiaoou Li[3], and Wen Yu[4]

[1] Posgrado e Investigacíon, UAEM-Texcoco, Av. Jardín Zumpango s/n
Fraccionamiento El Tejocote, Edo. Mex., C.P. 56259
jcervantesc@uaemex.mx

[2] Department of Control Science & Engineering, Tongji University
Cao'an Road 4800, Shanghai, 201804 China
dshuang@tongji.edu.cn

[3] Departmento de Computación, CINVESTAV-IPN, Mexico City, Mexico
lixo@cs.cinvestav.mx

[4] Departamento de Control Automático, CINVESTAV-IPN, Mexico City, Mexico
yuwen@ctrl.cinvestav.mx

**Abstract.** This paper presents a method for classification of imbalanced splice-site classification problems, the proposed method consists of the generation of artificial instances that are incorporated to the dataset. Additionally, the method uses a genetic algorithm to introduce just instances that improve the performance. Experimental results show that the proposed algorithm obtains a better accuracy to detect splice-sites than other implementations on skewed data-sets.

**Keywords:** SVM, Skewed datasets, Classification DNA splice sites.

## 1 Introduction

Recognizing boundaries of exons and introns is a challenging task in DNA sequence analysis. To identify exons into DNA sequences present a computational challenge due to the genes in many organisms splices of different way. Moreover, most of gene datasets are imbalanced and the bulk of classifiers generally performs poorly on imbalanced datasets because making the classifier too specific may make it too sensitive to noise and more prone to learn an erroneous hypothesis. Moreover, sometimes an instance can be treated as noise and ignored completely by the classifier if the dataset is imbalanced. Consequently, an effective detection of splice sites requires not just to know features, dependencies, relationship of nucleotides in the splice site surrounding region or an effective encoding method, but also a good method which tackles the disadvantage of imbalanced in datasets. In this paper, we use a novel SVM approach to detect splice sites in imbalanced datasets. The proposed method generates new synthetic instances in a similar form of SMOTE [5], the key idea of this model is to introduce artificial instances in the

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 342–349, 2013.

region of positive SV, decreasing the skew of the margin and improving the generalization capacity. The proposed technique not only modifies the margin also modifies the region of the minority class improving the generalization power of the classifier. However, to introduce incorrectly new synthetic instances can reduce the classifier performance because this is a sensible region to small changes. To avoid this fundamental issue, we incorporate a Genetic Algorithm which guides the search in the sensible region generating intelligently new synthetic instances. The proposed algorithm, tackles the disadvantage of imbalanced data-sets with SVM. The rest of the paper is organized as following: Section 2 shows the SVM imbalanced problem. Section 3 focuses on explaining the methodology of proposed SVM classification algorithm. Section 4 shows experimental results. Conclusions are given in Section 5.

## 2 Related Work

SVM has received considerable attention due to its optimal solution, discriminative power and performance. SVM has been applied in many fields, some SVM algorithms have been used in splice site detection with acceptable accuracies. There are a lot of works about Splice sites detection with several methods in the literature. However, the works most representative of splice sites detection with SVM are [1] [3] [4]. Baten [1] uses SVM with polynomial kernel to obtain an effective detection of splice sites, Cheng [2] uses SVM to predict mRNA polyadenylation sites [poly(A) sites], the method helps to identify genes, define gene boundaries, and elucidate regulatory mechanisms, [3] and [4] use SVM to detect splice-junction (intron-exon or exon-intron) sites in DNA sequences. In all this works, the accurate splice-site detection is a critical component of all analytic techniques. However, before mentioned methods do not consider datasets with high imbalance. Lately has been showed that SVM performance drops significantly with imbalanced data-sets. Some important algorithms based on Undersampling, Oversampling or SMOTE techniques[5] had been developed to tackle this problem. SMOTE over-samples the minority class by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the $k$ minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the $k$ nearest neighbors are randomly chosen. The SMOTE technique is better than under-sampling and over-sampling and a promising technique to tackle this problem. Some other proposals inspired in SMOTE can be seen in [6].

Methods based on Genetic Algorithms (GA) have also been pursued to tackle imbalanced problems. Since evolutive methods provide state-of-the-art techniques for many of todays data engineering applications, the use of evolutive methods to understand imbalanced learning has naturally attracted growing attention recently. Zou et al. [7] use a GA to balance the data-sets. In [8] the authors propose a classification system using hierarchical fuzzy rule and a genetic algorithm to select the most important rules and to eliminate conflicting rules or rules which perturb the performance. Garcia et al. [9] implement an algorithm which performs an optimized selection of previously defined generalized examples obtained by a heuristic. An excellent state of the art about imbalanced classification can be found in [10]. Despite the early proposed methods to improve the performance, these algorithms use GA's to balance the data-sets, obtain rules or to select instances intelligently, but not to generate new instances as the proposed algorithm. The proposed algorithm permits to create new instances and evaluate the discriminative power of these new instances in the data set. The region where the instances are created (region of SV) retains valuable information, but is necessary to use a GA to guide the search of best instances.

## 3 Methodology

The proposed algorithm is based in the sparse property of SVM, where the solution is given for a small subset from the original data-set called Support Vectors (SV). Formally, given a data set $\{(x_i, y_i)\}_{i=1}^{n}$ and separating hyperplane $f(x) = w_i^T x + b = 0$, the shortest distance from separating hyperplanes to the closest positive example and closest negative example in the non separable cases are given by

$$\gamma_+ = \min \gamma_i, \forall \gamma_i \in class + 1 \tag{1}$$

$$\gamma_- = \min \gamma_i, \forall \gamma_i \in class - 1 \tag{2}$$

where $\gamma_i$ is given by

$$\frac{y_i(w_i^T K \langle x_i \cdot x_j \rangle + b_i)}{\|w\|} \tag{3}$$

Margin is the optimal separating hyperplane obtained by training a SVM and it is given by $\gamma = \gamma_+ + \gamma_-$. This algorithm takes advantage of this fact, the key idea of this model is to introduce artificial instances from positive SV, It permits not only modify the margin also modify the region of the minority class and decrease the skew of the margin, but also improve the generalization capacity. Is clear that, introduce new synthetic instances in this region can affect negatively the SVM performance by introducing noise in the data-set. However, in this paper we use a GA to guide the search of the best regions and include just the best data instances in the margin region. Figure 1 shows the framework of the proposed method.

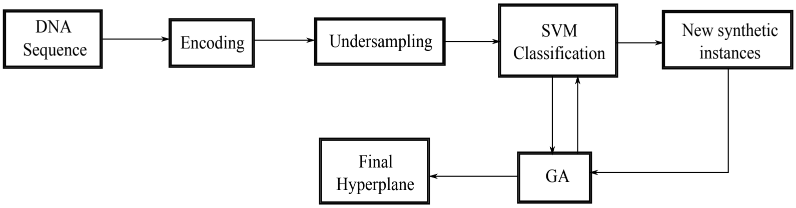


**Fig. 1.** Stages of the proposed algorithm

### 3.1 DNA Encoding

DNA sequences are given as strings of nucleotides and is necessary to encode it. Sparse encoding is a widely used encoding schema which represents each nucleotide with four bits: $A \to 1000, C \to 0100, G \to 0010$ and $T \to 0001$ [11]. We use 18 additional features with the sparse encoding schema. The first 16 components define the nucleotide pairs into a DNA sequence, which are defined by $\beta = \{(x_{AA}), (x_{AC}), (x_{AG}), (x_{AT}), \ldots, (x_{TA}), (x_{TC}), (x_{TG}), (x_{TT})\}$. When some nucleotide pair is in the sequence, it is marked with 1 and an absence of this pair is marked with 0. The last two components correspond to the informative function of each triples in the sequence ranked by their *F-value*. For each triple, we specify its location relative (pre and post) and its mean frequency between exons and decoys $\mu_k^+ - \mu_k^-$ respectively.

The *F-value* criterium is that used by Golub et al [12]. For each triple $x_k$,$k = 1, ..., n$, we calculated the mean $\mu_k^+(\mu_k^-)$ and the standard deviation $\sigma_k^+(\sigma_k^-)$ using positive and negative examples. The *F-value* criterium is given by

$$F(x_k) = \left| \frac{\mu_k^+ - \mu_k^-}{\sigma_k^+ + \sigma_k^-} \right| \tag{4}$$

where $x_k$ is the $k - esime$ triple, the *F-value* serves as a simple heuristic for ranking the triples according to how well they discriminate. The last point in the vector is represented by the relative presence of each triple of nucleotides.

This encoding schema allows to obtain the nucleotides of each sequence, showing the importance of some pairs in the sequence, and obtaining the importance of each triple at the begin and at the end of each sequence.

### 3.2 Classification Algorithm

The first step in the proposed algorithm consists in encode the DNA sequence, we use the method described early. In the next step, the algorithm obtains subsets from the entire data-set. To separate input data set, 70% of examples from data set are selected as training set labeled as $tr$. We select $tr$ with 70%, $tf$ 15% and $te$ 15% of input data maintaining almost equal proportion in class distribution over the data. For instance, if there are two class values (say $X^-$ and $X^+$) in a classification problem $P$ with 1000 examples in total, and the number of examples of class-types: $X^-$ and $X^+$ are respectively 800 and 200. Then, 560 and 140 examples of class-types $X_{tr}^-$ and $X_{tr}^+$ respectively are assumed to be included into $tr$ by random selection, and $X_{tf}^-$,$X_{tf}^+$, $X_{te}^-$ and $X_{te}^+$ with 120, 30, 120 and 30 examples respectively. Figure 1 show the steps of proposed algorithm which are described in detail in the algorithms 1 and 2. $X_{tr}^+$ and $X_{tr}^-$ are used to train a SVM and to find an introductory hyperplane $H_1$ $(X_r^+, X_r^-)$, from $H_1$ we obtain the SV $x_{svi}^-$ and $x_{svi}^+$ and generate new synthetic examples from it. We use the SMOTE technique to generate the first population of new synthetic instance $x_{svg}$. which is given by $x_{svg} = x_{svi}^+ + \delta \cdot (x_{svi-n}^+ - x_{svi}^+)$, where $x_{svg}$ denotes one synthetic instance, $x_{svi-n}^+$ is the nearest neighbors of $x_{svi}^+$ in the positive class, and $\delta \in [0, 1]$. This procedure is repeated for all the positive instances. The initial

**Algorithm 1.** General SVM classification procedure

```
Input: Nucleotides Sequence Output: Improved hyperplane $H_f : (X_{te}^{+}, X_{te}^{-})$

1. Encode the nucleotides sequences $\{x_i \in X : y = \pm 1\}, i = 1, \ldots, n$
2. From $X^{+}$ and $X^{-}$ obtain $X_{tr}^{+}$, $X_{tr}^{-}$, $X_{tf}^{+}, X_{tf}^{-}$, $X_{te}^{+}$, $X_{te}^{-}$ with 70%, 15% and 15%
   respectively
3. Train $SVM$ with $\left(X_{tr}^{+}, X_{tr}^{-}\right) \rightarrow H_1$
4. Obtain SVs $x_{svi}^{-}$ and $x_{svi}^{+}$ from $H_1$
5. Obtain initial population according to (3.2).
6. Obtain best data points $(X_{GA}^{+}, X_{GA}^{-})$ using the GA (Algorithm 2)
7. Obtain final hyperplane $trainSVM(X_{GA}^{+}, X_{GA}^{-}) \rightarrow H_f$
```

population is conformed by $x_{svi}^{-} \cup x_{svj}^{+} \cup x_{svg}$. It is manipulated using several genetic operators to improve the population in each iteration and optimizing the solution, i.e. DNA sequences are slightly modified from the DNA sequences with best discrimination power improving the classifier performance, this process is obtained by the GA defined in algorithm 2.

Second algorithm describes the functioning of the GA. We used a gray coding to represent each individual in the population and the fitness function.

Genetic operators can find a solution in a small space by crossover operators, and explore new areas in the space by mutation operators. The fitness function ensures the evolution towards optimization by the fitness score for each DNA sequence with high discriminative power in the population. The process continues until a predefined termination criterion has been met.

In the proposed technique, we use the F-measure as fitness, it provides a way to arrive the search solutions, and also controls the selection process. F-measure is defined by

$$\frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

where precision $= \frac{TP}{TP+FP}$ and recall $=\frac{TP}{TP+FN}$, $TP$ represents true positive rate defined by the fraction of true positives out of the positives and $FP$ false positive rate defined by the fraction of false positives out of the negatives.

Selection is based in ranking selection with elite preserving. Each individual survives in the next generation in proportion to the rank of its fitness value. The best individual in the population is made to remain to the next generation in order to prevent the best individual from being eliminated by stochastic genetic drifts.

In the proposed algorithm, we used crossover and mutation operators. Crossover operator unifies the genetic information of two individuals (parents), obtained by selection operator, and creates two new individuals (children) called as offspring. We use two points crossover. A crossover operator permits the fitness function to evolve towards optimization. The mutation operator helps to find the global optimal solution to the problem. It is called exploration operator. We use a crossover probability of $p_c = 0.9$, and a mutation probability

**Algorithm 2.** GA to generate artificial data of the minority class

**Input:** Initial population $X_{svg} = (x_{svg1}, x_{svg2}, \ldots, x_{svgm})$, Max generation. **Output:** Best data instances $(X_{GA}^{+}, X_{GA}^{-})$

```
1. m(k)=m(0)=m
2. for i=1 to m(k)
3.     $H_a \leftarrow trainSVM\left(x_{svi}^{-} \cup x_{svi}^{-} \cup x_{svg}(i)\right)$
4.     Obtain fitness from $H_a$ with $\left(X_{tf}^{+}, X_{tf}^{-}\right)$ by (5).
5. end for
6. Generate new population $X_{Nsvg}$ by selection, crossover and mutation.
7. Add the best individual in the current population $X_{svg}$ to the newly generated
   $X_{Nsvg}$ to form the next population.
8. $m(k)$ = size of new generation $X_{Nsvg}$
9. Return to 2 if the pre-specified stopping condition is not satisfied.
```

of $p_m = 1/n$, where $n$ is the string length for Gray coded. Final hyperplane is obtained until a stop criterion has been met.

Classical methods cannot decide which new instances will improve the SVM performance in imbalanced data-sets, because the search space is often huge, complex or poorly understood. GA has the ability to explore large and new areas. Finding new instances with discriminative power can be considered a GA search problem. The crossover and mutation operators realize the search exploratory and exploitative respectively. Thus, to use GA improves the SVM performance by generating artificial instances. The new instances obtained by the GA $(X_{GA}^{+}, X_{GA}^{-})$ contain information with high discriminative power helping to increase the classifier performance.

## 4 Experimental Results

We conducted experiments on some imbalanced and balanced intron-exon data-sets taken from $http : //www.raetschlab.org/suppl/MITBookSplice/files/$, *www.archive.ics.uci.edu* and $http : //big.crg.cat/bioinformatics_and_genomics/$ Table 1 shows details of these data-sets. We compared our method against: Under-Sampling, Over-Sampling and SMOTE techniques. The proposed method and the methods before mentioned are implemented in Matlab. To evaluate classifiers on skewed data-sets, require to use an adequate metric. We report the results with True Positive Rate (TPR), False Positive Rate (FPR), Area Under the Curve (AUC) and F-measure metrics. In all the experiments, we used 10-fold cross validation.

Table 1, shows data-sets used in experimental results, length of DNA sequence (ls), imbalance ratio (r) and size of exons, acceptors, donor (positive instances) and introns, decoys (negative instances).

Table 2, shows the results obtained in our experiments. The first column shows the data-set used and next columns show the experimental results obtained over

**Table 1.** Imbalanced ratio and size of the data-sets used in experimetal results

| *Dataset* | *ls* | *size* | *r* |
|---|---|---|---|
| Nobgrors (No_23) | 23 | 111827 | 1:1 |
| Nobgrors (No_09) | 9 | 110824 | 1:1 |
| Nobgrors (Starts) | 20 | 9299 | 1:1 |
| Nobgrors (Stops) | 15 | 11077 | 1:1 |
| Acc_23 (Ac_23) | 23 | 124728/374184 | 2:1 |
| Acc_39 (Ac_39) | 39 | 120000/360000 | 2:1 |
| Donor_9 (Do_09) | 9 | 120000/360000 | 2:1 |
| Genbank64 (EI_60) | 60 | 767/2422 | 3.15:1 |
| Genbank64 (IE_60) | 60 | 768/2421 | 3.15:1 |
| C. elegans (Ac_60) | 60 | 2785/91546 | 31:1 |
| C. elegans (Do_60) | 60 | 2785/89163 | 31:1 |

the 11 datasets with four different metrics measure methods, as well as the results obtained using the proposed method, Under-sampling, Over-sampling and SMOTE approaches. The best result for each classifier is highlighted in bold. We report the average on 30 runs for the proposed method. In all the cases, the proposed method got the highest F-measure (we used F-measure as fitness function in the GA), it suggests that the GA works well as a search engine that helps to find perfectly what new instances improve the classifier performance. In experimental results obtained not only F-measure performance is better, but also sometimes AUC-ROC and TP measures are improved. Moreover, the improvement can be carried in balanced data-set (Table 2).

The experimental results show that the proposed algorithm helps to improve the classification accuracy. The proposed algorithm helps to reduce the false positive rate (see Table 2 -EI_60, Ac_60, Do_60, Ac_23-) or helps to increase the true positives rate (IE_60, Do_60, Ac_23, No_09) with by adding artificial data. The improvement in the performance depends directly of the fitness function used. To use F-measure as fitness function helps to improve effectively FP rate too, but sometimes to improve can affect the AUC-ROC and TP rate due to imbal-

**Table 2.** Comparison of the proposed method (PM) with some measure metrics against other techniques for imbalanced data-sets

| *Measure* | Undersampling | | | | Oversampling | | | | Smote | | | | PM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *TP* | *FP* | *ROC* | *Fm* | *TP* | *FP* | *ROC* | *Fm* | *TP* | *FP* | *ROC* | *Fm* | *TP* | *FP* | *ROC* | *Fm* |
| EI_60 | 0.979 | 0.097 | 0.941 | 0.918 | 0.992 | 0.037 | 0.983 | 0.968 | 0.99 | 0.038 | 0.99 | 0.965 | 0.995 | **0.009** | 0.994 | **0.99** |
| IE_60 | 0.905 | 0.015 | 0.972 | 0.919 | 0.939 | **0.005** | 0.983 | 0.929 | 0.941 | 0.012 | 0.978 | 0.946 | 0.973 | 0.029 | 0.979 | **0.958** |
| Ac_60 | 0.958 | 0.359 | 0.856 | 0.96 | 0.974 | 0.318 | 0.917 | 0.963 | 0.965 | 0.142 | 0.97 | 0.973 | 0.969 | **0.128** | 0.983 | **0.973** |
| Do_60 | 0.952 | 0.335 | 0.807 | 0.953 | 0.972 | 0.357 | 0.861 | 0.93 | 0.972 | 0.323 | 0.971 | 0.963 | 0.976 | **0.243** | 0.974 | **0.977** |
| Ac_23 | 0.593 | 0.203 | 0.802 | 0.580 | 0.577 | 0.212 | 0.754 | 0.564 | 0.589 | 0.205 | 0.801 | 0.573 | 0.604 | **0.198** | 0.788 | **0.593** |
| Ac_39 | 0.411 | 0.245 | 0.676 | 0.446 | 0.42 | 0.223 | 0.726 | 0.451 | 0.433 | **0.222** | 0.732 | 0.46 | 0.468 | 0.25 | 0.712 | **0.493** |
| Do_09 | 0.628 | 0.187 | 0.777 | 0.596 | 0.614 | 0.193 | 0.777 | 0.610 | 0.631 | 0.175 | 0.812 | 0.598 | 0.618 | **0.171** | 0.814 | **0.605** |
| No_23 | 0.926 | 0.104 | 0.926 | 0.926 | 0.894 | 0.107 | 0.954 | 0.894 | 0.914 | 0.091 | 0.969 | 0.91 | 0.914 | **0.087** | 0.969 | **0.934** |
| No_09 | 0.925 | 0.075 | 0.974 | 0.925 | 0.938 | **0.051** | 0.970 | 0.938 | 0.929 | 0.071 | 0.976 | 0.929 | 0.943 | 0.057 | 0.973 | **0.942** |
| Starts | 0.753 | 0.247 | 0.833 | 0.753 | 0.752 | 0.248 | 0.836 | 0.752 | 0.77 | **0.228** | 0.850 | 0.770 | 0.827 | 0.230 | 0.857 | **0.832** |
| Stops | 0.629 | 0.371 | 0.629 | 0.629 | 0.611 | 0.389 | 0.658 | 0.611 | 0.63 | **0.370** | 0.684 | 0.630 | 0.635 | 0.379 | 0.688 | **0.640** |

ance ratio. Therefore, to obtain a fitness function that improves the measures whithout loss in a metric on imbalanced data-sets can be a future research work.

## 5 Conclusions

In this paper, we present a novel SVM classification approach for detection of splice sites. The proposed approach obtains new synthetic instances from the SVs obtained in a first stage and includes just the instances that improve the SVM performance in the data-set. The algorithm uses a GA to evaluate and obtain better instances in each iteration. Experiments done with DNA sequences, show that the information adjoined by the synthetic instances, help to improve the SVM performance. However, the cost of evaluating each solution in the population is very high and despite the good accuracy obtained its complexity is prohibitive in large data sets.

## References

1. Baten, A., Chang, B., Halgamuge, S., Li, J.: Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 7, S15 (2006)
2. Yiming, C., Robert, M.M., Bin, T.: Prediction of mRNA polyadenylation sites by support vector machine. Bioinformatics 22(19), 2320–2325 (2006)
3. Damaevicius, R.: Splice Site Recognition in DNA Sequences Using K-mer Frequency Based Mapping for SVM with Power Series Kernel. In: CISIS 2008, pp. 687–692 (2008)
4. Jing, X., Doina, C., Susan, B.: Exploring Alternative Splicing Features Using SVM. In: Proc. 2008 IEEE Int. Conf. on Bioinf. and Biomed, pp. 231–238 (2008)
5. Chawla, N., Bowyer, K., Hall, L.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 321–357 (2002)
6. Nguyen, H., Cooper, E., Kamei, K.: Borderline over-sampling for imbalanced data classification. Int. J. Knowl. Eng. Soft Data Paradigm 3(1), 4–21 (2011)
7. Zou, S., Huang, Y., Wang, Y., Wang, J., Zhou, C.: SVM learning from imbalanced data by GA sampling for protein domain prediction. In: ICYCS 2008, pp. 982–987 (2008)
8. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50(3), 561–577 (2009)
9. García, S., Derrac, J., Triguero, I., Carmona, C.J., Herrera, F.: Evolutionary-based selection of generalized instances for imbalanced classification. Knowledge-Based Systems 25(1), 3–12 (2012)
10. Haibo, H., Garcia, E.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21(9), 1263–1284 (2009)
11. Zhang, X.H.-F., Heller, K.A., Hefter, I., Leslie, C.S.: Sequence Information for the Splicing of Human Pre-mRNA Identified by SVM Classification. Genome Research 13, 2637–2650 (2003)
12. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)

# Speaker Verification Using Accumulative Vectors with Support Vector Machines

Manuel Aguado Martínez, Gabriel Hernández-Sierra, and José Ramón Calvo de Lara

Advanced Technologies Application Center, Havana, Cuba
{maguado,gsierra,jcalvo}@cenatav.co.cu

**Abstract.** The applications of Support Vector Machines (SVM) in speaker recognition are mainly related to Gaussian Mixtures and Universal Background Model based supervector paradigm. Recently, has been proposed a new approach that allows represent each acoustic frame in a binary discriminant space. Also a representation of a speaker - called accumulative vectors - obtained from the binary space has been proposed. In this article we show results obtained using SVM with the accumulative vectors and Nuisance Attribute Projection (NAP) as a method for compensating the session variability. We also introduce a new method to counteract the effects of the signal length in the conformation of the accumulative vectors to improve the performance of SVM.

**Keywords:** speaker recognition, binary values, accumulative vectors, Support Vector Machine, Nuisance Attribute Projection.

## 1 Introduction

Currently SVM is one of the most robust and powerful discriminative classifier in speaker recognition. The applications of SVM are mainly related to Gaussian Mixtures and Universal Background Model (GMM/UBM) based supervector paradigm [1, 2]. Generally a supervector is obtained by concatenating the means of the adapted GMM models. However, these approaches show limitations associated with the GMM/UBM paradigm. First, it is difficult to exploit temporal or sequential information. Second, the supervector space don´t allows working directly with discriminant aspects of speaker.

A new approach that attempts to reduce these limitations was proposed in [3]. It deals directly with the speaker discriminant information in a discrete and binary space. Our method to obtain the binary representation and then the accumulative vectors is similar to [4], it only differs on the normalization process used for reducing the susceptibility of the accumulative vectors to the signal length. At this point we introduce a new method for successfully accomplish this task since it shows better performance combined with the SVM.

In this article we used SVM as a classifier to work with the accumulative vectors and then we compared the results with those obtained with the GMM/UBM based supervector paradigm. We also use Nuisance Attribute Projection (NAP) as a method

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 350–357, 2013.

for compensating the session variability because this algorithm intend to reduce the susceptibility of SVM kernel to this problem [5].

This paper is organized as follows. Section 2 explains the process to obtain the accumulative vectors presented in [4]. Section 3 briefly describes SVM paradigm and NAP algorithm. In Section 4 are introduced the proposals: a new method for scaling the accumulative vectors and the use of SVM as a classifier for accumulative vectors. Section 5 describes the experimental setup, presents the results obtained and show advantages and disadvantages of the proposed approach. Finally, section 6 gives some conclusions.

## 2 Accumulative Vectors

The process to obtain the accumulative vectors is mainly composed by three components. First, a UBM is trained to divide the acoustic space in acoustic classes. Second, a set of Gaussians components is incorporated to each component or acoustic class of the UBM. These components are known as "speaker specificities" and the set of those as generator model. Finally, for an acoustic frame, each speaker specificity is evaluated and it corresponding binary value is established.

The role of the generator model is to highlight the speaker specificities. As mentioned, each acoustic class of the UBM is represented by a set of Gaussian Components. Those specificities are obtained from the adapted models of the training set, by matching the $i$ components of the adapted models to the $i$ component of the UBM. Since the specificities number is assumed to be very large, it is necessary to reduce it, selecting the most important [4]. As a result the number of specificities per acoustic class could not be the same.

For obtaining the binary representation of a given speaker, first we took each acoustic frame and determine its posterior probability related with each Gaussian component of the UBM by a process similar to Maximum a Posteriori (MAP) [6]. Then the $K$ components with the highest probability were selected, the specificities of these components are the ones represented in the binary vector. We use $K = 3$ based on previous results presented in [4]. Then for each component is compute the likelihood of each acoustic frame with all the corresponding specificities. The equations for determine the posterior probability and the likelihood are detail described in [7]. Finally a binary vector is created by set in 1 the components of the vector corresponding to the specificities with the higher likelihood. These are known as the "active components". After that, a binary vector for each acoustic frame is obtained. Pooling these vectors we have a binary matrix that represents a given speaker. The accumulative vector is then obtained by setting the component of the vector corresponding to a given specificity to the number of activations.

## 3 Support Vector Machines

At the most basic level, SVM is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In the speaker verification, one class is the vector of the target speaker (labeled as +1), and the other class is composed for the vectors of a background population (labeled as -1). Using this information, SVM intends to find a separating hyperplane that maximizes the margin of separation between these two classes. This is an optimization problem defined by:

$$Min \ \ \frac{1}{2}||w||^2 + C\sum_{i} \varepsilon_i \tag{1}$$

$$s.a \ \ t_i(w' * x_i \ - b) - 1 + \varepsilon_i \geq 0 \ \forall_i \tag{2}$$

$$\varepsilon_i \geq 0 \ \forall_i \tag{3}$$

where $w$ is a orthogonal vector to the separating hyperplane, $x_i$ is the accumulative vector $i$, $t_i$ is the class of the vector $i$, $\varepsilon_i$ are the "slack variables" (allows for violations of the constraints since in practice the data is not linearly separable), $b$ and $C$ are constants.

Here $\sum_i \varepsilon_i$ is the penalty or loss function and could be interpreted as a measure of "how bad" the violations are. The constant $C$ controls the tradeoff between penalty and margin.

This optimization problem is solved in a space of dimension higher than the original space, due to the solution is easier to find in it. To achieve this transformation, SVM use a kernel function $K(x_i, x_j)$. The kernel function should satisfy the Mercer condition [8] (The Kernel should be positive semi definite) and therefore can be expressed as:

$$K(x_i, x_j) = \langle \theta(x_i), \theta(x_j) \rangle = \theta(x_i)^T * \theta(x_j) \tag{4}$$

were $\langle \theta(x_i), \theta(x_j) \rangle$ is the inner product of two vectors. Given a test vector $y$ the discriminant function of SVM is given by:

$$f(y) = w^T * \ \theta(y) + b = \sum_{s=1}^{S} a_s t_s \theta(x_s)^T * \theta(y) + b \tag{5}$$

were $x_s$ are the support vectors determined in the optimization process and $S$ is the number of support vectors.

## 4 Proposed Methods

In order to improve the results obtained with the similarity measure Intersection and Symmetric Difference (ISDS) [4] we propose to use SVM as a classifier of the accumulative vectors. We train a model for each target speaker using its accumulative vector and a set of background vectors.

We first obtained the generator model described in [4] and extracted the accumulative vectors of the target speakers from its corresponding signal. We took a set of background speakers and extract the corresponding accumulative vectors to be used as impostors in the SVM training process. These background speakers are labeled as -1 and are used to train all the target speakers' models.

Something that has negative impact in the use of accumulative vectors with SVM is their direct dependency with the number of frames of their corresponding signals. For that, before feeding accumulative vectors into SVM we transform them by the procedure described in 4.1.

To improve the results obtained with the SVM we use NAP as a technique to compensate the session variability presented in the accumulative vectors. Therefore we apply the NAP transformation to the accumulative vectors used for the SVM training before starting the training process. We assume that the accumulative vectors holds session variability information and we confirm that in the results. The NAP procedure is similar to the presented in [5] and is described in section 4.2. Finally we use a standard linear kernel to train the support vector machines.

### 4.1 Scaling the Accumulative Vectors

As we explain in section 2, a binary vector is obtained from each acoustic frame of a given signal. As result the numbers of binary vectors extracted from an acoustic signal depends on the length of it. Since the accumulative vectors are obtained from these binary vectors and their represent the number of times that each specificity was activated, the accumulative vectors of two acoustic signals from the same speaker will be very different if one signal is bigger than the other.

To deal with this problem, in [4] each accumulative vector is divided by the sum of the accumulative values in it. But we face a problem with this method: the resulting accumulative values are too small, and therefore, this phenomenon causes loss of significance in the data during the training of SVM.

To address this trouble we propose to divide each accumulative vector for the number of frames of its corresponding signal. As result, each accumulative value will be equal to the average that specificity was activated by frame. Then the new accumulative vectors are obtained by:

$$s_{acc} = \frac{s_{acc}}{N_f} \tag{6}$$

where $s_{acc}$ is the accumulative vector and $N_f$ is the number of frames of its corresponding signal. Then the new accumulative values are not too small, and with this method we outperform the proposal in [4], using SVM.

### 4.2 Nuisance Attribute Projection (NAP)

Nuisance Attribute Projection is a compensation technique that successfully removes the session variability in SVM supervectors [5, 9] and we use it with accumulative vectors. This algorithm is not specific to some kernel and can be applied to any kind of supervectors.

NAP makes the hypothesis that the channel variations tend to lie in a low-dimensional subspace of a speaker $s$ and projecting out these dimensions, most of the speaker-dependent information in $s$ will be unaffected. This transformation is achieved by:

$$s'_{acc} = s_{acc} - U(U^t s_{acc}) \tag{7}$$

where $U$ is the projection matrix and $s_{acc}$ is the accumulative vector of a given speaker. By orthonormality this transformation is idempotent [9]. This means that it is not necessary to transform the test accumulative vectors.

To obtain the U matrix we need a dataset with several speakers and several sessions for each one of them. With this dataset the procedure to obtain U is the following:

1. Extract an accumulative vector of dimension $D_{acc}$ for each session of the training set.
2. Scale these accumulative vectors using the method described in 4.1.
3. For each speaker, calculate the mean accumulative vector and then subtracts this mean from all of its accumulative vectors. Pooling all these accumulative vectors is obtained a large matrix D.
4. Now perform a Principal Component Analysis (PCA) on D to obtain the $D_{NAP}$ principal eigenvectors.
5. The result matrix is the projection matrix U.

In the matrix D most of the speaker variability presented in the accumulative vectors has been removed, however it holds the intersession variability.

## 5 Experiments

For all the signals used in the experiments we extract 19 Linear Frequency Cepstral Coefficients (LFCC) with the log energy. We add 20 delta coefficients and 10 delta-delta coefficients for a total of 50 features.

A UBM with 512 components was trained using 1661 speakers from NIST SRE 2005. Using this, we train the generator model as was described in [4] with 2450 multilingual signals of 124 speakers from NIST SRE 2004. This set of signals also was used to estimate the NAP projection matrix. To train the SVM we use a subset of 500 signals from the ones selected from NIST SRE 2004.

We use 0.001 as activation threshold to obtain the accumulative vectors. This means that a position in a binary vector was set to 1 if its corresponding likelihood is bigger than this value.

For the test we use det7 core condition test of NIST SRE 2008. This test has 1270 target speakers and 2528 unknown signals. We make 6615 verifications based on this test.

### 5.1 Results

Firstly we conduct a set of experiments without channel compensation to adjust the parameter C. The results are shown in Table 1.

**Table 1.** Equal Error Rate (EER) and Detection Cost Function (DCF) results for speaker verification using SVM without channel compensation to adjust the C parameter

| C | EER | DCF |
|---|---|---|
| 100 | 10.022% | 0.0491 |
| 250 | 8.428% | 0.0441 |
| 500 | 7.561% | 0.0416 |
| 750 | 7.742% | 0.0406 |
| 1000 | 7.742% | 0.0405 |
| 2500 | 7.289% | 0.0407 |
| 5000 | 7.289% | 0.0403 |
| 6000 | 7.253% | 0.0405 |
| **7500** | **7.061%** | 0.0399 |
| 8000 | 7.205% | 0.0398 |
| 9000 | 7.289% | 0.0401 |
| 10000 | 7.289% | **0.0395** |

For highest values of C we can see a stable behavior in EER. Although the better performance was obtained for C = 7500. We choose this value to adjust the dimension of the NAP matrix projection.

**Table 2.** Equal Error Rate (EER) and Detection Cost Function (DCF) results for speaker verification using SVM with channel compensation for different dimension of the NAP projection matrix and C=7500

| $D_{NAP}$ | EER | DCF |
|---|---|---|
| 40 | 6.735% | 0.0350 |
| 60 | 6.378% | 0.0355 |
| 100 | 6.525% | 0.0343 |
| 200 | 6.378% | 0.0345 |
| 350 | 6.169% | **0.0325** |
| 450 | 6.039% | 0.0329 |
| 550 | 6.150% | 0.0331 |
| **600** | **5.975%** | 0.0337 |
| 700 | 6.150% | 0.0362 |

In Table 2 we show that the best result of our system is obtained with the dimension of NAP projection matrix equal to 600 for C=7500. Therefore the use of NAP improves the results of the SVM by about of 1%. It proves that the accumulative vectors holds information related to session variability.

Finally for comparison purposes we select and develop two different experiments. First an experiment with the similarity measure ISDS [4] was conducted. We apply the normalization of the accumulative vectors described in [4] because ISDS is adjusted to work with this method.

At last a set of experiments using the state of art algorithm i-vector [10] with the compensation techniques Within Class Covariance Normalization (WCCN) [11] and Linear Discriminant Analysis (LDA) [10] was presented. The estimations of the matrixes associated with this experiment uses NIST 2004 speaker set previous described. The rank of the i-vector total variability matrix was equal to 400 and the LDA dimension was set in 390.

**Table 3.** Equal Error Rate (EER) and Detection Cost Function (DCF) comparison of our proposal with others

| $D_{NAP}$ | EER | DCF |
|---|---|---|
| SVM C=7500 $D_{NAP} = 600$ | 5.975% | 0.0337 |
| ISDS | 11.690% | 0.0486 |
| i-vector | 7.092% | 0.0309 |
| i-vector + LDA | 5.828% | 0.0290 |
| i-vector+WCCN | 6.655% | 0.0286 |
| i-vector+WCCN+LDA | 5.922% | 0.0283 |

Table 3 shows that our proposal outperforms the similarity measure ISDS and therefore the base line of the accumulative vectors. Also the results are very close to those obtained with the better techniques of the state of art applied to the GMM/UBM based supervector paradigm. Although the experiments only show a slight improvement on a single dataset, the proposed approach seems promising due to that the accumulative vectors paradigm is relatively new, just like its previously mentioned possibility of working directly with the discriminative information of a speaker.

A major drawback of the realized experiments is that we only have one sample of each target speaker and therefore we trained his corresponding SVM model with the accumulative vector extracted from that signal. The use of more than one sample should improve the results obtained. Also the selection of the background signals used to train the SVMs is very crucial. Nevertheless the low cost process of training and scoring the SVM models, its high discriminative power and the advantages relative to the accumulative vectors paradigm, compensate the mentioned disadvantages.

## 6 Conclusions and Future Work

In this paper we introduce a new scaling method of accumulative vectors because we obtain better results using SVM with it than with the reported in [4]. The obtained results prove that this method successfully removes the effects of signal length in the accumulative vectors. We also show that the accumulative vectors hold information about the session variability and it can be reduced by applying the NAP compensation technique. We demonstrate also that the results obtained with our proposal are much closed to those obtained with the state of art techniques but using a binary representation of a speaker that allows working directly with it discriminative characteristic and temporal information.

In the future we will try to use others compensation techniques instead on NAP to remove the session variability in the accumulative vectors and make a comparison with the results obtained, just like, use PLDA instead of LDA for comparison purpose. Also in [4] was proposed a trajectory model that represents the temporal information of a speaker and extract more than one accumulative vector, so we intend in futures work to exploit the information relative to this model by using SVM. Furthermore we intend to run more experiments in different datasets to enhance the robustness of our proposal.

## References

1. Campbell, W., et al.: Support vector machines for speaker and language recognition. Computer Speech and Language 20, 210–229 (2006)
2. Campbell, W., Sturim, D., Reynolds, D.: Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters 13, 308–311 (2006)
3. Anguera, X., Bonastre, J.F.: A novel speaker binary key derived from anchor models. In: Proc. Interspeech, pp. 2118–2121 (2010)
4. Hernández-Sierra, G., Bonastre, J.-F., Calvo de Lara, J.R.: Speaker recognition using a binary representation and specificities models. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 732–739. Springer, Heidelberg (2012)
5. Solomonoff, A., Campbell, W.M., Boardman, I.: Advances in Channel Compensation for SVM speaker Recognition. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 629–632 (2005)
6. Hautamaki, V., Kinnunen, T., Karkkainen, I., Tuononen, M., Saastamoinen, J., Franti, P.: Maximum a Posteriori estimation of the centroid model for speaker verification (2008)
7. Bonastre, J.F., Bousquet, P.M., Matrouf, D.: Discriminant binary data representation for speaker recognition. In: Proc. ICASSP, pp. 5284–5287 (2011)
8. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines (2000)
9. Campbell, W., Sturim, D., Reynolds, D.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 637–640 (2005)
10. Dehak, N., et al.: Front-End Factor Analysis For Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing 19(4), 788–798 (2010)
11. Hatch, A.O., Kajarekar, S., Stolcke, A.: Within-Class Covariance Normalization for SVM-based Speaker Recognition. In: Proc. ICSLP, pp. 1471–1474 (2006)

# Multimodal Biometric Fusion: A Study on Vulnerabilities to Indirect Attacks

Marta Gomez-Barrero, Javier Galbally,
Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group–ATVS, EPS, Universidad Autonoma de Madrid,
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{marta.barrero,javier.galbally,julian.fierrez,javier.ortega}@uam.es

**Abstract.** Fusion of several biometric traits has traditionally been regarded as more secure than unimodal recognition systems. However, recent research works have proven that this is not always the case. In the present article we analyse the performance and robustness of several fusion schemes to indirect attacks. Experiments are carried out on a multimodal system based on face and iris, a user-friendly trait combination, over the publicly available multimodal Biosecure DB. The tested system proves to have a high vulnerability to the attack regardless of the fusion rule considered. However, the experiments prove that not necessarily the best fusion rule in terms of performance is the most robust to the type of attack considered.

**Keywords:** Security, vulnerabilities, multimodality, iris recognition, face recognition, fusion schemes.

## 1 Introduction

Being able to automatically recognise people is of the utmost importance for many applications, such as regulating international border crossings or performing financial transactions on-line. Traditional security technologies required the use of PINs or tokens. Biometrics proposes a change of paradigm, from "what you know" or "what you have" to "who you are": forget about passwords, you are your own key [1].

However, as any other security technology, biometrics are exposed to external attacks which could compromise their integrity [2]. It is therefore essential to understand the threats to which they are subjected and to analyse their vulnerabilities in order to prevent possible attacks and increase their benefits for the users. External attacks to biometric systems are commonly divided into: *direct attacks* (also known as *spoofing attacks*), carried out against the sensor, and *indirect attacks*, directed to some of the inner modules of the system. In the last recent years important research efforts have been conducted to study the vulnerabilities of biometric systems to both direct and indirect attacks [3–5].

This new concern which has arisen in the biometric community regarding the security of biometric systems has led to the appearance of several international

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 358–365, 2013.

projects, like the European TABULA RASA [6] and BEAT [7], which base their research on the security through transparency principle: in order to make biometric systems more secure and reliable, their vulnerabilities need to be analysed and useful countermeasures need to be developed.

In this scenario, biometric multimodality has been regarded as an effective way of increasing the robustness of biometric-based security systems against external attacks. Combining the information offered by several traits would force an eventual intruder to successfully break several unimodal modules instead of just one. However, it has already been proven that this is not necessarily true for the case of spoofing attacks [8–10].

But are all fusion schemes equally robust to indirect attacks? If not, are the system performance and the robustness somehow correlated? In the present work we try to answer those questions using several score-level fusion schemes and a multimodal indirect attack already proven to be very successful in [11].

The paper is structured as follows: the attacking algorithm is summarized in Sect. 2. The system attacked, with the different fusion rules considered, is presented in Sect. 3, while the experimental protocol followed and the performance evaluation of the system are described in Sect. 4. The results obtained are shown in Sect. 5. Finally conclusions are drawn in Sect. 6.

## 2 Hll-Climbing Attack to Multimodal Recognition Systems

In order to attack the multimodal verification system using the different fusion schemes considered, the algorithm detailed in [11] will be used, which may be summarized as follows. Consider the problem of finding a $(K + L)$-dimensional vector $\mathbf{x}$ of real (size $K$) and binary (size $L$) values which, compared to an unknown template $\mathcal{C}$ (in our case related to a specific client), produces a similarity score higher than a certain threshold $\delta$, according to some matching function $J$, i.e., $J(\mathcal{C}, \mathbf{x}) > \delta$.

The problem stated above may be solved by dividing the vector $\mathbf{x}$ into its real-valued ($\mathbf{x}_{\text{real}}$) and binary parts ($\mathbf{x}_{\text{bin}}$) and alternately optimizing each of them. In order to optimize each of the parts, two different sub-algorithms will be used: *i*) a hill-climbing based on the uphill simplex to attack the real-valued segment; and *ii*) a hill-climbing attack based on a genetic algorithm to break the binary segment. Thus, the steps followed by the multimodal attack are:

1. Generate a synthetic template ($\mathbf{x}$) randomly initializing the real-valued ($\mathbf{x}_{\text{real}}$) and binary ($\mathbf{x}_{\text{bin}}$) segments, of lengths $K$ and $L$, respectively. Then compute the similarity score $s = J(\mathcal{C}, \mathbf{x})$, which will be iteratively maximised.
2. Leaving one of the segments unaltered, optimize the other segment of the template using the appropriate sub-algorithm until one of the stopping criteria of the sub-algorithm is fulfilled.
3. Change the optimization target to the segment which was previously left unaltered and go back to step 2.

The algorithm stops when: *i*) the verification threshold is reached (i.e., access to the system is granted), or *ii*) the total number of iterations exceeds a previously fixed value (i.e., the attack has failed).

It should be noted that the number of executions of each sub-algorithm is not fixed, and may vary depending on the user account at hand. That number can even be zero for one of the sub-algorithms, meaning that optimizing the other part of the template is enough to break the account.

For further details on the multimodal attack and on each of the two sub-algoritms, the reader is referred to [11].

**Notation.** Since the multimodal attack will be tested against a face- and iris-based multimodal system, we will henceforth denote the number of times the real-valued hill-climbing is executed as $N_{\text{face}}$, and the number of times that the binary-valued hill-climbing is executed as $N_{\text{iris}}$. Similarly, the real-valued segment of the template $\mathbf{x}$ will be denoted as $\mathbf{x}_{\text{face}}$, and the binary part as $\mathbf{x}_{\text{iris}}$.

## 3 Multimodal Verification System

The multimodal verification system evaluated in this work is the fusion of two unimodal systems, namely: *i*) the iris recognition system developed by L. Masek[1] [12], which is widely used in related publications; and *ii*) an Eigenface-based face verification system, used, for instance, to present initial face verification results for the Face Recognition Grand Challenge [13].

Given an input vector $\mathbf{x}$, the multimodal system performs the following tasks in order to obtain the final score, $s$:

1. Compute the similarity scores obtained by the face ($s_{\text{face}}$) and iris ($s_{\text{iris}}$) traits, as given by the unimodal matchers.
2. Normalize the scores $s_k$, with $k = \{\text{face}, \text{iris}\}$, using hyperbolic tangent estimators (its robustness and high efficiency are proven in [14]). This way, the normalised scores $s'_k$ lie in the interval $[0, 1]$.
3. Finally, both normalised scores are fused. Several fusion schemes have been considered [15, 16]:

$$\begin{array}{rr} \text{Sum rule} : s = s'_{\text{face}} + s'_{\text{iris}} & \text{Product rule} : s = s'_{\text{face}} \times s'_{\text{iris}} \\ \text{Max rule} : s = \max\{s'_{\text{face}}, s'_{\text{iris}}\} & \text{Min rule} : s = \min\{s'_{\text{face}}, s'_{\text{iris}}\} \end{array}$$

## 4 Database and Experimental Protocol

The experiments are carried out on the face and iris subcorpora included in the Desktop Dataset of the Multimodal Biosecure Database [17], which comprises voice, fingerprints, face, iris, signature and hand of 210 users, captured in two time-spaced acquisition sessions. This database was acquired thanks to the joint

[1] `www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html`

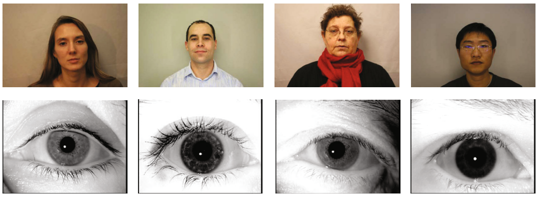


**Fig. 1.** Typical samples of the face and iris images available in the Desktop Dataset of the multimodal BioSecure database

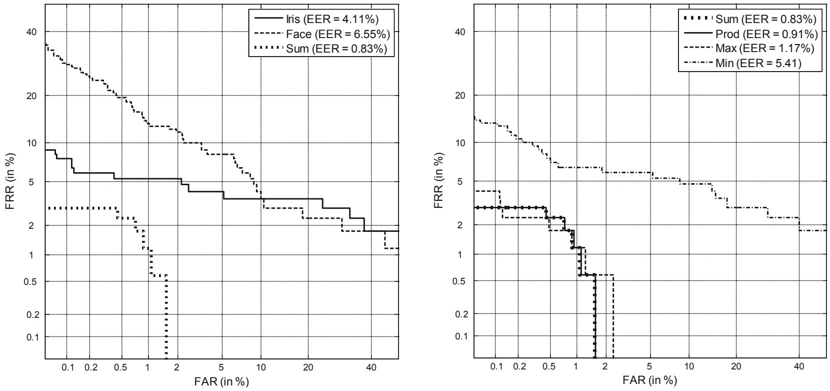


**Fig. 2.** DET curves for the unimodal systems and the fusion rule with the best performance (left) and for all the fusion rules considered (right), with their corresponding EER

effort of 11 European institutions and has become one of the standard benchmarks for biometric performance and security evaluations. It is publicly available through the BioSecure Association[2].

The face subset used in this work includes four frontal images (two per session) with an homogeneous grey background, and captured with a reflex digital camera without flash (210 × 4 = 840 face samples), while the iris subset includes four grey-scale images (two per session as well) per eye, all captured with the Iris Access EOU3000 sensor from LG. In the experiments only the right eye of each user has been considered, leading this way as in the face case to 210 × 4 = 840 iris samples. Typical samples may be seen in Fig. 1.

### 4.1 Performance Evaluation

The database is divided into: *i*) a training set comprising the first three samples of 170 clients, used as enrolment templates for each sub-system; and *ii*) an evaluation set formed by the fourth image of the previous 170 users (used to

[2] `http://biosecure.it-sudparis.eu/AB`

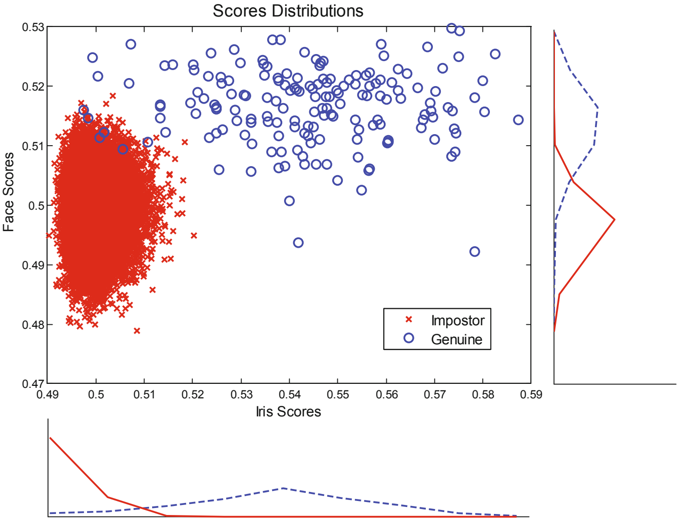


**Fig. 3.** Genuine and impostor distributions for the face ($y$ axis) and iris ($x$ axis) recognition systems

compute the genuine scores), and all the 4 images of the remaining 40 users (used to compute the impostor scores). The final score given by the multimodal system is the average of the scores obtained after matching the input template $\mathbf{x}$ to the three face and iris templates of the client model $\mathcal{C}$.

The attacking algorithm is evaluated at three operating points with FAR = 0.1%, FAR = 0.05%, and FAR = 0.01%, which correspond to a low, medium, and high security application according to [18].

As described in Sect. 3, several fusion rules are considered in the present study. The verification performance of the unimodal and multimodal combinations considered are shown in Fig. 2, where the Detection Error Tradeoff (DET) curves are depicted. As may be observed, the best performance is achieved for the sum rule (EER = 0.83%), while the worst one is shown for the min rule (EER = 5.41%).

In Fig. 3, the genuine and impostor distributions are shown.

### 4.2 Experimental Protocol for the Attacks

The performance of the attack will be evaluated in terms of: $i$) Success Rate (SR) which is the expected probability of breaking a given account, indicating how dangerous the attack is (the higher the SR, the bigger the threat); and $ii$) Efficiency (Eff) defined as the inverse of the average number of matchings needed to break an account, thus giving an estimation of how easy it is for the attack to break into the system in terms of speed (the higher the Eff, the faster the attack). The SR is computed as the ratio between the number of broken accounts ($A_B$) and the total number of accounts attacked ($A_T = 170$): SR = $A_B/A_T$, and the

**Table 1.** Eff and SR for the different fusion rules considered

| **FAR** | **Sum** | | **Prod** | | **Max** | | **Min** | |
|---|---|---|---|---|---|---|---|---|
| | SR | Eff ($\times 10^{-4}$) | SR | Eff ($\times 10^{-4}$) | SR | Eff ($\times 10^{-4}$) | SR | Eff ($\times 10^{-4}$) |
| 0.10% | 100% | 1.9372 | 100% | 1.9144 | 100% | 1.3231 | 100% | 2.3134 |
| 0.05% | 100% | 1.8218 | 100% | 1.7863 | 100% | 1.2060 | 100% | 2.0602 |
| 0.01% | 100% | 1.3702 | 100% | 1.3616 | 100% | 1.0220 | 100% | 1.7657 |

**Table 2.** Number of user accounts broken after attacking each part of the template a fixed number of times specified by $N_{\text{face}}$ and $N_{\text{iris}}$ (see Sect. 2)

| **FAR** | **Sum** ($N_{\text{face}} + N_{\text{iris}}$) | | | | | | **Prod** ($N_{\text{face}} + N_{\text{iris}}$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $1+0$ | $1+1$ | $2+1$ | $2+2$ | $3+2$ | $3+3$ | $1+0$ | $1+1$ | $2+1$ | $2+2$ | $3+2$ | $3+3$ |
| 0.10% | 0 | 153 | 9 | 7 | 0 | 1 | 0 | 161 | 5 | 4 | 0 | 0 |
| 0.05% | 0 | 155 | 8 | 7 | 0 | 0 | 0 | 158 | 6 | 6 | 0 | 0 |
| 0.01% | 0 | 117 | 27 | 21 | 2 | 3 | 0 | 118 | 27 | 19 | 3 | 3 |

| **FAR** | **Max** ($N_{\text{face}} + N_{\text{iris}}$) | | | | | | **Min** ($N_{\text{face}} + N_{\text{iris}}$) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $1+0$ | $1+1$ | $2+1$ | $2+2$ | $3+2$ | $3+3$ | $1+0$ | $1+1$ | $2+1$ | $2+2$ | $3+2$ | $3+3$ |
| 0.10% | 5 | 118 | 14 | 30 | 1 | 2 | 0 | 127 | 19 | 4 | 13 | 0 |
| 0.05% | 2 | 102 | 15 | 36 | 4 | 9 | 0 | 101 | 37 | 7 | 20 | 0 |
| 0.01% | 0 | 90 | 5 | 54 | 8 | 10 | 0 | 58 | 53 | 8 | 38 | 0 |

Eff is computed as $\text{Eff} = 1/\left(\sum_{i=1}^{A_B} n_i/A_B\right)$, where $n_i$ is the number of matchings needed to bypass each of the broken accounts.

## 5 Results

The experiments have two different goals, namely: *i*) analyse the robustness against indirect attacks of different fusion rules, and *ii*) study to what extent the vulnerabilities of a multimodal recognition system based on face and iris are correlated to the verification performance.

### 5.1 Vulnerabilities Evaluation

In Table 1, the performance of the attack in terms of the SR and the Eff is shown. As may be observed, the SR is 100% in all cases: all accounts are broken, regardless of the fusion scheme considered. However, not all the fusion schemes are equally robust in terms of speed: the Eff for the min rule is the highest one, being therefore the least robust fusion scheme. On the other hand, while the Eff for the sum and product rules is very similar, for the max rule it is considerably lower. Therefore, for applications where the robustness to this kind of attacks is more important than having an optimal performance (EER rises from 0.83% with the sum rule, to 1.17% with the max rule), the max rule should be considered.

For all the user accounts attacked, each sub-algorithm was executed between 0 and 3 times. Therefore, there are six possible cases regarding the number of those executions ($N_{\text{face}}+N_{\text{iris}}$). In the particular case when $N_{\text{iris}} = 0$, the account was broken after the first execution of the real-valued hill-climbing, therefore not needing to attack the binary part. The number of accounts that fall into each category is shown in Table 2. As may be observed, most accounts are broken after optimizing each part of the template only once.

In Sect. 4.2, Eff was defined as the inverse of the average number of comparisons needed to break an account. Therefore, the lower the Eff, the higher the number of comparisons needed. As could be expected, the lower the FAR at the operating point tested, the higher the number of users for which more executions of each sub-algorithm were needed.

However, when we compare the results shown in Table 2, we observe two different behaviours:

- For the sum, product and max rules, as expected, the lower the Eff, the higher the number of users for which two or even three executions of each sub-algorithm were needed.
- For the min rule, which presented the highest Eff for the attack (see Table 1), the number of users requiring three executions of the real-valued sub-algorithm is the highest. This means that the genetic sub-algorithm saturates quickly, and therefore the general attacking scheme starts attacking the face part of the template: as stated in [11], the genetic sub-algorithm needs considerably more comparisons than the hill-climbing based on the uphill simplex, leading this quick change to a higher Eff.

## 6 Conclusions

In the present article we have analysed the robustness of different multimodal score-level fusion rules (sum, product, max and min) to indirect attacks. We have then explored to what extent there is a correlation between the vulnerabilitiy level and the performance of the multimodal system. A multimodal system based on face and iris, a trait combination commonly regarded as user-friendly, working on a publicly available multimodal database, was used in the experiments.

The experiments showed that the multimodal attack achieves a Success Rate of 100% in all cases, regardless of the operating point or the fusion rule considered. However, the Efficiency of the algorithm varies, and from that variation some criteria for choosing a fusion rule for the multimodal system were inferred.

Even though the results presented here are based on simple fusion rules, the experimental framework can be easily extended to more complex architectures. Future work considering other biometric modalities and fusion schemes will be carried out in order to reach a deeper understanding of the behaviour of multimodal biometric systems under indirect attacks.

Works such as the one presented here emphasize the importance of developing appropriate template protection countermeasures that minimize the effects of the studied attacks. Some countermeasures have been proposed to counterfeit spoofing attacks, such as [19]. However, the application of those measures against indirect attacks is not straightforward, since they work on raw biometric traits instead of preprocessed templates.

**Acknowledgements.** This work has been partially supported by projects Contexts (S2009/TIC-1485) from CAM, Bio-Challenge (TEC2009-11186) and Bio-Shield (TEC2012-34881) from Spanish MINECO, TABULA RASA (FP7-ICT-257289) and BEAT (FP7-SEC-284989) from EU, and *Cátedra UAM-Telefónica*. Marta Gomez-Barrero is supported by a FPU Fellowship from Spanish MECD.

## References

1. Jain, A.K., et al.: Biometrics: a tool for information security. IEEE TIFS 1(2), 125–143 (2006)
2. Schneier, B.: Inside risks: the uses and abuses of biometrics. Commun. ACM 42, 136 (1999)
3. Galbally, J., et al.: Evaluation of direct attacks to fingerprint verification systems. Telecommunication Systems 47, 243–254 (2011)
4. Galbally, J., et al.: On the vulnerability of face verification systems to hill-climbing attacks. PR 43, 1027–1038 (2010)
5. Akhtar, Z., et al.: Robustness analysis of likelihood ratio score fusion rule for multimodal biometric systems under spoof attacks. In: Proc. ICCST, pp. 1–8 (2011)
6. TABULA RASA: Trusted biometrics under spoofing attacks (2013)
7. BEAT: Biometrics evaluation and testing (2013)
8. Rodrigues, R., et al.: Evaluation of biometric spoofing in a multimodal system. In: Proc. IEEE BTAS (September 2010)
9. Johnson, P.A., et al.: Multimodal fusion vulnerability to non-zero effort (spoof) imposters. In: Proc. WIFS (2010)
10. Akhtar, Z., et al.: Spoof attacks in mutimodal biometric systems. In: Proc. IPCSIT, vol. 4, pp. 46–51. IACSIT Press (2011)
11. Gomez-Barrero, M., et al.: Efficient software attack to multimodal biometric systems and its application to face and iris fusion. PRL (2013), doi:10.1016/j.patrec.2013.04.029
12. Masek, L., Kovesi, P.: Matlab source code for a biometric identification system based on iris patterns. Master's thesis, University of Western Australia (2003)
13. Phillips, J., et al.: Overview of the face recognition grand challenge. In: Proc. IEEE CVPR, pp. 947–954 (2005)
14. Jain, A.K., et al.: Score normalization in multimodal biometric systems. PR 38, 2270–2285 (2005)
15. Kittler, J., et al.: On combining classifiers. IEEE TPAMI 20(3), 226–239 (1998)
16. Fierrez, J.: Adapted Fusion Schemes for Multimodal Biometric Authentication. PhD thesis, Universidad Politecnica de Madrid (2006)
17. Ortega-Garcia, J., et al.: The multi-scenario multi-environment BioSecure multimodal database (BMDB). IEEE TPAMI 32, 1097–1111 (2010)
18. ANSI-X9.84-2001: Biometric information management and security (2001)
19. Marfella, L., et al.: Liveness-based fusion approaches in multibiometrics. In: Proc. IEEE BIOMS (2012)

# Gait-Based Gender Classification Using Persistent Homology

Javier Lamar Leon[1], Andrea Cerri[2,⋆], Edel Garcia Reyes[1], and Rocio Gonzalez Diaz[3]

[1] Patterns Recognition Department, Advanced Technologies Application Center,
7a ♯ 21812 e/ 218 y 222, Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba
{jlamar,egarcia}@cenatav.co.cu

[2] IMATI - CNR, : via de Marini 6, 16149 Genova, Italy
andrea.cerri@ge.imati.cnr.it

[3] Applied Math Dept., School of Computer Engineering,
Campus Reina Mercedes, University of Seville, Seville, Spain
rogodi@us.es

**Abstract.** In this paper, a topological approach for gait-based gender recognition is presented. First, a stack of human silhouettes, extracted by background subtraction and thresholding, were glued through their gravity centers, forming a 3D digital image $I$. Second, different *filters* (i.e. particular orders of the simplices) are applied on $\partial K(I)$ (a simplicial complex obtained from $I$) which capture relations among the parts of the human body when walking. Finally, a *topological signature* is extracted from the persistence diagram according to each filter. The measure cosine is used to give a similarity value between topological signatures. The novelty of the paper is a notion of robustness of the provided method (which is also valid for gait recognition). Three experiments are performed using all human-camera view angles provided in CASIA-B database [1]. The first one evaluates the named topological signature obtaining 98.3% (lateral view) of correct classification rates, for gender identification. The second one shows results for different human-camera distances according to training and test (i.e. training with a human-camera distance and test with a different one). The third one shows that upper body is more discriminative than lower body.

**Keywords:** gait-based recognition, topology, persistent homology, gender classification.

## 1 Introduction

Gender human classification can be obtained based on face [1], voice [2] or gait [3, 4]. Dynamic features when the people walk give the possibility to identify

[⋆] Supported by the Italian CNR Flagship project INTEROMICS, InterOmics PB05, research unit WP15.

[1] http://www.cbsr.ia.ac.cn/english/GaitDatabases.asp1

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 366–373, 2013.

persons and their gender at a distance, without any interaction from the subjects [5–7]. This fact can improve the performance of surveillance system intelligent, the analysis of customer information in trade centers, and it can reduce the false positive rate during reidentification of an individual on a wide network cameras. People not only observe the global motion properties while human walk, but they detect motion patterns of local body parts. For instance, females tent to swing their hips more than their shoulders. On the contrary, males tent to swing their shoulders more than their hips [8]. Moreover, males have in general wider shoulders than females [9]. An experiment for human observers to analyze the contributions of different parts of the human body (lower body, upper body and whole body) and to study their discriminative power appears in [3]. According to that experiment, upper body contributes more than lower body to gender classification. In fact, 94.35% and 67.86% of correct classification rates, for upper and lower body, respectively, were obtained. In this paper, a modified version of the topological gait signature given in [10, 11] is presented, which is valid for gait and also for gender classification. Besides, an important contribution of the paper are arguments for the robustness of the signature with respect to small input-data perturbations (i.e., perturbations on the stack of silhouettes) are presented. We test this topological signature on the CASIA-B database and compare our method with existing ones for gender recognition.

The rest of the paper is organized as follows. Section 2 is devoted to describe the method for obtaining the topological signature and arguments for its robustness. Experimental results are then reported in Section 3. We conclude this paper and discuss some future work in Section 4.

## 2 Topological Signature for Gender Classification

In this section, we briefly explain how the topological signature for gait and gender classification is obtained. As we will see below, the filters (ordering of simplices) are given by using functions defined on the simplicial complex $\partial K(I)$ and associated to the given view directions. These functions will be used later for sketching robustness of the topological signature for gait and gender recognition with respect to "small" input-data perturbations. Persistent homology obtained from these filters are represented here in *persistence diagram* format [12].

### 2.1 The Simplicial Complex $\partial K(I)$

First, the foreground (person) is segmented from the background by applying background modeling and subtraction. The sequence of resulting silhouettes is analyzed to extract one *subsequence of representation*, which include at least a gait cycle [5].

The 3D binary digital picture $I = (\mathbb{Z}^3, B)$ (where $B \subset \mathbb{Z}^3$ is the foreground), is built by stacking silhouettes of a subsequence of representation, aligned by their gravity centers ($gc$). See Fig. 1.a and Fig. 1.b. The 3D cubical complex $Q(I)$ associated to $I$ is constructed as follow: Visit all the points $v = (i, j, k) \in B$ from down to up and from left to right.

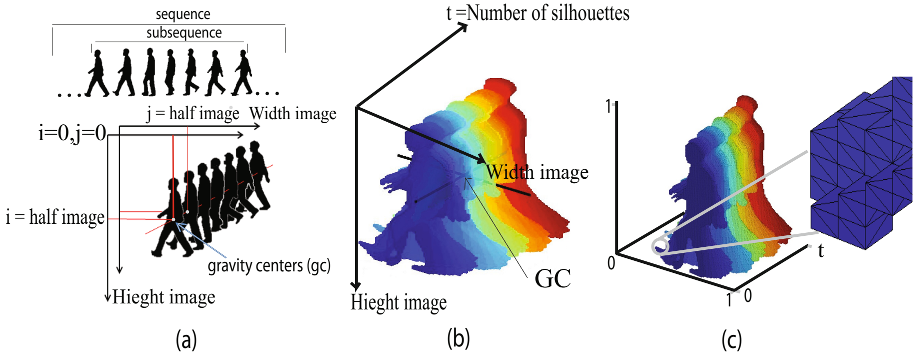


**Fig. 1.** (a) Silhouettes aligned by their *gc*. (b) $I = (\mathbb{Z}^3, B)$ obtained from the silhouettes ($GC$ is the gravity center of $I$). (c) The border simplicial complex $\partial K(I)$.

If the 7 neighbors of $v$, $\{(i+1,j,k),(i,j+1,k),(i,j,k+1),(i+1,j+1,k),(i+1,j,k+1),(i,j+1,k+1),(i+1,j+1,k+1)\}$, are also in $B$ then, the unit cube formed by these 8 vertices together with all its faces (vertices, edges and squares) are added to $Q(I)$. The simplicial complex $\partial K(I)$ is constructed by selecting all the squares of $Q(I)$ that are faces of exactly one cube in $Q(I)$ and subdividing such squares in two triangles. The faces of each triangle (vertices and edges) are also added to $\partial K(I)$ (see Fig. 1.c). Finally, coordinates of the vertices of $\partial K(I)$ are *normalized* to coordinates $(x, y, t)$, where $0 \leq x, y \leq 1$ and $t$ is the number of silhouette of the subsequence of representation.

### 2.2 Filters for $\partial K(I)$

The topology of $\partial K(I)$ is, in general, very poor. However, in this subsection we present how, using persistence diagrams, it is possible to get a topological signature from $\partial K(I)$ that captures relations among the parts of the human body when walking, and is robust against small input-data perturbations.

When a view direction $d$ is chosen, two filters for $\partial K(I)$ are obtained as follows. All vertices belonging to $\partial K(I)$ are associated with two *filtering functions* $f_+$ and $f_-$. For each vertex $v \in \partial K(I)$, $f_+(v)$ is the distance between $v$ and the plane normal to $d$ and passing through the origin of the reference frame, while $f_-(v) = -f_+(v)$. Edges and triangles are associated to the smallest value that $f_+$ (resp. $f_-$) assumes on their vertices. Being the simplices of $\partial K(I)$ finite in number, we can determine a minimum value for $f_+$, say $f_{\min}$, and a maximum one, $f_{\max}$. It is now possible to induce two filters on $\partial K(I)$ by ordering its simplices according to increasing values of $f_+$ and $f_-$, respectively. Denote these filters by $K_{[f_{\min}, f_{\max}]} = \{\sigma_1, \ldots, \sigma_k\}$ and $K_{[-f_{\max}, -f_{\min}]} = \{\sigma'_1, \ldots, \sigma'_k\}$.

### 2.3 Persistence Diagrams and Topological Signatures

Given a simplicial complex $K$, a filtering function $f$, and a filter $\{\sigma_1, \ldots, \sigma_k\}$ for $K$, if $\sigma_i$ completes a $p-$cycle ($p$ is the dimension of $\sigma_i$) when $\sigma_i$ is added to

$K_{i-1} = \{\sigma_1, \ldots, \sigma_{i-1}\}$, then a $p-$homology class $\gamma$ *is born at time* $i$; otherwise, a $(p-1)-$homology class *dies at time* $i$. The difference between the birth and death time of a homology class is called its *index persistence*, which quantifies the significance of a topological attribute. If $\gamma$ never dies, we set its persistence as well as its index persistence to infinity. Drawing a point $(i, j)$ for a homology class that is born at time $i$ and dies at time $j$, we get the $p-$persistence diagram of the filtration, denoted as $Dgm(f)$. It represents a $p-$homology class by a point whose vertical distance to the diagonal is the persistence. Since always $i < j$, all points lie above the diagonal (see [12]).

In this paper, persistence diagrams are first computed for $K_{[f_{\min}, f_{\max}]}$ and $K_{[-f_{\max}, -f_{\min}]}$. Then, the diagrams are explored according to a uniform sampling. More precisely, given a positive integer $n$, compute the integer $h = \lfloor \frac{k}{n} \rfloor$ representing the width of the "window" we use to analyze the persistence diagram. Indeed, for $i = 1, \ldots, n$, *the* $i-$*reduced persistence diagram of* $K_{[f_{\min}, f_{\max}]}$ (resp. $K_{[-f_{\max}, -f_{\min}]}$) show

(a) Homology classes that are born after $(i-1) \cdot h$ and before $i \cdot h$. Let $\ell$ be the time when such homology class is born. Its *reduced life-length* is $i \cdot h - \ell$.

Having the reduced persistence diagrams on hand, we can now compute two $n-$dimensional vectors for $K_{[f_{\min}, f_{\max}]}$ (resp. for $K_{[-f_{\max}, -f_{\min}]}$) whose $i-$entry corresponds to:

1. the sum of the reduced life-lengths for the $0-$homology classes $sumH_0$
2. the sum of the reduced life-lengths for the $1-$homology classes $sumH_1$.

These two collections of two $n-$dimensional vectors, represent *the topological signature for a gait subsequence associated with a fixed view direction.*

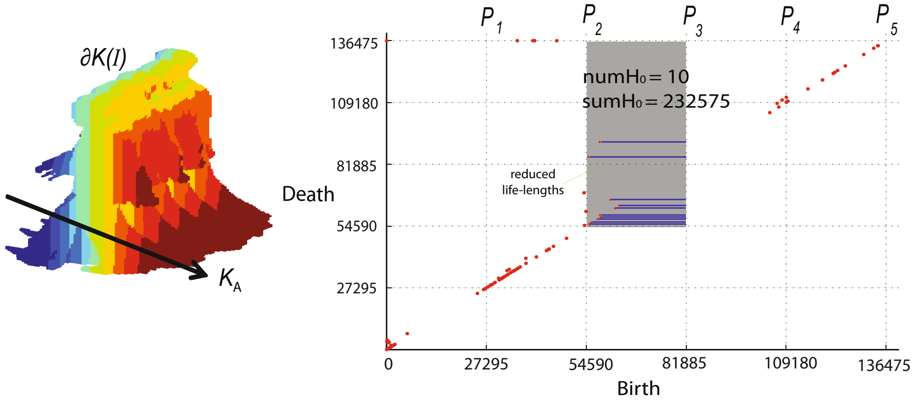


**Fig. 2.** An example of computation of the first element of a topological signature

For example, consider $K_A$ given in Fig. 2 which consists in 136475 triangles. We perform $n = 5$ uniform cuts on the $0-$persistence diagram. The sum of the reduced life-lengths for the 0−homology classes ($numH_0 = 10$) that were born in time $54560 \leq t < 81885$ are $sumH_0 = 232575$ (blue lines in Fig. 2). The first element of the topological signature: $V_1$ is, $\{473625, 813786, 232575, 10039, 203958\}$.

### 2.4 Comparing Topological Signatures

The topological signatures for two gait subsequences associated with a fixed view direction, say $V = \{V_1, \ldots, V_4\}$ and $W = \{W_1, \ldots, W_4\}$, can be compared according to the following procedure: for every $i = \{1, \ldots, 4\}$ compute:

$$S_i = \frac{V_i \cdot W_i}{\|V_i\| \cdot \|W_i\|}. \tag{1}$$

which is the cosine of the angle between the vectors $V_i$ and $W_i$. Observe that $0 \leq S_i \leq 1$ since the entries of both vectors are always non-negative. Then, the total similarity value for two gait subsequences, $O_1$ and $O_2$, considering a fixed view direction, is the sum of the 4 similarity measures computed before:

$$S(O_1, O_2) = S_1 + S_2 + S_3 + S_4. \tag{2}$$

### 2.5 Robustness

In this subsection, we briefly sketch a notion of robustness for our topological signature with respect to small input-data perturbations. Fix a view direction $d$ and an associated filtering function $f$. The assumption here is that the input-data perturbations can be modeled as perturbations of the function $f$. We could think, for example, of small perturbations in fixing $d$, as well as noise in the computation of $f$. More precisely, consider two functions $f, g : \partial K(I) \to \mathbb{R}$ such that

$$\max_{\sigma \in \partial K(I)} |f(\sigma) - g(\sigma)| \leq \varepsilon,$$

with $\varepsilon$ being a small positive real number. Let $K_f = \{\sigma_1, \ldots, \sigma_k\}$ and $K_g = \{\sigma'_1, \ldots, \sigma'_k\}$ be the filters associated with the increasing values of $f$ and $g$, respectively[2]. Assume also that all homology classes fulfill either condition $(a)$: Such an assumption is actually mild, and can be achieved quite easily in practice (e.g., by slightly perturbing the values of $f$). Then, the stability of persistence diagrams [13] implies that the birth- and the death-times, with respect to $g$, of each homology class, cannot differ more than $\varepsilon$ from those with respect to $f$. Therefore, if $\varepsilon$ is sufficiently small, it follows that

- If a homology class fulfills condition $(a)$ for $K_f$, the same occurs for $K_g$.

Moreover, the same stability result in [13] implies that new homology classes, living no longer than $2\varepsilon$, could appear, as well as old classes living shorter than $2\varepsilon$ may vanish. These events could sensibly change the number of homology classes satisfying condition $(a)$. Nevertheless, considering such classes according to their reduced life-length, as specified above, guarantees the robustness of our topological signatures.

[2] Similar arguments hold if considering filters associated with the decreasing values of $f$ and $g$.

## 3 Experimental Results

In this section, we show the performance of the proposed method on gait sequences from the CASIA-B database, which contains 124 subjects, 91 males and 31 females. There are 6 walking sequences for each person. CASIA-B database provides image sequences with background subtraction for each person.

To avoid bias, 31 males were randomly selected. The 62 subjects were then divided in 31 disjoint sets, each containing 2 subjects (a male and a female). Only one of these 31 sets was sued to test. The remaining 30 sets were used for training. The correct classification rate (CCR) is the average of the 31 possible combinations.

The experimental protocol was made according to [3, 4]. In this experiment, a subsequence of representation corresponds to the whole sequence, which has two gait cycle as average. We have fixed $n = 24$ and used 3 view directions. The first one is vertical (i.e. parallel to axis $y$). The second one forms 45 degrees with axes $x$ and $y$ and 90 degrees with axis $t$. The third one is parallel to axis $t$. See Fig. 3. In each experiment, the results of our method are compared with the methods presented in [3, 4].

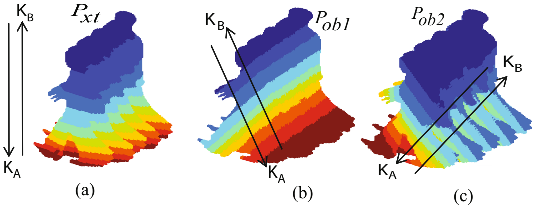


**Fig. 3.** View directions used in the experiments

### 3.1 Experiment 1

The aim of this experiment is to evaluate the topological signature for gender classification. Table 1 shows the 31−fold-cross-validation of CCR for the whole body using the 11 view directions provided by the CASIA-B database. The first line of the table refers to the camera view angle. This way, 0 degrees means that the person is in front to the camera and walking to the camera, 90 degrees means that the person is walking lateral to the camera (*lateral view*), and 180 degrees means that the person is back to the camera and walking away the camera. We can see that the topological signature provides better results for the lateral view. This agrees with [7].

### 3.2 Experiment 2

In this experiment we show that our topological signature is robust with respect to scaling. Images form CASIA-B database of size 320 × 240 are scaled to

**Table 1.** Correct classification rates (CCR in %) for the whole body

| Method | 0 | 18 | 36 | 54 | 72 | 90 | 108 | 126 | 144 | 162 | 180 | **Avg** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Avg computer [3] | | | | | | 95.97 | | | | | | |
| Avg human observers [3] | | | | | | 95.45 | | | | | | |
| Gabor + MMI [14] | | | | | | 96.8 | | | | | | |
| MCRF [4] | | | | | | 98.3 | | | | | | |
| Our method | 83.6 | 92.3 | 92.6 | 93.0 | 95.6 | 98.3 | 94.3 | 94.0 | 92.4 | 92.5 | 94.1 | **93.0** |

160 × 120. Table 2 shows results considering different scales for training and test. For example, if images of size 320 × 240 are used for training and images of size 160 × 120 are using for test, then we obtain 98.0% of CCR. Nevertheless, if images of size 160 × 120 are used for training and images of size 320 × 240 are using for test, then we obtain 95.6% of CCR.

**Table 2.** Correct classification rates (CCR in %) using different sizes of the images for training and test

| | | **Test** | |
|---|---|---|---|
| | | 320 × 240 *images* | 160 × 120 *images* |
| **Training** | 320 × 240 *images* | 98.3 | 98.00 |
| | 160 × 120 *images* | 95.6 | 97.5 |

### 3.3 Experiment 3

The aim of this experiment is to compare gender classifications using only upper or lower body. According to Table 3, our results confirm that upper body contributes more than lower body to gender classification for both original ($320x240$) and scaled ($160x120$) images. This agrees with the results obtained by human observers in [3].

**Table 3.** Correct classification rates (CCR in %) for lower and upper body from lateral view for original (320 × 240) and scaled (160 × 120) images

| Method | CCR (lateral view) |
|---|---|
| Human observers (lower body) (320 × 240 images) [3] | 67.86 |
| Our method (lower body)(320 × 240 images) | 88.1 |
| Our method (lower body) (160 × 120 images) | 87.0 |
| | |
| Human observers (upper body) (320 × 240 images) [3] | 94.35 |
| Our method (upper body)(320 × 240 images) | 96.0 |
| Our method (upper body) (160 × 120 images) | 95.5 |

## 4 Conclusion and Future Work

In this paper, a representation based on topological invariants, previously used for gait based human identification at a distance, is used for a gender classification task. Arguments for the robustness of the method with respect to small input-data perturbations are given. It should be noticed that the view direction should be selected according to the camera view angle to improve the results. The method has been implemented in C++ and has been tested in real-time real-life scenery in [11]. Our future work consists in trying to improve our results for camera view angles different to lateral view selecting the appropriate view direction, and to adapt our method to occlusions.

## References

1. Golomb, B.A., Lawrence, D.T., Sejnowksi, T.J.: SEXNET: A neural network identifies sex from human faces. In: Lippmann, R.P., Moody, J.E., Touretzky, D.S. (eds.) Advances in Neural Information Processing Systems, vol. 3, pp. 572–579. Morgan Kaufmann Publishers, Inc. (1991)
2. Harb, H., Chen, L.: Gender identification using a general audio classifier. In: Proceedings of the 2003 International Conference on Multimedia and Expo, ICME 2003, vol. 1, pp. 733–736. IEEE (2003)
3. Yu, S., Tan, T., Huang, K., Jia, K., Wu, X.: A study on gait-based gender classification. IEEE Trans. Image Processing 18(8), 1905–1910 (2009)
4. Hu, M., Wang, Y., Zhang, Z., Zhang, D.: Gait-based gender classification using mixed conditional random field. IEEE Transactions on Systems, Man, and Cybernetics, Part B 41(5), 1429–1439 (2011)
5. Nixon, M.S., Carter, J.N.: Automatic recognition by gait. Proc. of IEEE 94(11), 2013–2024 (2006)
6. Kale, A., Sundaresan, A., Rajagopalan, A.N., Cuntoor, N.P., Chowdhury, A.K.R., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Trans. Image Processing 13(9), 1163–1173 (2004)
7. Goffredo, M., Carter, J., Nixon, M.: Front-view gait recognition. In: Biometrics: Theory, Applications and Systems, September 29-October 1, pp. 1–6 (2008)
8. Aggarwal, J.K., Cai, Q.: Human motion analysis: A review. Computer Vision and Image Understanding 73(3), 428–440 (1999)
9. Mather, G., Murdoch, L.: Gender discrimination in biological motion displays based on dynamic cues. Proceedings of Biological Sciences 258(1353) (1994)
10. Lamar-León, J., García-Reyes, E.B., Gonzalez-Diaz, R.: Human gait identification using persistent homology. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 244–251. Springer, Heidelberg (2012)
11. Lamar, J., Garcia, E., Gonzalez-Diaz, R., Alonso, R.: An application for gait recognition using persistent homology. Electronic Journal Image-A 3(5) (2013)
12. Edelsbrunner, H., Harer, J.: Computational Topology - an Introduction. American Mathematical Society (2010)
13. Cohen-Steiner, D., Edelsbrunner, H., Harer, J.: Stability of persistence diagrams. Discrete & Computational Geometry 37(1), 103–120 (2007)
14. Hu, M., Wang, Y., Zhang, Z., Wang, Y.: Combining spatial and temporal information for gait based gender classification. In: 20th International Conference on Pattern Recognition (ICPR), pp. 3679–3682. IEEE (2010)

# Iris-Biometric Fuzzy Commitment Schemes under Image Compression

Christian Rathgeb[1], Andreas Uhl[2], and Peter Wild[2]

[1] Hochschule Darmstadt - CASED, Darmstadt, Germany
christian.rathgeb@h-da.de
[2] Dept. of Computer Sciences, University of Salzburg, Austria
{uhl,pwild}@cosy.sbg.ac.at

**Abstract.** With the introduction of template protection techniques, privacy and security of biometric data have been enforced. Meeting the required properties of irreversibility, i.e. avoiding a reconstruction of original biometric features, and unlinkability among each other, template protection can enhance security of existing biometric systems in case tokens are stolen. However, with increasing resolution and number of enrolled users in biometric systems, means to compress biometric signals become an imminent need and practice, raising questions about the impact of image compression on recognition accuracy of template protection schemes, which are particularly sensitive to any sort of signal degradation. This paper addresses the important topic of iris-biometric fuzzy commitment schemes' robustness with respect to compression noise. Experiments using a fuzzy commitment scheme indicate, that medium compression does not drastically effect key retrieval performance.

## 1 Introduction

*Biometric cryptosystems* and *cancelable biometrics* are classes of template protection schemes designed to maintain recognition accuracy [10] while protecting biometric information as standardized in ISO/IEC 24745 in case standard encryption (using AES, etc.) is not an option (e.g., there is no secure hardware environment). Their two critical properties are referred to as irreversibility (original biometric templates can not be retrieved in any way from stored reference data) and unlinkability (different versions of protected templates can not be cross-matched against each other), making them - generally - highly sensitive towards changes in environmental recording conditions and signal degradation which may be caused by compression algorithms [3].

The contribution of this work is the investigation of the impact of image compression on the performance of iris fuzzy commitment schemes (FCSs) [11], biometric cryptosystems which represent instances of biometric key-binding. We employ a representative selection of lossy image compression standards for biometric data compression (JPEG, JPEG XR and JPEG 2000), i.e. images are compressed after sensing and before normalization reflecting, e.g. remote-processing with mobile data acquisition on low-powered devices. Fig. 1 illustrates

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 374–381, 2013.

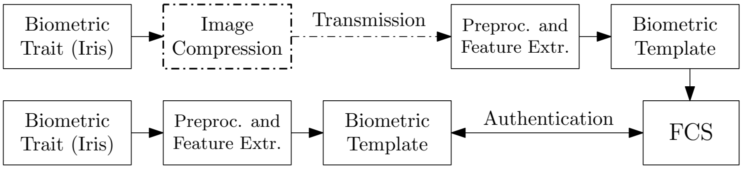


**Fig. 1.** Supposed scenario: compressed images are transmitted and applied in a template protection system based on the FCS

the processing chain. Experimental studies are carried out on an iris-biometric database employing different feature extractors to construct FCSs. It is found that the incorporation of image compression standards to FCSs reveal key retrieval rates, comparable to the performance of original recognition algorithms even at high compression levels. This paper is organized as follows: In Sect. 2 related works regarding FCSs and compression of biometric data are reviewed. Subsequently, a comprehensive evaluation on the effect of image compression standards on an iris-biometric FCS is presented in Sect. 3. Finally, a conclusion is drawn in Sect. 4.

## 2 Fuzzy Commitment Schemes

A FCS is a bit commitment scheme resilient to noise and proposed in [11]. Given a witness $x \in \{0,1\}^n$ representing a binary biometric feature vector and a set $C$ of error correcting codewords of length $n$, a FCS can be modeled as a function $F$, applied to commit $x$ with a codeword $c \in C$. Instead of storing the original feature vector, $x$ is concealed using a hash function $h(x)$. In order to reconstruct $x$, an offset $\delta \in \{0,1\}^n, \delta = x - c$ is calculated: $F(c,x) = \big(h(x), x - c\big)$. Since biometric signals $x$ are rarely reproduced exactly in different sensing operations, it is demanded, that any $x'$ sufficiently "close" to $x$ according to an appropriate metric (e.g. Hamming distance), should be able to reconstruct $c$ using the difference vector $\delta$. If for small fixed threshold $t$ (lower bounded by the according error correction capacity) the inequality $\|x - x'\| \leq t$ holds, $x'$ yields a successful de-commitment of $F(c,x)$ for any $c$. In order to accomplish this task, Hadamard codes (for elimination of bit errors originating from the natural biometric variance) and Reed-Solomon codes (correct burst errors resulting from distortions) can be applied [8]. Otherwise $c$ can not be reconstructed ($h(c) \neq h(c')$) yielding a key error.

FCSs have been applied to several different biometric modalities. Hao *et al.* [8] applied FCS to iris biometrics using relatively long (140-bit) keys with Hadamard and Reed-Solomon error correction codes. Bringer *et al.* introduce 2D iterative min-sum decoding for error correction decoding in an iris-based FCS, which gets close to a theoretical bound. Rathgeb and Uhl [18] present a technique to rearrange iris-codes in a way that FCS error correction capacities are exploited more effectively. Zhang *et al.* [23] propose a bit masking and code concatenation scheme to improve the accuracy of iris-based FCSs. In [19] a feature level fusion technique for increasing efficiency in a FCS is presented. Nandakumar

**Table 1.** Experimental results of FCSs proposed in literature

| Ref. | Modality | FRR/ FAR | Key Bits | Remarks |
|---|---|---|---|---|
| [8] | Iris | 0.47/ 0 | 140 | ideal images |
| [2] | | 5.62/ 0 | 42 | short key |
| [18] | | 4.64/ 0 | 128 | – |
| [19] | | 5.56 / $\leq$0.01 | 128 | fusion |
| [21] | Fingerprint | 0.9/ 0 | 296 | user-specific tokens |
| [16] | | 12.6/ 0 | 327 | – |
| [22] | Face | 3.5/ 0.11 | 58 | >1 enroll. sam. |
| [1] | | 7.99/ 0.11 | >4000 | user-specific tokens |
| [15] | Online Sig. | EER >9 | >100 | >1 enroll. sam. |

*et al.* [16] quantize the Fourier phase spectrum of a minutia set to derive a binary fixed-length representation for a FCS. Teoh *et al.* [21] apply a non-invertible projection based on a user-specific token randomized for a FCS based on dynamic quantization transformation from a multichannel Gabor filter and Reed-Solomon codes, similar to the approach in Ao and Li [1] based on face biometrics. Another face-based FCS is introduced in [22] based on bit selection to detect most discriminative features from binarized real-valued face features. Maiorana and Campisi [15] introduce a FCS for on-line signatures. Table 1 lists a summary of FCSs approaches.

It is important to note, that both, standardization and a variety of independent studies deal with compression. Current ISO/ IEC 19794 ("Biometric Data Interchange Formats") on standardized image compression in biometrics (fingerprint, face, and iris image data are covered) defines JPEG 2000 to be the recommended format for lossy compression (in previous editions also JPEG compression was supported). ANSI/NIST-ITL 1-2011 ("Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information") supports PNG and JPEG 2000 for lossless compression and JPEG 2000 only for applications tolerating lossy compression. While in the biometric community, lossy fingerprint compression attracted most researchers (e.g. [20]), also lossy compression of face [5] and iris image data has been discussed. For the latter case, [4,6,9,17] are early works covering an assessment on recognition accuracy for standard approaches covering different IREX formats (K3 for compression of cropped iris images, K7 for ROI-masked and cropped images, K16 referring to unsegmented polar format). In [7,12,13] methods to adapt compression techniques (customizing quantization tables, ROI-coding) for advanced iris recognition are examined. The attention of most techniques is focused on lossy compression, since bit-rate savings are more significant as compared to lossless techniques.

## 3 Image Compression in Iris-Biometric FCS

### 3.1 Experimental Setup

Experiments are carried out on CASIA-v3-Interval iris database[1]. At preprocessing the iris of a given sample image is segmented and normalized to a rectangular

[1] CASIA Iris Image Database, `http://www.idealtest.org`

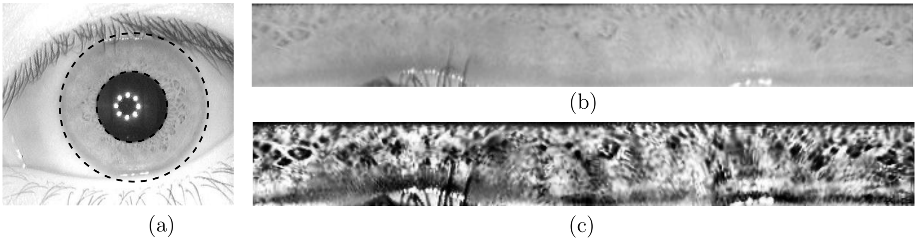

(a)
(b)
(c)

**Fig. 2.** Preprocessing and feature extraction: (a) segmented iris image (b) unwrapped iris texture and (c) preprocessed iris texture after enhancement

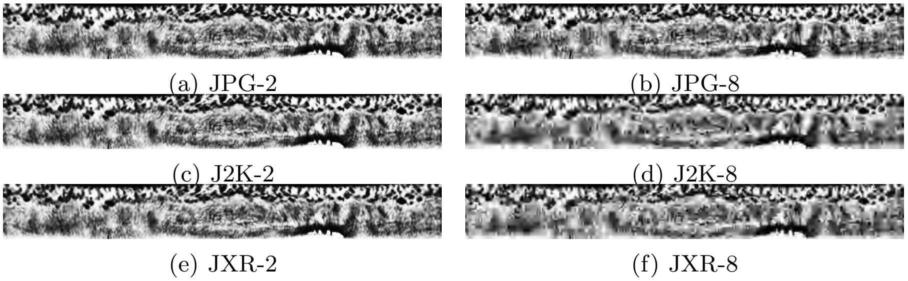

(a) JPG-2
(b) JPG-8
(c) J2K-2
(d) J2K-8
(e) JXR-2
(f) JXR-8

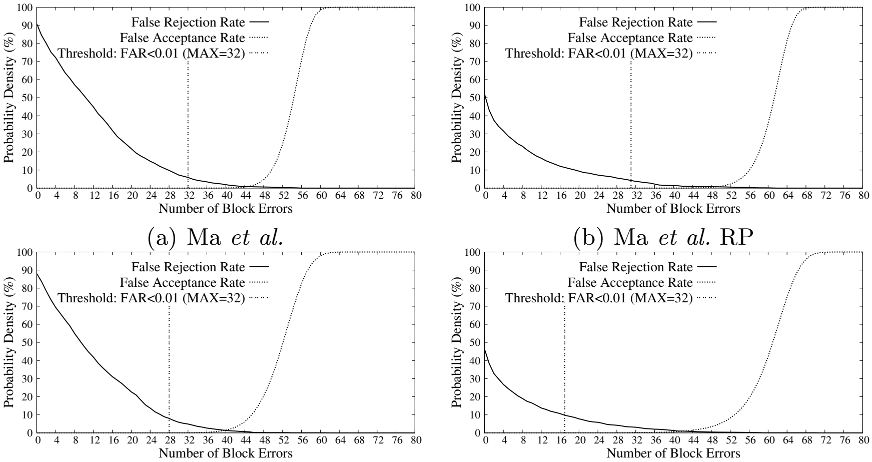
**Fig. 3.** Image Compression: (a)-(l) different levels of JPEG (JPG), JPEG 2000 (J2K), and JPEG XR (JXR) compression

texture of 512 × 64 pixel, see Fig. 2. In the feature extraction stage we employ custom implementations[2] of two different algorithms extracting a binary iris-code each: *Ma et al.* refers to the algorithm described in [14], which employs a dyadic wavelet transform on a stripified version of the iris texture. A $512 \times 20 = 10240$ bit code is generated for two fixed subbands encoding positions of all local minima and maxima. *Masek* refers to the open-source implementation of a 1D Daugman-like feature extraction[3] using convolution with Log-Gabor filters. By encoding the phase angle with 2 bits, again a 10240 bit iris-code is generated.

The applied FCS follows the approach in [8]. For both feature extraction algorithms, Ma et al. and Masek, Hadamard codewords of 128-bit and a Reed-Solomon code $RS(16, 80)$ are applied, which provided the best experimental results for a binding of 128-bit cryptographic keys: a $16{\cdot}8 = 128$ bit cryptographic key $R$ is prepared with a $RS(16, 80)$ Reed-Solomon code (which is capable of correcting $(80 - 16)/2 = 32$ block errors). All 80 8-bit blocks are processed by Hadamard encoding, expanding the length of codewords from length $n$ to $2^{n-1}$ (i.e. from 80 128-bit codewords to a 10240-bit bitstream). This way, up to 25% of bit errors can be detected and corrected. As a result, the bitstream is bound to the iris-code using the XOR operation and the commitment of the original key $h(R)$ is calculated using the hash function. At authentication, the key is retrieved by XORing an extracted iris-code with the first part of the commitment.

[2] USIT - University of Salzburg Iris Toolkit, `http://www.wavelab.at/sources/`

[3] L. Masek: Recognition of Human Iris Patterns for Biometric Identification, Master's thesis, University of Western Australia, 2003.



(a) Ma *et al.*

(b) Ma *et al.* RP

(c) Masek

(d) Masek RP

**Fig. 4.** Performance rates: (a)-(d) FCSs based on the algorithm of Ma *et al.* and Masek without applying image compression

Decoding using Hadamard and Reed-Solomon codes usually correct biometric variation and burst errors. In case the hashed versions are equal $(h(R') = h(R))$, the correct key $R$ is released, otherwise an error message is returned. Bringer [2] report, that a random permutation of bits in iris-codes improves key retrieval rates. We consider two types of FCSs, one in which iris-codes are left unaltered and one in which a single random permutation is applied to each iris-code of the database, denoted by FCS RP.

### 3.2 Image Compression

In the proposed case study image compression is applied to IREX K16 preprocessed iris textures. After image compression feature extraction is applied and resulting iris-codes are used to retrieve keys from stored commitments, where commitments are generated using un-compressed iris textures (see Fig. 1). That is, the proposed scenario provides a fair ground truth, *i.e.* by applying image compression to segmented iris textures the obtained key retrieval rates remain comparable. Different image compression standards are applied: (1) JPEG (ISO/IEC 10918): the well-established DCT-based method of compressing images, (2) JPEG 2000 (ISO/IEC 15444): a wavelet-based image compression standard, and (3) JPEG XR (ISO/IEC 29199-2): which, like JPEG 2000, generally provides better quality than JPEG but is more efficient than JPEG-2000, with respect to computational effort. For each standard, eight different compression levels with fixed bitrate are considered. In Fig. 3 examples of these compression levels are illustrated.

### 3.3 Performance Evaluation

Experimental results for both feature extractors and FCSs according to different compression levels are summarized in Table 2, including average PSNRs

**Table 2.** Summarized experiments for both feature extraction methods and FCSs under various JPG, J2K and JXR image compression levels

| | | | **Ma** *et al.* | | | | | **Masek** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Original | FCS | | FCS RP | | Original | FCS | | FCS RP | |
| Comp. | ∅ PSNR | ∅ Size | FRR at FAR≤0.01 | FRR at FAR≤0.01 | Corr. blocks | FRR at FAR≤0.01 | Corr. blocks | FRR at FAR≤0.01 | FRR at FAR≤0.01 | Corr. blocks | FRR at FAR≤0.01 | Corr. blocks |
| None | – | 1.00 | 2.54 % | 5.90 % | 32 | 3.72 % | 31 | 6.59 % | 8.01 % | 28 | 9.15 % | 17 |
| JPG-1 | 42.5 dB | 0.63 | 3.16 % | 6.94 % | 32 | 5.01 % | 31 | 8.75 % | 10.27 % | 27 | 10.81 % | 17 |
| JPG-2 | 37.2 dB | 0.49 | 3.37 % | 6.79 % | 32 | 4.40 % | 32 | 9.11 % | 10.11 % | 27 | 10.57 % | 17 |
| JPG-3 | 31.3 dB | 0.32 | 3.57 % | 6.75 % | 32 | 4.47 % | 32 | 9.95 % | 10.17 % | 27 | 10.11 % | 18 |
| JPG-4 | 28.9 dB | 0.26 | 3.62 % | 7.25 % | 32 | 4.41 % | 32 | 9.42 % | 10.19 % | 27 | 10.03 % | 18 |
| JPG-5 | 25.8 dB | 0.17 | 3.81 % | 6.94 % | 32 | 4.09 % | 32 | 9.83 % | 10.89 % | 27 | 9.80 % | 19 |
| JPG-6 | 24.3 dB | 0.13 | 4.50 % | 7.56 % | 32 | 4.71 % | 32 | 9.80 % | 10.42 % | 27 | 10.73 % | 17 |
| JPG-7 | 22.1 dB | 0.08 | 4.65 % | 7.72 % | 32 | 4.63 % | 32 | 9.54 % | 10.50 % | 27 | 10.03 % | 18 |
| JPG-8 | 20.2 dB | 0.05 | 5.55 % | 8.18 % | 32 | 4.86 % | 32 | 10.93 % | 11.58 % | 27 | 11.35 % | 18 |
| J2K-1 | 43.1 dB | 0.63 | 2.94 % | 7.43 % | 32 | 4.67 % | 32 | 8.65 % | 11.28 % | 26 | 10.25 % | 17 |
| J2K-2 | 39.6 dB | 0.49 | 3.04 % | 7.42 % | 32 | 4.27 % | 32 | 8.89 % | 9.83 % | 27 | 9.12 % | 18 |
| J2K-3 | 34.6 dB | 0.32 | 3.32 % | 6.97 % | 32 | 4.04 % | 31 | 9.29 % | 8.77 % | 28 | 8.62 % | 20 |
| J2K-4 | 30.7 dB | 0.26 | 3.71 % | 7.02 % | 32 | 4.32 % | 32 | 9.47 % | 9.19 % | 28 | 9.59 % | 19 |
| J2K-5 | 28.4 dB | 0.17 | 3.88 % | 6.51 % | 32 | 4.36 % | 32 | 9.58 % | 10.43 % | 27 | 9.13 % | 19 |
| J2K-6 | 24.9 dB | 0.13 | 3.96 % | 7.39 % | 32 | 4.02 % | 32 | 9.94 % | 12.41 % | 26 | 9.84 % | 20 |
| J2K-7 | 23.1 dB | 0.08 | 4.21 % | 7.28 % | 32 | 4.66 % | 32 | 10.05 % | 11.95 % | 26 | 10.02 % | 18 |
| J2K-8 | 21.9 dB | 0.05 | 4.55 % | 7.49 % | 32 | 5.21 % | 32 | 10.43 % | 10.23 % | 27 | 10.33 % | 17 |
| JXR-1 | 44.3 dB | 0.63 | 2.72 % | 6.82 % | 32 | 4.23 % | 32 | 9.75 % | 9.83 % | 27 | 9.13 % | 18 |
| JXR-2 | 40.9 dB | 0.49 | 3.09 % | 6.95 % | 32 | 3.78 % | 32 | 9.92 % | 9.97 % | 27 | 9.64 % | 17 |
| JXR-3 | 34.1 dB | 0.32 | 3.83 % | 6.22 % | 32 | 4.12 % | 32 | 10.05 % | 10.85 % | 26 | 10.09 % | 18 |
| JXR-4 | 32.9 dB | 0.26 | 4.79 % | 6.95 % | 32 | 4.34 % | 32 | 10.13 % | 9.55 % | 27 | 9.11 % | 19 |
| JXR-5 | 28.5 dB | 0.17 | 4.92 % | 7.58 % | 32 | 4.65 % | 32 | 10.61 % | 9.02 % | 28 | 9.08 % | 19 |
| JXR-6 | 25.1 dB | 0.13 | 5.03 % | 7.04 % | 32 | 4.70 % | 32 | 10.74 % | 11.98 % | 26 | 10.88 % | 17 |
| JXR-7 | 21.7 dB | 0.08 | 5.12 % | 8.16 % | 32 | 4.92 % | 32 | 11.48 % | 10.44 % | 27 | 10.76 % | 18 |
| JXR-8 | 22.9 dB | 0.05 | 5.18 % | 9.44 % | 32 | 5.79 % | 32 | 11.60 % | 14.92 % | 26 | 11.96 % | 18 |

caused by image compression, resulting filesizes and the number of corrected block errors after Hadamard decoding (*i.e.* error correction capacities may not handle the optimal amount of occurring errors within intra-class key retrievals). The FRR of a FCS defines the percentage of incorrect keys returned to genuine subjects. By analogy, the FAR defines the percentage of correct keys retrieved by non-genuine subjects. It is assumed that all subjects are registered under favorable conditions, *i.e.* commitments constructed using unaltered templates are de-committed applying degraded templates (*i.e.* computed from compressed data). For the recognition algorithm of Ma *et al.* and Masek FRRs of 2.54% and 6.59% are obtained at a FAR of 0.01% where the Hamming distance is applied as dis-similarity metric. Focusing on the feature extraction of Ma *et al.* FCSs provide FRRs of 5.90% in the original version and 3.72%, in the case case a random permutation is applied. FRRs are lower bounded by error correction capacities, *i.e.* bit-level error correction is applied more effectively if errors are distributed rather uniformly (see Fig. 4 (a) and (b)). With respect to the feature extraction of Masek, applying a random permutation does not improve the key retrieval rate obtaining FRRs of 8.01% and 9.15%, respectively.

For all applied compression standards a continuous significant degradation of recognition accuracy with respect to applied levels of compression is observed for both of the original iris recognition algorithms (see Table 2, column "Original HD"). For the highest compression levels FRRs of 5.55%, 4.55%, and 5.18% are obtained at FARs less than 0.01% for the JPEG (JPG), JPEG 2000 (J2K), and JPEG XR (JXR) compression standard for the algorithm of Ma *et al.*. For the feature extraction of Masek FRRs of 10.93%, 10.43%, and 11.60% are achieved at FARs less than 0.01% for the highest compression levels, *i.e.* recognition accuracy is significantly affected for high compression levels, while low compression levels almost maintain recognition accuracy of the schemes applied without any compression (*e.g.* JPG-1, J2K-1, and JXR-1). In contrast, while FCSs based on both feature extraction methods suffer from degradation in key retrieval rates, too, performance improves for average compression levels. It is found that incorporating image compression, at certain compression levels, improves key retrieval rates obtaining FRRs of $\sim$ 4.50% and 10.00% (RP), since, on average, extracted iris-codes are even more alike, *i.e.* image compression tends to blur iris textures (see Fig. 3) which is equivalent to de-noising. FCSs RP partially outperform the original recognition algorithms at higher compression levels. All types of investigated FCSs appear rather robust to a certain extent of image compression. As expected, the JPEG 2000 and JPEG XR compression standards provide higher image quality at certain file sizes with respect to PSNRs. However, higher quality according to PSNR values does not coincide with obtained recognition rates nor with key retrieval rates achieved by the applied FCSs, especially at higher compression levels (*e.g.* JPG-8 compression leads to better performance than J2K-8 or JXR-8 for the FCS RP of Ma *et al.*, even if JPG-8 provides lower quality in terms of PSNR). Uncompressed preprocessed iris textures exhibit a file size of 32.4 kB. According to ISO/IEC 19794-6 compressed iris images should reveal a file size of 25-30 kB in “rectilinear” format (and 2 kB in “polar” format as suggested in the older standard version, respectively). For the proposed FCSs acceptable rates are achieved for transferred iris textures of less than 2 kB (see Table 2), e.g. for J2K at FARs less than 0.01% FRRs of 5.21% and 10.33 % are obtained for FCSs RP, applying the algorithm of Ma *et al.* and Masek, where compressed iris textures exhibit a filesize of $32.4 \times 0.05 = 1.62$ kB (J2K–7).

## 4 Conclusion

This work investigated compression effects of IREX K16 iris images in a FCS. For all tested compression techniques JPEG, JPEG 2000 and JPEG XR, the application of compression induced a slight impact on key retrieval in case of high compression rates. However, in case of medium and slight compression, results were almost unaffected and at certain levels, compression with its de-noising effects was found to improve key retrieval. While this behavior is most likely due to the scenario employed (compression is applied after segmentation), recent studies highlight the critical impact of compression on segmentation. Nevertheless, the result illustrates a resilience of FCS for compression artifacts despite being claimed to be sensitive to noise.

## References

1. Ao, M., Li, S.Z.: Near infrared face based biometric key binding. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 376–385. Springer, Heidelberg (2009)
2. Bringer, J., Chabanne, H., Cohen, G., Kindarji, B., Zémor, G.: Theoretical and practical boundaries of binary secure sketches. IEEE Trans. on Information Forensics and Security 3, 673–683

3. Cavoukian, A., Stoianov, A.: Biometric encryption: The new breed of untraceable biometrics. In: Biometrics: fundamentals, theory, and systems. Wiley (2009)
4. Daugman, J., Downing, C.: Effect of severe image compression on iris recognition performance. IEEE Trans. on Inf. Forensics and Sec. 3(1), 52–61 (2008)
5. Delac, K., Grgic, M., Grgic, S.: Effects of JPEG and JPEG2000 compression on face recognition. In: Singh, S., Singh, M., Apte, C., Perner, P. (eds.) ICAPR 2005. LNCS, vol. 3687, pp. 136–145. Springer, Heidelberg (2005)
6. Grother, P.: Quantitative standardization of iris image formats. In: Proc. of the Biometrics and Electronic Signatures (BIOSIG 2009). LNI, pp. 143–154 (2009)
7. Hämmerle-Uhl, J., Prähauser, C., Starzacher, T., Uhl, A.: Improving compressed iris recognition accuracy using JPEG2000 roI coding. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1102–1111. Springer, Heidelberg (2009)
8. Hao, F., Anderson, R., Daugman, J.: Combining Cryptography with Biometrics Effectively. IEEE Trans. on Computers 55(9), 1081–1088 (2006)
9. Ives, R.W., Broussard, R.P., Kennell, L.R., Soldan, D.L.: Effects of image compression on iris recognition system performance. Journal of Electronic Imaging 17, 11015 (2008), doi:10.1117/1.2891313
10. Jain, A.K., Nandakumar, K., Nagar, A.: Biometric template security. EURASIP J. Adv. Signal Process 2008, 1–17 (2008)
11. Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: Sixth ACM Conference on Computer and Communications Security, pp. 28–36 (1999)
12. 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.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1091–1101. Springer, Heidelberg (2009)
13. Kostmajer, G.S., Stögner, H., Uhl, A.: Custom JPEG quantization for improved iris recognition accuracy. In: Gritzalis, D., Lopez, J. (eds.) SEC 2009. IFIP AICT, vol. 297, pp. 76–86. Springer, Heidelberg (2009)
14. Ma, L., Tan, T., Wang, Y., Zhang, D.: Efficient Iris Recogntion by Characterizing Key Local Variations. IEEE Trans. on Image Processing 13(6), 739–750 (2004)
15. Maiorana, E., Campisi, P.: Fuzzy commitment for function based signature template protection. IEEE Signal Processing Letters 17, 249–252 (2010)
16. Nandakumar, K.: A fingerprint cryptosystem based on minutiae phase spectrum. In: Proc. of IEEE Workshop on Information Forensics and Security (WIFS) (2010)
17. Rakshit, S., Monro, D.M.: An evaluation of image sampling and compression for human iris recognition. IEEE Trans. Inf. Forensics and Sec. 2, 605–612 (2007)
18. Rathgeb, C., Uhl, A.: Adaptive fuzzy commitment scheme based on iris-code error analysis. In: Proc. of the 2nd Europ. Workshop on Visual Inf. Proc. (EUVIP 2010), pp. 41–44 (2010)
19. Rathgeb, C., Uhl, A., Wild, P.: Reliability-balanced feature level fusion for fuzzy commitment scheme. In: Int'l Joint Conf. on Biometrics, pp. 1–7 (2011)
20. Sherlock, B.G., Monro, D.M.: Optimized wavelets for fingerprint compression. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 1996), Atlanta, GA, USA (May 1996)
21. Teoh, A., Kim, J.: Secure biometric template protection in fuzzy commitment scheme. IEICE Electron. Express 4(23), 724–730 (2007)
22. Van der Veen, M., Kevenaar, T., Schrijen, G.-J., Akkermans, T.H., Zuo, F.: Face biometrics with renewable templates. In: SPIE Proc. on Security, Steganography, and Watermarking of Multimedia Contents, vol. 6072, pp. 205–216 (2006)
23. Zhang, L., Sun, Z., Tan, T., Hu, S.: Robust biometric key extraction based on iris cryptosystem. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1060–1069. Springer, Heidelberg (2009)

# Person Re-identification Using Partial Least Squares Appearance Modeling

Gabriel Lorencetti Prado[1], William Robson Schwartz[2], and Helio Pedrini[1]

[1] Institute of Computing,
University of Campinas, Campinas, Brazil, 13083-852
gabriel.prado@students.ic.unicamp.br, helio@ic.unicamp.br
[2] Department of Computer Science,
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
william@dcc.ufmg.br

**Abstract.** Due to the large areas covered by surveillance systems, employed cameras usually lack intersection of field of view, refraining us from mapping the location of a person in a camera to another one. Therefore, when a subject appears in a camera, a person re-identification method is required to discover whether the subject has been previously identified in a different camera. Even though several approaches have been proposed in the literature, person re-identification is still a challenging problem due to appearance variation between cameras, changes in illumination, pose variation, and low quality data, among others. To reduce the effect of the aforementioned difficulties, we propose a person re-identification approach that models the appearance of the subjects based on multiple samples collected from multiple cameras and employs person detection and tracking to enhance the robustness of the method. Experiments conducted on three public available data sets demonstrate improvements over existing methods.

**Keywords:** person re-identification, partial least squares, appearance-based modeling, person detection, object tracking.

## 1 Introduction

Person re-identification consists on tracking multiple people in a camera network, registering their trajectories on each camera and assigning them consistent identifiers across the network. It is still an open problem in the Computer Vision area, as many challenges are faced when designing a robust re-identification system.

In the single camera case, issues include occlusion, pose variation and lighting conditions. Besides these difficulties, the major problem in a multi-camera system is to maintain a correspondence of the people tracked across cameras and people re-entering the camera field of view (FOV). When the camera calibration is not given or there is no overlapping field of views, the solution is even harder to be found, since homography-based methods such as [1] do not apply.

Several approaches have been proposed to address the re-identification problem and can be labeled in terms of different features. The classification under

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 382–390, 2013.

single or multiple shot methods is the most common and refers to whether the method considers only one image to create a person representation or a group of images to perform it. In the following paragraphs, we present a brief review of works more related to the proposed approach. An extended description and discussion of person re-identification methods can be found in [2,3].

Some approaches based on a single image are [4,5,6]. In all these methods, the problem is treated as a classification task. Considering the multiple shot case, many distinct methods have been proposed. Interest point based approaches are presented in [7,8]. Some methods using person appearance modeling are given in [7,9,10]. Other strategies include the use of body regions [11], particle filtering [12], global descriptors [13], a master-slave approach [14], and finally a transfer approach to the scenario where only a subset of people is considered for [15]. Such methods present environmental constraints that limit their usage in a general condition. Unlike these works, our method introduces an appearance-based modeling that can be applied in a wider variety of real world scenarios.

This paper proposes a method to tackle the re-identification problem under multiple shot context built upon the work proposed in [5,6]. Similarly to these methods, appearance models based on Partial Least Squares (PLS) regressions [16] are used to grant the system a powerful discriminative characteristic. However, in our method, these models are associated to the information obtained from person tracking and detection to enhance the robustness of the original method. Moreover, we do not require that a person gallery is given, that is, classes in which people will be classified are not known in advance. Hence, our approach aims at the general scenario, in which cameras do not need to share FOV, camera calibration is ignored, and synchronized frames are not required.

The proposed technique is evaluated on three video data sets, which were also used for the ICPR 2012 Contest - *People tracking in wide baseline camera networks* [17]. The approach is compared with the method proposed in [5] to demonstrate the necessity of a tracking module to achieve better results for the person re-identification problem.

## 2 Proposed Method

Our approach to the re-identification problem is based on a full body appearance-based modeling via PLS regression [16]. The proposed method is composed of five stages, which are explained in details after a brief description of the Partial Least Squares regression technique. Figure 1 illustrates an overview of the method.

### 2.1 Partial Least Squares

In PLS regression, the input consists of a collection of classes in which the samples are classified and a collection of high dimensional feature vectors for each class. The method then significantly reduces the dimensionality of the feature vectors by creating variables (latent variables), which are obtained as linear combination of the original ones.

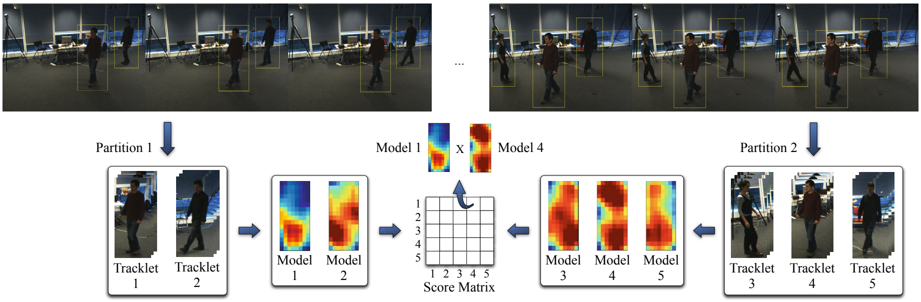


**Fig. 1. Visual description of the proposed method.** People are first detected by a pedestrian detector based on Partial Least Squares (PLS), as shown in the top images.People detected in continuous frames are then grouped into tracklets.These tracklets are put into partitions, where tracklets known to belong to different people are put in the same partition. A PLS model is then created for each tracklet in a one-against-all approach using the remaining tracklets from its partition as counter-examples. Afterwards, a score matrix is built by evaluating the matching between each pair of models. Finally, the score matrix is used to discover which models (therefore, tracklets) belong to the same person by removing the pairs with higher matching scores and assigning them to the same person.

The dimension reduction is performed in such a way that the covariance between the classes (subject's identifiers) and their feature vectors is maximized. The result is a collection of weight vectors that can be used to reduce the dimensionality of new feature vectors and classify them in the low dimensionality space using a regression-based approach [18].

### 2.2 Person Re-identification Based on Partial Least Squares

The goal of the first step is to obtain the detection windows for people contained in each frame of the video sequence. To do so, we first train a person detector based on a PLS appearance model that works with two classes (positive and negative person samples) [19]. The training is achieved by cropping the samples into overlapping blocks and extracting low-level features from each of them.

Detection windows are found by decomposing each frame into samples of increasing sizes and then extracting the same features from each of them. A non-maximum suppression is applied in the results to clean up redundant detection windows in multiple scales. A more detailed description can be found in [19].

The next step is to group detection windows from sequential frames of the same camera into tracklets. This is performed by tracking each detection window with a Kalman filter [20]. This approach also allows the tracking to fill in missing detections when they occur during a small number of frames.

Tracklets from different detections of the same frame naturally correspond to different people. This information is maintained throughout the method by keeping a set of tracklet partitions, where each partition contains tracklets that correspond to different people. A new partition is created whenever a tracklet is lost due to missing detections or a tracked person walks off the camera range.

For each tracklet partition, a new PLS appearance model is created for each tracklet in a one-against-all approach [5], using the remaining tracklets in the same partition as counter-examples. Then, all tracklets are pairwise matched against each other using the generated models. The matching results are used to build a matrix of matching scores between each tracklet pair.

The next step consists on successive removals of the maximum value from the matrix, marking the corresponding tracklets as belonging to the same person. Auxiliary maps are used in this stage and are described as follows.

For each tracklet, a similarity map ($S$) and a distinctivity map ($D$) are used. As soon as the tracklet is created, it is inserted into its own similarity map (Equation 1a). All other tracklets in the same partition ($P$) are inserted into its distinctivity map (Equation 1b). When a matching is found in the score matrix, both maps for the two involved tracklets are merged (Equations 2a) unless one tracklet is found in the other distinctivity map. This new similarity map is used to update each of the distinctivity maps for the tracklets found in the new distinctivity map (Equation 2b).

Initialization equations:

$$S_i \leftarrow i, \quad \forall i \in P_p \qquad (1a) \qquad\qquad D_i \leftarrow j, \quad \forall i, j \in P_p \mid i \neq j \qquad (1b)$$

Update equations:

$$\begin{aligned} S_{\text{new}} &\leftarrow S_i \cup S_j & D_{\text{new}} &\leftarrow D_i \cup D_j & \text{(2a)} \\ S_k &\leftarrow S_{\text{new}}, \quad \forall k \in S_{\text{new}} & D_k &\leftarrow D_{\text{new}}, \quad \forall k \in S_{\text{new}} & \\ & & D_k &\leftarrow D_k \cup S_{\text{new}}, \quad \forall k \in D_{\text{new}} & \text{(2b)} \end{aligned}$$

Finally, when no more matrix removals are possible, unique identifiers are assigned to each tracklet according to the information held in the similarity maps, in order to obtain the final tracklet-person matching.

## 3 Experimental Results

In this section, we present the evaluation of our method for the task of person re-identification using three data sets and its comparison to a baseline method [5]. Experiments were run offline, so the method is not ready for real time usage.

For the detection stage, a full body person detector was trained using cropped person images from the INRIA Person Dataset [21]. Only person images larger than 100 pixels were considered. The adopted features for both detection and modeling stages were histogram of oriented gradients [21] and gray-level co-occurrence matrices [22] to obtain edge and texture information, respectively.

Our re-identification approach was evaluated on three public video data sets, described as follows and summarized in Table 1. The first data set was CAT data set [23], in which people are mainly detected from only one of the cameras at a time. The second data set was sequence S7 of PETS2006 [24], where there is a larger number of frames with multiple person detections. The last data set considered was the Techgate data set. Unlike the previous data sets, cameras on this one are very close to the subjects.

**Table 1.** Summary of the data sets. Frames per camera denotes the number of frames available per camera; Camera sync indicates if the camera frames are synchronized; FOV Overlap indicates the amount of overlapping in the camera field of view.

| | **CAT** | **PETS2006** | **Techgate** |
|---|---|---|---|
| Number of cameras | 4 | 4 | 6 |
| Number of people | 4 | 56 | 4 |
| Frames per second | 15 | 25 | 15-30 |
| Frames per camera | 710 | 3401 | 3037-4415 |
| Camera sync | Yes | Yes | No |
| FOV Overlap | Low | High | High |

For the single camera case, selected metrics were Correct Detected Track (CDT), Track Detection Failure (TDF), False Alarm Track (FAT), and Track Fragmentation (TF) [25]. Best results are achieved when CDT is high and TDF, FAT and TF are low. Results for these metrics are shown in Tables 2-4.

For the multi-camera case, Crossing Fragments (X-Frag), Crossing ID Switches (X-IDS), Returning Fragments (R-Frag) and Returning ID Switches (R-IDS) [9] were selected. For all these metrics, lower values indicate better results. Results for the multi-camera metrics are shown in Table 5.

In Tables 2-5, metrics are displayed in percentages instead of absolute values and TF is shown as an average measure over all tracks in order to have a more general view of the results. Tracks shorter than 1 second were discarded, as suggested in the ICPR 2012 Contest [17] to avoid influence of short tracks.

The method described in [5] is used as baseline. We evaluate both methods with the same detections described above. Unlike ours, the baseline method requires a person gallery to classify the detected people. It is also used by the method to build person models in a one-against-all schema. Therefore, a gallery was created for each data set using person detections from the respective ground truth. For each person, 5 samples were randomly selected. To avoid influence from the training, results are an average of 10 executions with different galleries.

For the CAT data set, the proposed approach achieved accurate results for the single camera tracking, as it can be seen in Table 2. The CDT metric is quite high and the others present low values as expected for a good performance.

**Table 2.** Single camera metrics for each camera $C_i$ of CAT data set

| **CAT** | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| CDT | 45.6% | 27.2% | 10.8% | 26.7% |
| TDF | 54.4% | 72.8% | 89.2% | 73.3% |
| FAT | 0.0% | 0.0% | 0.0% | 0.5% |
| TF | 0.3 | 0.2 | 0.0 | 0.0 |

(a) Baseline Results

| **CAT** | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| CDT | 100% | 94.4% | 76.9% | 83.3% |
| TDF | 0.0% | 5.6% | 23.1% | 16.7% |
| FAT | 0.0% | 0.6% | 7.7% | 6.3% |
| TF | 0.5 | 1.3 | 0.2 | 1.2 |

(b) Proposed Method Results

On the PETS2006 data set results, presented in Table 3, it can also be noted that the proposed method performance is higher than the baseline method. However, it was lower than the CAT data set due to the large number of people in the scene. In the second camera, performance is much lower since this camera is positioned at a quite higher place. Detections are therefore smaller and, consequently, fewer features can be extracted to perform the appearance modeling.

**Table 3.** Single camera metrics for each camera $C_i$ of PETS2006 data set

| **PETS** | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| CDT | 5.1% | 1.5% | 3.0% | 0.6% |
| TDF | 94.9% | 98.1% | 97.0% | 99.4% |
| FAT | 0.0% | 0.1% | 0.0% | 0.0% |
| TF | 0.2 | 0.0 | 0.1 | 0.0 |

(a) Baseline Results

| **PETS** | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| CDT | 89.8% | 37.3% | 83.0% | 85.2% |
| TDF | 10.2% | 62.7% | 17.0% | 14.8% |
| FAT | 2.2% | 3.3% | 1.0% | 1.8% |
| TF | 0.6 | 0.5 | 1.0 | 0.4 |

(b) Proposed Method Results

In Techgate data set, an issue opposite to the found in PETS2006 data set occurs at some detections. The proximity of some people with the camera does not allow the full body person detector to find them, also reducing the system performance. Nevertheless, our method still produces better results.

**Table 4.** Single camera metrics for each camera $C_i$ of Techgate data set

| **Tech** | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| CDT | 12.5% | 13.8% | 22.9% | 9.0% | 22.0% | 26.4% |
| TDF | 87.5% | 86.2% | 77.1% | 91.0% | 78.0% | 73.6% |
| FAT | 0.9% | 0.4% | 0.1% | 0.0% | 2.3% | 0.1% |
| TF | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.5 |

(a) Baseline Results

| **Tech** | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ |
|---|---|---|---|---|---|---|
| CDT | 75.0% | 92.3% | 73.5% | 50.0% | 60.0% | 71.8% |
| TDF | 25.0% | 7.7% | 26.5% | 50.0% | 40.0% | 28.2% |
| FAT | 5.5% | 5.5% | 1.3% | 1.6% | 14.8% | 1.5% |
| TF | 1.2 | 1.3 | 0.8 | 0.4 | 0.0 | 1.0 |

(b) Proposed Method Results

Finally, for the multi-camera metric results shown in Table 5, it can be seen that our method demonstrates superior results for the X-Frag and R-Frag metrics. However, it is not as good as the baseline with metrics X-IDS and R-IDS.

**Table 5.** Multi-camera metrics for each data set

| | **CAT** | **PETS** | **Tech** |
|---|---|---|---|
| X-Frag | 95.7% | 100% | 98.3% |
| X-IDS | 0.0% | 0.0% | 0.1% |
| R-Frag | 94.9% | 99.8% | 96.9% |
| R-IDS | 0.1% | 0.0% | 0.1% |

(a) Baseline Results

| | **CAT** | **PETS** | **Tech** |
|---|---|---|---|
| X-Frag | 58.0% | 85.4% | 73.2% |
| X-IDS | 15.2% | 10.1% | 10.9% |
| R-Frag | 56.4% | 80.8% | 71.3% |
| R-IDS | 24.7% | 12.5% | 12.5% |

(b) Proposed Method Results

From the results presented above, it is clear that results for the proposed approach overcomes the baseline ones in all the considered data sets.

For the CDT metric, results are highly superior due to the addition of tracking as shown in Tables 2-4. This feature makes our method more robust against change of ids in the same tracklet, which causes CDT metric to achieve a higher performance. The same applies for the TDF metric.

Results are also low for the other single camera metrics, pointing out that not only our method is able to discover more correct tracks, but it also does not add false positive tracks to the results. Actually, for both methods, this metric depends on false positive detections, which are shared by them both.

In the results for the multi-camera tracking metrics X-Frag and R-Frag, it can also be observed that our method achieves better results when compared to the baseline, as depicted in Table 5. For the R-IDS and X-IDS metrics, our method is also able to keep them low, however, higher than the baseline. This fact can be explained because baseline results are low due to the small amount of detected tracks, reflected in the CDT and TDF metrics. Furthermore, these metrics are also highly dependent on the detection results, which were clearly higher for the CAT data set when compared to the others.

## 4 Conclusions and Future Work

In this work, we presented a novel appearance-based modeling method for person re-identification across multiple cameras. Our system does not have any restriction on the camera network configuration, such as calibration, FOV intersection or frame synchronization. Therefore, it is sufficiently general to be used in many data sets with different characteristics, as demonstrated in the experiments. Moreover, it can be easily modified to use other low level features.

The feasibility of the proposed approach was demonstrated by its execution on three different data sets and the obtained results are evaluated in terms of current state-of-art metrics for both single and multiple camera tracking. Experimental results demonstrated that it outperforms the baseline method regarding most of considered metrics for both single and multiple camera cases.

As future work, we plan to investigate how to update PLS models online, as new person associations are discovered during the matching stage.

**Acknowledgements.** The authors are grateful to FAPESP, FAPEMIG, CNPq, and CAPES for the financial support and to the organizers of the ICPR 2012 Contest - *People tracking in wide baseline camera networks* [17] for providing the data sets.

## References

1. Hu, W., Hu, M., Zhou, X., Tan, T., Lou, J., Maybank, S.: Principal Axis-Based Correspondence between Multiple Cameras for People Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 663–671 (2006)
2. Mazzon, R., Tahir, S.F., Cavallaro, A.: Person re-identification in crowd. Pattern Recognition Letters 33(14), 1828–1837 (2012)
3. Satta, R.: Appearance Descriptors for Person Re-identification: a Comprehensive Review. CoRR abs/1307.5748 (2013)
4. Hirzer, M., Beleznai, C., Köstinger, M., Roth, P.M., Bischof, H.: Dense Appearance Modeling and Efficient Learning of Camera Transitions for Person Re-Identification. In: International Conference on Image Processing (2012)

5. Schwartz, W.R., Davis, L.S.: Learning Discriminative Appearance-Based Models Using Partial Least Squares. In: Brazilian Symposium on Computer Graphics and Image Processing (2009)
6. Schwartz, W.R.: Scalable People Re-Identification Based on a One-Against-Some Classification Scheme. In: International Conference on Image Processing (2012)
7. Gheissari, N., Sebastian, T.B., Hartley, R.: Person Reidentification Using Spatiotemporal Appearance. In: Computer Vision and Pattern Recognition (2006)
8. Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In: International Conference on Distributed Smart Cameras (2008)
9. Kuo, C.-H., Huang, C., Nevatia, R.: Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 383–396. Springer, Heidelberg (2010)
10. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-Shot Person Re-identification by HPE Signature. In: International Conference on Pattern Recognition (2010)
11. Bazzani, L., Cristani, M., Murino, V.: Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identification. Computer Vision and Image Understanding 117(2), 130–144 (2013)
12. Li, M., Chen, W., Huang, K., Tan, T.: Visual Tracking via Incremental Self-tuning Particle Filtering on the Affine Group. In: Computer Vision and Pattern Recognition (2010)
13. Cai, Y., Pietikäinen, M.: Person Re-identification Based on Global Color Context. In: Koch, R., Huang, F. (eds.) ACCV 2010 Workshops, Part I. LNCS, vol. 6468, pp. 205–215. Springer, Heidelberg (2011)
14. Alahi, A., Vandergheynst, P., Bierlaire, M., Kunt, M.: Cascade of Descriptors to Detect and Track Objects Across Any Network of Cameras. Computer Vision and Image Understanding 114(6), 624–640 (2010)
15. Zheng, W.S., Gong, S., Xiang, T.: Transfer Re-identification: From Person to Set-based Verification. In: Computer Vision and Pattern Recognition (2012)
16. Wold, H.: Partial Least Squares. Encyclopedia of Statistical Sciences 6, 581–591 (1985)
17. International Conference on Pattern Recognition Contest - People tracking in wide baseline camera networks (2012), `http://www.wide-baseline-camera-network-contest.org`
18. Schwartz, W.R., Guo, H., Choi, J., Davis, L.S.: Face Identification Using Large Feature Sets. IEEE Transactions on Image Processing 21(4), 2245–2255 (2012)
19. Schwartz, W.R., Kembhavi, A., Harwood, D., Davis, L.S.: Human Detection Using Partial Least Squares Analysis. In: International Conference on Computer Vision (2009)
20. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. Technical report, Department of Computer Science, University of North Carolina, USA (1995)
21. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Computer Vision and Pattern Recognition (2005)
22. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics SMC 3, 610–621 (1973)

23. CAT Project, `http://www.cat-project.at`
24. Thirde, D., Li, L., Ferryman, J.: Overview of the PETS2006 Challenge. In: International Workshop on Performance Evaluation of Tracking and Surveillance (2006)
25. Yin, F., Makris, D., Velastin, S.A.: Performance Evaluation of Object Tracking Algorithms. In: International Workshop on Performance Evaluation of Tracking and Surveillance (2007)

# A New Iris Recognition Approach Based on a Functional Representation

Dania Porro-Muñoz, Francisco José Silva-Mata, Victor Mendiola-Lau,
Noslen Hernández, and Isneri Talavera

Advanced Technologies Application Center (CENATAV) - Cuba
{dpmunoz,fjsilva,vmendiola,nhernandez,italavera}@cenatav.co.cu

**Abstract.** This paper proposes the introduction of annular Zernike polynomials for representing iris images data. This representation offers notables advantages like representing the images on a continuous domain that allows the application of Functional Data Analysis techniques, preserving their original nature. In addition, it provides a significant dimensionality reduction of the data, while it still has a high discriminative power. The proposed approach also deals with the occlusion problems that can be present in this type of images. In order to corroborate the effectiveness of the introduced approach, identification experiments were carried out. Iris international databases were used. Some of them are characterized by the presence of severe occlusion problems. Results have shown high recognition accuracy.

**Keywords:** Iris recognition, Functional Data Analysis.

## 1 Introduction

Eyes texture has practically unrepeatable patterns among human beings, even twins, which makes the iris a biometric entity usable for people identification [1]. In iris recognition research area, there are still many open problems that require of innovative solutions related to the steps of iris recognition(image capture, eye localization, segmentation, noise detection, normalization, feature extraction and matching)[2].

Traditionally, iris features have been represented by high-dimensional vectors. However, the fact that digital images data are recorded discretely, though their nature is continuous is ignored. Therefore, instead of representing these images by vectors, it might be more appropriate to represent the iris image data by an underlying continuous function, in a way that this representation can be closer to the original nature of the data. Using functions for these images representation brings many advantages, such as: data is represented as a whole, the revelation of the dynamic aspects of the original data, the ability of analyzing some of the most significant features of the function, for example, the monotony, differentiability and smoothness, and a substantial dimensionality reduction of the data[3]. An image can be represented as a function $f(x,y)$ over

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 391–398, 2013.

a spatial domain [4]. Finding such function that better approximates the discrete observed data is one of the key points of this approach. Using functional basis expansion is a common issue for this representation. The selection of the basis set and the estimation of the appropriate number of coefficients must be performed thinking in obtaining a sufficiently discriminative representation with a minimum dimension. The proposed solution for the particular case of iris images, demonstrates the validity of the approach and traces a new general methodology for analyzing biometric images.

The paper is organized as follows. In Section 2 the iris recognition process is explained, and our introduced approach is detailed. In Section 3 the performed experiments and results are described. Conclusions and future research topics are drawn in Section 4.

## 2 Iris Recognition Process

Iris recognition process consists on the general steps(image capture, eye localization, segmentation, noise detection, normalization, feature extraction and matching). Nevertheless, these steps present some peculiarities when Functional Data Analysis(FDA) approach is used. The general diagram presented in Fig. 1 shows the necessary steps to perform an image recognition task based on functional data analysis. This helps for a better understanding of the process.

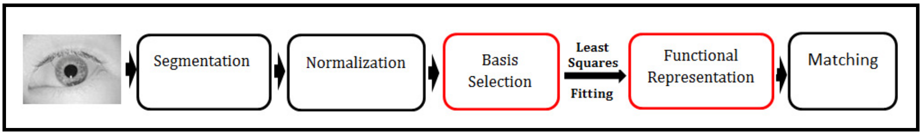


**Fig. 1.** A general description of the iris image analysis using FDA

### 2.1 Segmentation and Normalization

The segmentation step is applied to separate the iris region from the other part of the eye image. The main goal of this step is to extract the region of interest(ROI), containing the maximum amount of pixels with valid information, and the minimum with irrelevant information. During the normalization stage, the isolated region of the iris is reduced to specific dimensions(specified by the radial and angular resolution) using the coordinates transformation.

For these two steps of the iris recognition process(segmentation and normalization), we used the third module(Iris recognition) of the known Video-based Automated System for Iris Recognition (VASIR) [5]. This system segments the iris region using their own segmentation approach. Those regions are then extracted and normalized based on the known *"rubber-sheet"* method. Each point within the iris region is assigned a pair of real coordinates $(r, \theta)$ contained in a rectangle where the radius $r$ lies on the unit interval $[0,1]$ and $\theta$ is the angle over $[0,2\Pi]$.

Before constructing the functional representation, it must be ensured that it is the least affected by the factors that can be present in the ROI. This factors could be: dilation, specular reflection, iris resolution, motion blur, camera diffusion, presence of eyelids, eyelashes, and others[6]. In particular, the occlusion by eyelashes and eyelids significantly affects the behavior of this type of representation.

Some solutions use binary masks to face these occlusion problems [7]. However, most of them are based on representations that use local features. Other solutions have restricted the region of interest of the iris domain[8], by taking into account its statistical behavior with respect to the presence of occlusions. This region should be the least affected by occlusions of eyelids and eyelashes. In our research, we built our representation after selecting different regions of the iris images, as explained below. In Fig.2(a), half of the iris region was removed, keeping the region comprised in $[-\pi/4,+\pi/4]$ and $[3\pi/4,5\pi/4]$(upper and bottom sectors indicated with a cross )[7]. In Fig.2(b) it is shown that those sectors selected in Fig.2 a) were reduced, in order to decrease the incidence of the eyelids and eyelashes comprised in $[-\pi/4,+\pi/8]$ and $[7\pi/8,5\pi/4]$. Finally, in Fig.2 c), a 1/4 of the rectangular region is selected, minimizing even more the incidence of occlusion as is proposed in [8]. For each of the selected regions, all the images were reduced to an specific dimension 32x128, 32x48 and 16*64 respectively.

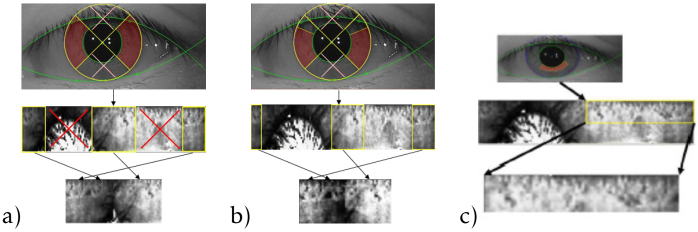


**Fig. 2.** Three different regions with different of occlusion levels

## 2.2 Representation

After the image has been normalized, we proceed to the construction of the functional representation. The method consists in approximating(smoothing) each image by a linear combination (weighted sum) of basis functions, which is a common strategy for achieving this type of representation [3].

$$I_f(x,y) = \sum_{k=1}^{K} c_k b_k(x,y), \tag{1}$$

where $\{b_k(x,y)\}_k$ denotes the set of $k$ bases functions and $\{c_k\}_k$ represents the coefficients of the expansion.

Finally, from the discrete observations $I(x,y)$ i.e, the iris normalized images, a functional representation $\hat{I}_f(x,y)$, is obtained. Images will now be described by the coefficients $\{c_k\}$. It has been demonstrated that working with these coefficients is strictly equivalent to working directly on the $b_k$ functions[9].

One important step to consider in this process is to choose the best basis set to expand the function. There are several criteria for selecting the basis set, which are explained bellow.

**Basis Selection and Determination of the Number of Coefficients.** Some of the principal criteria that are considered for basis selection are: the geometry of the domain [10], computational complexity, differentiability, periodicity of the event to model, the ratio of speed of convergence, completeness (understood as the ability to represent any function with high precision and with enough terms coefficients). For our selection, we will mainly focus on the geometry of the domain. Since we are interested in approximating functions of two variables, we need bivariate basis functions for the expansion. The annular Zernike polynomials are a basis functions set defined over an unitary annulus, which is suitable for modeling the iris domain due to its shape.

Any function $I_f(\rho,\theta)$ defined over a two-dimensional space can be approximated using the annular Zernike polynomials bases[11]. Thus, the equation for expanding a sector of an annular iris region($\hat{I}_f(\rho,\theta,\epsilon)$) in terms of annular polynomials $Z_n^m(\rho,\theta,\epsilon)$ that are orthonormal over a unit annulus is:

$$\hat{I}_f(\rho,\theta,\epsilon) = \sum_{n}^{\infty}\sum_{m}^{n} C_n^m Z_n^m(\rho,\theta,\epsilon) \tag{2}$$

where $C_n^m$ represents the vector of coefficients and $Z_n^m$ represents the annular Zernike polynomials basis functions. The annular Zernike polynomials are similar to circular Zernike polynomials, except that are orthonormal in an annulus instead of a circle [11]. They are usually defined in polar coordinates $(\rho,\theta)$. The parameter $\rho \in [0,1]$, is the radial coordinate, and $\theta \in [0,2\pi]$, the azimuthal component. Annular Zernike polynomials have inner radius $\epsilon$ and outer radius 1, and thus the coordinate $\rho$ is subject to the restriction $0 \leq \epsilon \leq 1$. For the specific case of iris, the value $\rho$ is comprised between the pupillary boundary and the limbus boundary.

These polynomials are derived from the circular Zernike polynomials by Gram-Schmidt orthogonalization process[11]. Each Zernike polynomial is a tensor product of Fourier bases in the angular direction and a special type of Jacobi polynomials in the radial direction[11]. It consists of 3 components: a normalization factor, a radial component and one azimuthal component. Annular Zernike polynomials are then defined as follows:

$$Z_n^{\pm m} = \begin{cases} N_n^m R_n^{|m|}(\rho,\epsilon)\cos(m\theta) & if \quad m \geq 0 \\ -N_n^m R_n^{|m|}(\rho,\epsilon)\sin(m\theta) & if \quad m < 0., \end{cases} \tag{3}$$

For a given radial order or polynomial order $n$, the azimuthal frequency or Fourier order $m$ can only take values of $-n, -n+2, -n+4, ..., n$ [11]. $N_n^m$ is the normalization factor

$$N_n^m = \sqrt{\frac{2(n+1)}{1+\delta_{m0}}}, \quad \text{with} \quad \delta_{m0} = \begin{cases} 1 \text{ if} & m = 0 \\ 0 \text{ if} & m \neq 0 \end{cases}, \tag{4}$$

and $R_n^{|m|}(\rho, \epsilon)$ is the representation for the Jacobi polynomial, which for the general case can be expressed as:

$$\mathrm{R}_{2j+m}^m(\rho, \epsilon) = \left[ \frac{1-\epsilon^2}{2(2j+m+1)\mathrm{h}_j^m} \right]^{\frac{1}{2}} \rho^m \mathrm{Q}_j^m(\rho^2) \tag{5}$$

where $\left\{ Q_j^m(u) \right\}$[1] is a set of orthogonal polynomials obtained by the orthonormalization of the sequence $1, u, \ldots, u^j$ over the interval $(\epsilon^2, 1)$. It is defined as:

$$Q_j^m(u) = \begin{cases} R_{2j}^0(\rho, \epsilon) & \text{si } m = 0 \\ \frac{2(2j+2m-1)}{(j+m)(1-\epsilon^2)} \frac{h_j^{m-1}}{Q_j^{m-1}(0)} \sum_{i=0}^{j} \frac{Q_i^{m-1}(0) Q_i^{m-1}(u)}{h_i^{m-1}} & e.o.c \end{cases} \tag{6}$$

These expressions were taken from [11]. In order to obtain an estimate $\hat{I}_f(\rho, \theta)$ of $I_f(\rho, \theta)$ for each image $I(\rho, \theta)$, we need to estimate the coefficients in the expansion. This will be done by least squares fitting. Due to the fact that Zernike polynomials are orthonormal in the unit sphere, i.e.,

$$\frac{1}{\pi(1-\epsilon^2)} \int_\epsilon^1 \int_0^{2\pi} Z_n^m(\rho, \theta, \epsilon) Z_{n'}^{m'}(\rho, \theta, \epsilon) \rho d\rho d\theta = \delta_{nn'} \delta_{mm'} \tag{7}$$

the operations (e.g., inner products and norms) between functions expressed on this basis, get reduced to operations between their corresponding coefficients, which makes the computations easier. Thus, as it was explained at the beginning of this section, every analyzed image will be represented by the coefficients obtained from the linear combination of the annular Zernike polynomials, expressed in Eq.2. There is still one hyperparameter that needs to be established before doing the least square estimation of the coefficients, and it is the number of coefficients (or number of basis functions) in the expansion. This parameter has the role of a smoothing parameter. Statistically, keeping a few coefficients in the expansion is equivalent to conducting heavy amount of smoothing for the original data. It also determines the dimensionality reduction achieved with this representation. One way for determining this parameter is through the bootstrapping strategy [12].

[1] Note that $u = \rho^2$.

## 3 Experimental Results

To evaluate the performance of the proposed method, we used the CASIA-V2[13], UPOL[14], and MMU[15] databases. CASIA database version 2.0 consists of 1200 iris images from 60 different irises (subjects) with a resolution of 640 x 480 pixels. The MMU v1.0 database contains 450 iris images, which were collected from 45 subjects. There are 10 images from each subject. These are images of 320 x 240 pixels. This database has images with problems like specular reflections, off-axis and off-angle, blur, focus, non-uniform illumination, occlusions such as eyelids, eyelashes, glasses, contact lens, and hair.The UPOL database contains 384 images of 576 x 768, extracted from both eyes of 64 subjects (three images per eye). In our experiments, we used the VASIR method for the segmentation and normalization as we explained in Sec.1. Once we have images normalized, including the selection of the region avoiding the occlusions(explained in Sec.2), we represent them by using FDA, like is explained below: *a*) In order to obtain the optimal number of coefficients for the functional representation, we split the data sets in training (80% of the data) and test (20%). On each training set, a bootstrap method was then applied to find this optimal number. Results are then given by evaluating the test sets with the selected number of coefficients. It should be noted, that for the three cases of the selection of the region, all the resultant subregions have the same size. Taking this into account, the bootstrap process was applied to the UPOL database, and the optimal number obtained (48 coefficients) was used for all the databases.

*b*) The coefficients were estimated by the least squares fitting method on the annular Zernike polynomial basis using the normalized image. This process was repeated for the three different regions of interest selected with different affectations of occlusion.

*c*) The step *b*) was applied for every image in the databases, and then performed iris recognition by using an Euclidean distance between the coefficients. The recognition results were compared with Daugman[1] and Masek[16] representations. In Tables 1, 2 and 3, we show the results of the identification experiments with the images represented by the optimal number of coefficients (48) for each of the selected regions shown in Fig.2. In order to corroborate the importance of obtaining the optimal number of coefficients for a good representation, some experiments were carried out using 16 and 32 coefficients too. As it can be seen, the best results for the 3 databases, were obtained in Table 3

**Table 1.** Recognition results on CASIA V2, MMU V1 and UPOL databases, with region of interest represented in Fig.2(a)

| Database | 16 | 32 | 48 | Daugman | Masek |
|---|---|---|---|---|---|
| CASIA2 | 95.33 | 96.75 | 97.42 | 98.75 | 98.00 |
| MMU1 | 94.22 | 95.56 | 96.22 | 97.33 | 97.56 |
| UPOL | 97.22 | 98.96 | 99.74 | 100 | 100 |

**Table 2.** Recognition results on CASIA V2, MMU V1 and UPOL databases, with region of interest represented in Fig.2(b)

| **Database** | **16** | **32** | **48** | **Daugman** | **Masek** |
|---|---|---|---|---|---|
| MMU1 | 94.67 | 96.89 | 97.11 | 97.33 | 97.56 |
| CASIA2 | 96.83 | 98.25 | **98.75** | 98.75 | 98.00 |
| UPOL | 99.22 | 99.48 | 99.74 | 100 | 100 |

**Table 3.** Recognition results on CASIA V2, MMU V1 and UPOL databases, with region of interest represented in Fig.2(c)

| **Database** | **16** | 32 | 48 | **Daugman** | **Masek** |
|---|---|---|---|---|---|
| CASIA2 | 98 | **99.25** | **99.75** | 98.75 | 98.00 |
| MMU1 | 93.56 | 96.85 | 97.11 | 97.33 | 97.56 |
| UPOL | 99.22 | 99.48 | 99.74 | 100 | 100 |

(highlighted with black letter). In this case, the upper-right region of the image was used, according to the Fig.2(c), and then represented with the optimal number of coefficients. It should be noted that in this case our method overcomes the results from Daugman and Libor Masek in the Casia2 database(database with severe occlusions), while we achieved a dimensionality reduction of a 50% and 25% with respect to Daugman and Masek respectively. In the case of MMU1 and UPOL, although the results did not overcome the other methods, the achieved values of accuracy were high. The results from Table 2, representing the images that correspond to the region mentioned in Fig.2(b), are comparable to those of other methods, but it must still be highlighted the low dimensionality of our representation. Finally, analyzing the results from Table 1, it can be seen that occlusion factors affected slightly the accuracy results, but they are good overall. In general, it can be seen that our approach obtained good results with representations of very low dimensionality. It should also be noted that the best results were obtained with the representation constructed from the upper-right region of the iris image (Fig.2(c)), which is the less affected by occlusions. This shows somehow the sensitivity of the proposed representation to the occlusion problems. Therefore, the need of dealing with this problem (in this case by selecting the significative regions of the images) before obtaining the functional representation of the iris. In the 3 experiments, with the different selections of regions, the best accuracy was obtained with the optimal number of coefficients, mainly for those images where occlusion is present. This shows the importance of finding an optimal number of coefficients, for a good performance of our method.

## 4 Conclusions and Future Works

In this paper, we present a new functional iris representation based on annular Zernike polynomials. The main contributions of this work are: a high recognition accuracy in presence of severe occlusion problems while the dimensionality of the data is significantly reduced. It was demonstrated that with the proposed representation a high recognition rate can be obtained by just analyzing a small subregion of the iris image. For future work we should perform more experiments with other iris images databases to corroborate the performance of our method. Moreover, we are planning to extend this approach to other biometric images.

## References

[1] Daugman, J.: How iris recognition works. IEEE Trans. CSVT 14(1), 21–30 (2004)

[2] Jan, F., Usman, I., Agha, S.: Iris localization in frontal eye images for less constrained iris recognition systems. Digital Signal Processing 22, 971–986 (2012)

[3] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer, Heidelberg (2005)

[4] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall Upper Saddle River, New Jersey 07458 (2001)

[5] Lee, Y., Phillips, P.J., Micheals, R.J.: An automated video-based system for iris recognition. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1160–1169. Springer, Heidelberg (2009)

[6] Chaskar, U.M., Sutaone, M.S., Shah, N.S., Jaison, T.: Iris image quality assessment for biometric application. IJCSI International Journal of Computer Science Issues 9(3), 474–479 (2012)

[7] Vatsa, M., Singh, R., Noore, A.: Improving iris recognition performance using segmentation, quality enhancement, match score fusion, and indexing. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 38(4), 1021–1035 (2008)

[8] Lin, Z., Lu, B.: Iris recognition method based on the optimized gabor filters. In: 3rd International Congress on Image and Signal Processing (CISP 2010) (2010)

[9] Rossi, F., Delannay, N., Conan-Guez, B., Verleysen, M.: Representation of functional data in neural networks. Neurocomputing 64(2), 183–210 (2005)

[10] Boyd, J.P.: Chebyshev and Fourier Spectral Methods: Second Revised Edn. Springer (1989)

[11] Mahajan, V.N., Aftab, M.: Systematic comparison of the use of annular and zernike circle polynomials for annular wavefronts. Appl. Opt. 49(33), 6489–6501 (2010)

[12] Iskander, D.R., Collins, M.J., Davis, B.: Optimal modeling of corneal surfaces with zernike polynomials. IEEE Transactions on biomedical engineering 48, 87–95 (2001)

[13] Institute of Automation, Chinese Academy of Sciences, CASIA Iris Image Database (2004), http://www.sinobiometrics.com

[14] Dobes and Libor Machala.Upol iris image database (2004), http://phoenix.inf.upol.cz/iris

[15] Multimedia University, MMU Iris Image Database. (2004), http://pesona.mmu.edu.my/ccteo

[16] Masek, L.: Recognition of human iris pattern for biometric identification. Master's thesis, School of Computer Science and Software Engineering, The University of Western Australia (2003)

# Fusion of Facial Regions Using Color Information in a Forensic Scenario

Pedro Tome, Ruben Vera-Rodriguez, Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group-ATVS, EPS, Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{pedro.tome,ruben.vera,julian.fierrez,javier.ortega}@uam.es

**Abstract.** This paper reports an analysis of the benefits of using color information on a region-based face recognition system. Three different color spaces are analysed ($RGB$, $YC_bC_r$, $l\alpha\beta$) in a very challenging scenario matching good quality mugshot images against video surveillance images. This scenario is of special interest for forensics, where examiners carry out a comparison of two face images using the global information of the faces, but paying special attention to each individual facial region (eyes, nose, mouth, etc.). This work analyses the discriminative power of 15 facial regions comparing both the grayscale and color information. Results show a significant improvement of performance when fusing several regions of the face compared to just using the whole face image. A further improvement of performance is achieved when color information is considered.

**Keywords:** Face recognition, facial regions, forensics, color information, facial components, video surveillance, at a distance.

## 1 Introduction

Automatic face recognition systems are generally designed to match grayscale images of full faces. However, in practice, the full face is not always available, e.g., due to occlusions and other variability factors. On the other hand, in forensics, the examiners usually carry out a manual inspection of the color face images, focussing their attention not only on the grayscale full face but also on individual traits and color information. They carry out an exhaustive morphological comparison, analysing the face region by region (e.g., nose, mouth, eyebrows, etc.), even examining traits such as marks, moles, wrinkles, etc.

There are some previous works where grayscale facial region-based recognition is studied [1–3] but non of them focus their attention in the color regions normally considered by forensic experts. In this work, we have extracted facial components (called from now on facial regions) following forensic protocols from law enforcement laboratories, allowing us to study individually the different facial regions normally considered in current practice of forensic examiners. In particular, we address in this paper the problem of combining the most discriminative areas of the face for recognition using the available color information on a very challenging video surveillance scenario.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 399–406, 2013.

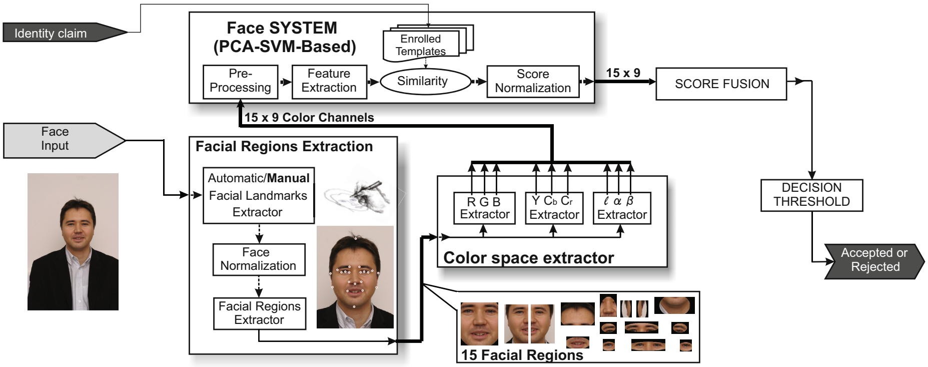


**Fig. 1.** Experimental framework diagram description

In contrast to traditional grayscale systems, this paper studies the discriminative power of each facial region using three color spaces: $RGB$, $YC_bC_r$, and $l\alpha\beta$. Fig. 1 summarizes the experimental framework followed.

Understanding how different facial regions from different color spaces are combined on a very challenging scenario has some remarkable benefits, for example: *i*) allowing investigators to work only with particular regions of the face in different color spaces, *ii*) improving the face recognition performance using all available information from color images, or *iii*) preventing that incomplete, noisy, and missing regions degrade the recognition performance. Further, a better understanding of the combination of facial regions in different color spaces should facilitate the study of facial regions-based face recognition. Therefore, the fusion of the different facial regions from different color spaces is performed achieving significant improvements of performance compared to a traditional face recognition system based only on the grayscale face as a whole.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of the automatic facial region extraction procedure and presents the color spaces analysed. Section 3 describes the experimental protocol followed, the database and the verification system adopted for the experiments. Section 4 reports an experimental fusion of different facial regions using different color spaces. Finally, Section 5 draws some conclusions of our work.

## 2 Facial Regions Extraction and Color Methodology

The proposed facial regions extraction framework is described in detail in [4] and extended in [3]. In this framework, two kinds of regions extraction are defined: *i*) based on human facial proportions, and *ii*) based on facial landmarks. For this work, the second extractor based on facial landmarks has been adopted. This extractor, based on facial landmarks manually located, allows to extract the facial regions with high precision. The final region extraction result is the set of 15 facial regions (see Table 1) based on forensic laboratories protocols[1] as shown in Fig. 2.

[1] Spanish Guardia Civil (DGGC), `http://www.guardiacivil.es/` and Netherlands Forensic Institute (NFI), `http://www.forensicinstitute.nl`

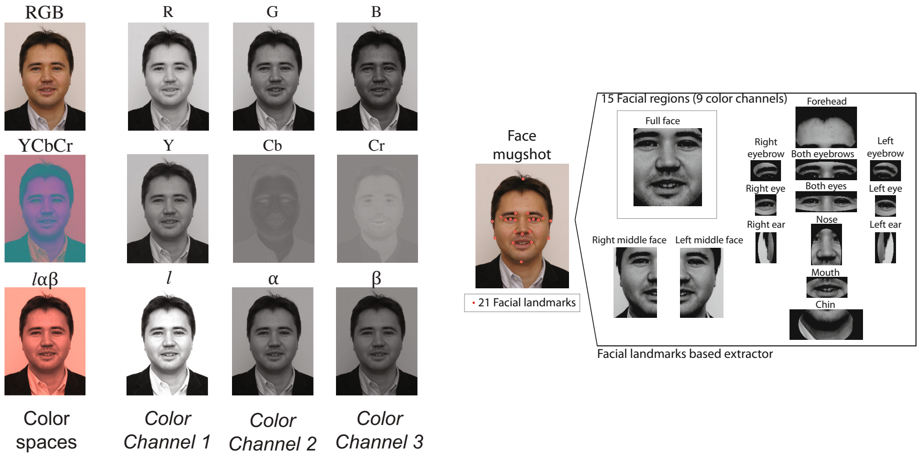


**Fig. 2.** (Left) Grayscale intensity values of faces for each color space analysed. (Right) Facial regions extraction based on facial landmarks extractor. The regions are extracted for the 9 color channels considered here.

**Table 1.** Facial regions id for each color channel and their sizes for extractor based on facial landmarks (height × width in pixels)

| Color Channel 1<br>Id Num. | Color Channel 2<br>Id Num. | Color Channel 3<br>Id Num. | Facial<br>Region | Facial Region<br>Size (h × w) |
|---|---|---|---|---|
| 1 | 16 | 31 | Chin | 75x181 |
| 2 | 17 | 32 | Left ear | 75x51 |
| 3 | 18 | 33 | Right ear | 75x51 |
| 4 | 19 | 34 | Left eyebrow | 51x75 |
| 5 | 20 | 35 | Right eyebrow | 51x75 |
| 6 | 21 | 36 | Both eyebrows | 51x151 |
| 7 | 22 | 37 | Left eye | 51x51 |
| 8 | 23 | 38 | Right eye | 51x51 |
| 9 | 24 | 39 | Both eyes | 51x151 |
| **10** | **25** | **40** | **Full face** | **192x168** |
| 11 | 26 | 41 | Forehead | 101x151 |
| 12 | 27 | 42 | Left middle face | 173x106 |
| 13 | 28 | 43 | Right middle face | 173x106 |
| 14 | 29 | 44 | Mouth | 51x101 |
| 15 | 30 | 45 | Nose | 101x75 |

There are some previous works where color spaces such as $RGB$ or $YC_bC_r$ have been used for face recognition [5, 6]. But, to the best of our knowledge, this is the first work where color information is used for face recognition using 15 facial regions.

When dealing with color images, the $RGB$ color space is commonly used. This color space is composed by three channels (red, green, and blue), which are correlated among them. The components that form the second color space considered $YC_bC_r$ are as follows: $Y$, luminance component, $C_b$, blue component ($B - Y$), and $C_r$, red component ($R - Y$) [7].

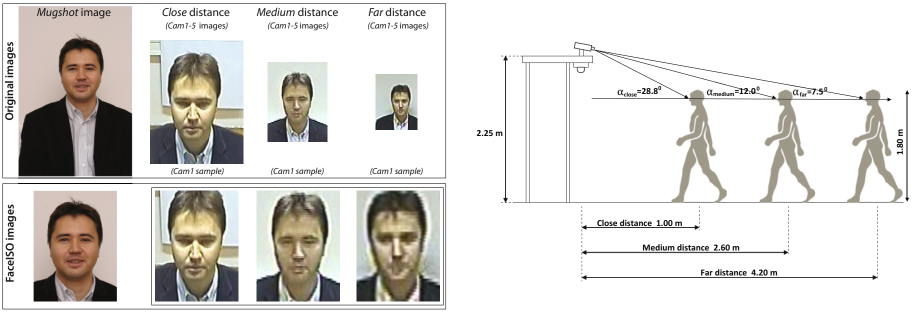


**Fig. 3.** (Left) SCface image samples of each datasets for *mugshot* and *Cam*1 images, and their corresponding normalized face for *close*, *medium*, and *far* distances. (Right) The three acquisitions distances: *close*, *medium* and *far*. Acquisition angle of each distance calculated for a subject with mean height of 1.80 meters.

Both $RGB$ and $YC_bC_r$ color spaces have correlated color channels among them. We also consider the $l\alpha\beta$ color space [8], which minimizes the perceptual correlation among the channels of an image. The parameter $l$ represents the luminance or brightness of the image and $\alpha$ and $\beta$ represent the chromatic content, i.e., the color information. Fig. 2 (left) shows an example of each color channel for these three color spaces considered in the experiments.

## 3 Experimental Protocol

Once each facial region has been extracted from each color channel, Principal Component Analysis (PCA) is computed obtaining eigen-regions. Then, similarity scores are computed in this PCA vector space (dimension 200, retaining an average of 98% of the energy of the original eigen-region space) using a Support Vector Machine (SVM) classifier with a linear kernel. The experimental protocol followed is described in more detail in [3].

The database used in our experiments SCface [9], (see Fig. 3 (left)), was divided into 3 subsets based on the subject ID: development (1-43), SVM training (44-87), and test (88-130). These three subsets were used for training the PCA features, as impostors in the training of SVMs, and for testing the final system performance, respectively. The procedure followed is summarized in Table 2.

Fig. 3 (left) shows an example of a mughost image, and the images acquired by one of the surveillance cameras. As can be seen there is a considerable scenario variation in terms of quality, pose and illumination. The change in the pose is specially important due to the different angles between the person and the cameras as shown in Fig. 3 (right). In this work a very challenge scenario of videosurveillance is studied considering a common case that a forensic examiner can find in practice: *mugshot vs CCTV* images. In addition, three distances between subject and camera typical in practical applications are analysed: *close*, *medium* and *far* distances (see Fig. 3 (right)).

**Table 2.** Partitioning of the SCface DB according to the *Mugshot vs CCTV images* protocol

| | **SCface DB (130 Subjects)** - Mugshot vs CCTV protocol | | |
|---|---|---|---|
| ***Subsets*** | 1...43 Subject<br>(43 Subjects) | 44...87 Subject<br>(44 Subjects) | 88...130 Subject<br>(43 Subjects) |
| *Mugshot* | Development set<br>(PCA subspace) | SVM Training<br>(Impostors) | SVM Training (Clients) |
| *Cam* 1<br>*Cam* 2<br>*Cam* 3<br>*Cam* 4<br>*Cam* 5 | | | Test<br>(Clients/Impostors) |

# 4 Facial Regions Fusion

This section describes the fusion of the 15 forensic facial regions extracted from a human face in comparison with the performance of the whole *face* region normally used in face recognition systems. The fusion is carried out at score–level combining the facial regions for the color channels considered here.

Before carrying out the fusion, scores of the different facial regions are first normalized to the $[0, 1]$ range using the tanh-estimators described in [10].

For this paper three different experiments were defined in order to analyse the potential of color information in a face recognition system: *i) Exp.1* Grayscale baseline system, where the grayscale facial regions are fused as the traditional face recognition systems. *ii) Exp.2* Fusion of color channels from each color space, (e.g. for $RGB$ color space, the channels $\{R, G, B\}$ are fused for each facial region considered). *iii) Exp.3* Fusion of all color channels, where all 9 available color channels are fused for each face region.

## 4.1 *Exp.1* Grayscale (Baseline System)

The fusion is carried out at the score–level for various combinations of grayscale regions. In particular, the 15 facial regions are fused using a parallel fusion approach based on the sum rule [11], starting from the most discriminative, then fusing this trait with the rest and keeping the best fusion of two regions, and continuing this process until all the regions are fused.

The fusion results obtained for the three distances are shown in Table 3 (*Exp.1*). As can be seen the system performance improves fusing several facial regions compared to just using the full *face* region.

*Close* and *medium* distance scenarios combine 7 facial regions to achieve the best result, but the *far* scenario needs to combine a total of 10 facial regions to obtain it. It is interesting to note that in the *close* scenario the best result is obtained with the fusion of inner and outer facial traits together with the full *face* (relative improvement of $56.7\%$ in the EER with respect to using only the full face).

Similarly, in the two other distances considered, the best fusion includes inner and outer parts of the face, and relative improvements of over $40\%$ in the EER are obtained with the fusion of regions compared to using only the full face.

**Table 3.** EER results for the score–level fusion obtained for sequential region fusion and the full face for the color channels of the three color spaces. In brackets is indicated the number of regions fused.

| | Color Space | *Close Scenario* | *Medium Scenario* | *Far Scenario* |
|---|---|---|---|---|
| | | Fusion (# Regions) – Full face | Fusion (# Regions) – Full face | Fusion (# Regions) – Full face |
| *Exp.1* | $Grayscale$ | 14.30% (7) – 33.10% | 12.90% (7) – 31.20% | 16.80% (10) – 28.90% |
| *Exp.2* | $RGB$ | 11.58% (12) – 32.19% | **10.79% (13)** – 30.21% | 14.61% (15) – 29.96% |
| | $YC_bC_r$ | 12.89% (16) – 29.50% | 12.65% (8) – 33.35% | 16.37% (21) – 31.72% |
| | $l\,\alpha\,\beta$ | **10.79% (12)** – 31.82% | 11.20% (16) – 31.09% | **14.50% (18)** – 28.93% |
| *Exp.3* | $ALL$ | **9.03% (27)** – 29.96% | **10.33% (22)** – 30.33% | **13.12% (39)** – 28.93% |

## 4.2 *Exp.2* Fusion of Three Color Channels

For the *Exp.2*, the score–level fusion is carried out fusing the three channels in a color space, i.e., $15\times3 = 45$ facial regions (as Table 1 shows) using a parallel fusion approach as in the previous experiment.

Table 3 (*Exp.2*) shows the fusion results for the three distances analysed. Fig. 4 shows the sequential fusion results obtained for the three distances and their corresponding color space with best performance ($l\alpha\beta$ for *close* and *far* distance, and $RGB$ for *medium* distance). Similar to the previous case the system performance improves fusing several facial regions compared to just using the *full face* region. It is interesting to note that the number of regions fused to obtain the best performance increases with the distance between the subject and the camera.

Comparing the fusion results with the baseline system based on grayscale facial regions, relative improvements of performance of 24.5%, 16.3%, and 13.7% for *close*, *medium* and *far* distance, are achieved respectively. These results support the utility of color information using facial regions to improve the performance of traditional face recognition systems.

## 4.3 *Exp.3* Fusion of All Color Channels

In this case, all facial regions from all color channels are combined following the same fusion methodology. In this case, we combine the 3 sets of 45 facial regions considered in the previous experiment, i.e., 135 facial regions in total.

Table 3 (*Exp.3*) shows the fusion results for this experiment. As can be seen this experiment achieves the best EER results for the three distances compared to the previous experiment. However this case needs to fuse more facial regions to achieve the best performance (approximately double than *Exp.2*), and just around 1% EER of improvement is achieved compared to *Exp.2*. Again, the increment of the acquisition distance increases the number of facial regions to be combined to achieve the best performance.

Similarly, in the three distances considered, the best fusion includes inner and outer parts of the face, and relative improvements of over 66% in the EER are obtained with the regions fusion compared to only using the *full face*.

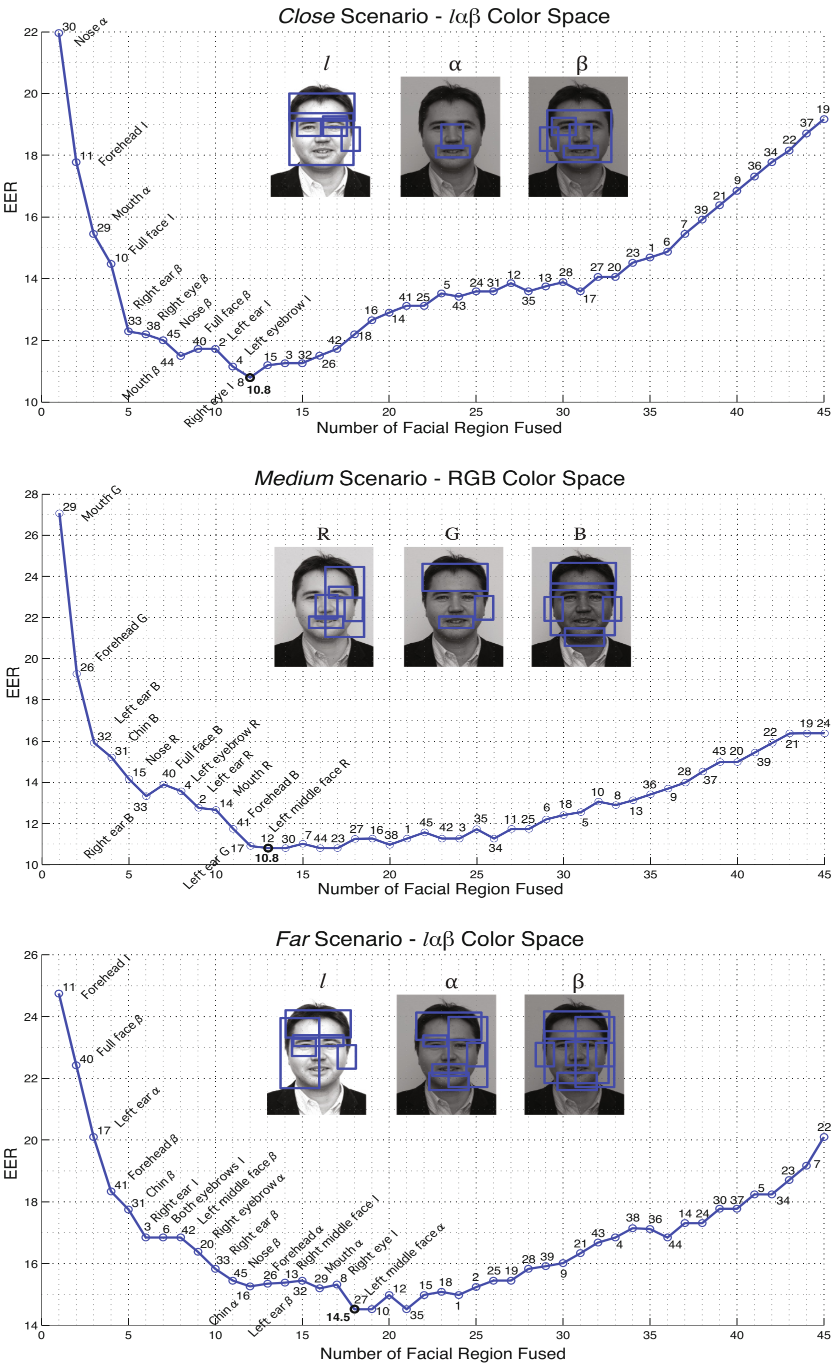


**Fig. 4.** EER for sum sequential fusion of the best combination of different facial regions for the best individual color space in each distance scenario: *close* ($l\alpha\beta$), *medium* ($RGB$) and *far* ($l\alpha\beta$)

## 5 Conclusions

This paper reports an study of the combination of 15 human facial regions extracted from three different color spaces on a very challenging scenario comparing mugshot versus CCTV images. The best fused performance of facial regions is compared with the *full face* region, which is the normal case in face recognition. Preliminary results show that a combination of a set of facial regions in different color spaces can significantly improve the system performance by a relative average improvement of over 66% for the three distances considered. The combination of facial regions with color information allows to improve the system performance with a relative improvement of over 20% comparing with the traditional face recognition systems using only grayscale information. The potential of fusion of facial regions on these scenarios has been demonstrated to significantly improve a traditional full face recognition system performance.

**Acknowledgements.** P. Tome is supported by a FPU Fellowship from Univ. Autonoma de Madrid. This work has been partially supported by contract with Spanish Guardia Civil and projects BBfor2 (FP7-ITN-238803), Bio-Challenge (TEC2009-11186), Bio-Shield (TEC2012-34881), Contexts (S2009/TIC-1485), TeraSense (CSD2008-00068) and “Cátedra UAM-Telefónica”.

## References

1. Ocegueda, O., Shah, S.K., Kakadiaris, I.A.: Which parts of the face give out your identity? In: IEEE Proccedings of CVPR, pp. 641–648 (2011)
2. Bonnen, K., Klare, B., Jain, A.K.: Component-based representation in automated face recognition. IEEE Transactions on Information Forensics and Security 8(1), 239–253 (2013)
3. Tome, P., Fierrez, J., Vera-Rodriguez, R., Ramos, D.: Identification using face regions: Application and assessment in forensic scenarios. Forensic Science International (submitted 2013)
4. Tome, P., Blazquez, L., Vera-Rodriguez, R., Fierrez, J., Ortega-Garcia, J., Exposito, N., Leston, P.: Understanding the discrimination power of facial regions in forensic casework. In: International Workshop on Biometrics and Forensics, Lisboa, Portugal (April 2013)
5. Singh, S.K., Chauhan, D.S., Vatsa, M., Singh, R.: A robust skin color based face detection algorithm, tamkang. Journal of Science and Engineering 6, 227–234 (2003)
6. Liu, Z., Liu, C.: Robust face recognition using color information. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 122–131. Springer, Heidelberg (2009)
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice-Hall, Inc., Upper Saddle River (2006)
8. Ruderman, D.L., Cronin, T.W., Chiao, C.C.: Statistics of coneresponses to natural images: implications for visual coding. Journal Optical Society of America 15, 2036–2045 (1998)
9. Grgic, M., Delac, K., Grgic, S.: Scface - surveillance cameras face database. Multimedia Tools Appl. 51(3), 863–879 (2011)
10. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition 38(12), 2270–2285 (2005)
11. Fierrez, J.: Adapted Fusion Schemes for Multimodal Biometric Authentication. PhD thesis, Univ. Politecnica de Madrid (2006)

# Facial Landmarks Detection Using Extended Profile LBP-Based Active Shape Models

Nelson Méndez, Leonardo Chang,
Yenisel Plasencia-Calaña, and Heydi Méndez-Vázquez

Advanced Technologies Application Center 7ma A ♯ 21406, Playa, Havana, Cuba
{nllanes,lchang,yplasencia,hmendez}@cenatav.co.cu

**Abstract.** The accurate localization of facial features is an important task for the face recognition process. One of the most used approaches to achieve this goal is the Active Shape Models (ASM) method and its different extensions. In this work, a new method is proposed for obtaining a Local Binary Patterns (LBP) based profile for representing the local appearance of landmark points of the shape model in ASM. The experimental evaluation, conducted on XM2VTS and BioID databases, shows the good performance of the proposal.

**Keywords:** facial landmarks, facial features detection, ASM, LBP.

## 1 Introduction

A fundamental step in the face recognition process is to model the face shape in an accurate manner; this allows one to find a correspondence of landmark points in face images for their posterior normalization or affine warping. However, the problem of accurately find these landmarks remains an open issue. The Active Shape Models (ASM) method [1] is one of the main approaches to automatically detect these points. ASM uses an appearance model to represent the image appearance around each landmark. In the original approach, this model is based on the so-called "gray-level profile", defined by the differences in the intensity values of adjacent pixels on a line centered at the landmark point. This is very sensitive to illumination changes and noise, not providing a sufficiently good description and discrimination of the local appearance.

Some extensions have been proposed in order to improve or replace the gray-level profile in ASM [2,3,4,5]. The Local Binary Patterns (LBP) operator [6] is a popular local appearance descriptor that can be considered for this task [5]. When using LBP, robustness to monotonic illumination changes and noise is achieved. LBP have been used in [4] and [5] for representing the local appearance in ASM, obtaining very good results in comparison with the original approach. In this work we propose a new method that uses LBP with ASM, aiming at better describing the local appearance of possible locations to which the landmark can shift. The rest of this paper is organized as follows. Section 2 describes the original ASM method and some ASM extensions related to this work. In

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 407–414, 2013.

Section 3, after a short introduction on LBP, the new proposal named EP-LBP ASM is presented, as well as some aspects related to its computation. Section 4 presents the experiments followed by concluding remarks in Section 5.

## 2 Active Shape Models (ASM)

ASM was first introduced by Cootes *et al.* [1] in 1995. It can be considered a deformable model that attempts to locate landmark points of known objects in images using a shape model, which describes the typical variations of the object shape, and a set of profile models that give a statistical representation of the image appearance around each point of the model.

Using ASM, the shape of an object in a 2D image is represented by a set of $n$ landmarks, concatenated into a single vector of dimension $2 \times n$. Then, given $M$ training samples, $M$ such vectors are generated and aligned to a common coordinate frame before performing the statistical analysis [1]. The aligned shapes can be then considered to form a distribution in a $2 \times n$ dimensional space, where the relationships between positions of every point can be modeled and learned. Finally, Principal Component Analysis (PCA) is used to approximate each shape in the training set, obtaining the so-called statistical shape model.

Given the statistical shape model, the ASM searches along profiles of each point the best match to the data in a new image. It is then necessary to have a model of the local appearance surrounding each landmark. In the original method proposed by Cootes *et al.* [1], the gray-level profile is represented by the normal to the shape boundary, passing through each landmark. This model is also learned from the training images. For searching and fitting the model in a new image, the learned mean shape is used as the initial shape. Then, each region is examined iteratively for searching the best shape and position parameters which best match the model to the image, until convergence is achieved.

### 2.1 ASM Extensions

There are several ASM extensions in the literature, most of them try to obtain a more discriminative and robust representation of landmarks appearance, which is also the aim of this paper.

The Combined-ASM method [2] is a combination of the original ASM gray-level profile with SIFT descriptor. Combined-ASM represents the local appearance of inner landmarks of the face using SIFT and uses the gray-level profile to represent the appearance of face border landmarks. This method provides a more discriminative description to inner points while maintains a description of face border that better describes edges. In Reg-ASM [3], regression is used to describe landmarks local appearance, in order to learn from false displacements in rectangular regions centered at landmarks. The use of regression allows one to infer causal relations, but may lead to false relations. Besides, it imposes a detailed labeling of true and false examples on the training dataset. A multi-resolution detector based on Multiple Kernel Learning (MKL) is proposed in [7], combining kernels from different resolutions in order to use more information. The Active Appearance Models (AAM) method [8] is another extension of ASM, which introduces a more detailed description of the appearance. However, ASM has shown to be more accurate than AAM for locating landmarks [9]. But there are some recent extensions of AAM, such as CLM [10], SOS [11] and TST [12] that outperform both AAM and ASM.

An extension of ASM based on the LBP descriptor was first proposed in [4]. The method, named ELBP-ASM, extracts different radius LBP descriptors [6] over circular subimages (gray scale image and gradient magnitude image) centered at each landmark. In order to retain spatial information, block-based LBP is used. Later, Marcel *et al.* [5] propose three LBP-based ASM approaches: profile-based LBP-ASM, square-based LBP-ASM and divided-square-based LBP-ASM. The profile-based LBP-ASM approach extracts LBP values from the normal profile of each landmark. Square-based LBP-ASM builds LBP histograms from a squared region centered at a given landmark. In divided-square-based LBP-ASM, the same square is used, but partitioned into four regions from which the LBP histograms are extracted and concatenated into a single feature histogram. The best results were achieved with the divided-square-based LBP-ASM [5].

In this work we propose a new Extended Profile LBP-based ASM (EP-LBP ASM) method, aiming to improve ASM landmark points localization. Although our proposal is also based on LBP, it differs substantially from previous approaches. In order to describe the appearance of not only the neighborhood of each landmark, but of the regions in the face image which define possible fittings in each iteration, we extract several LBP histograms of squared regions equally separated over profile normal.

## 3 ASM Using a New Extended Profile Based on LBP

In this section, we first briefly introduce the LBP operator. Next, we describe our proposal for modeling and fitting the shape model using a new Extended Profile based on LBP operator (EP-LBP).

### 3.1 Local Binary Patterns (LBP)

The LBP operator is a texture descriptor introduced in [6]. The operator and its different extensions have been widely applied in many computer vision applications, and particulary for face analysis. The original operator labels each pixel of an image by thresholding its $3 \times 3$ neighborhood with reference to its intensity value, and considering the result as a binary number. Since the operator only encodes the ordinal comparison (darker or brighter) between pixels intensities, it is considered to be invariant to monotonic illumination variations. On the other hand, when using this operator, the local appearance of an image is usually represented by histograms of the image regions, which makes the representation more robust to different kinds of noise in the image.

### 3.2 Landmarks Local Appearance Description with EP-LBP

In this paper we propose the EP-LBP as a more distinctive landmarks local appearance descriptor. Unlike other LBP-based approaches in literature (refer to Section 2.1) that describe circular or squared regions centered in the landmark point, we aim at describing those regions of the face image which define possible landmarks fittings in each iteration.

The profile $\varrho(p,k,d)$ of a landmark point $p$, is defined as the set of $k$ points over $p$ normal, centered at $p$ and separated by $d$ pixels. Figure 1 shows an example of the profile $\varrho(p,5,2)$, defined by a set of $k=5$ points in the line over $p$ normal, separated by $d=2$ pixels.

Once the profile is defined, for each point $\varrho_i \in \varrho(p,k,d)$, a LBP histogram, $HLBP_{\varrho_i,m}$, over a squared sized region of width $m$ and centered at point $\varrho_i$ is extracted. Then, the EP-LBP descriptor of a landmark point $p$, EP-LBP$_{(p)}$, is obtained by concatenating these $k$ histograms. In this work we use the original $LBP^{u2}_{(8,1)}$, so the number of LBP labels is 59 [6] and then the size of our EP-LBP descriptor for a given landmark is $59 \cdot k$ bins.

### 3.3 Shape Fitting Using EP-LBP

For shape fitting using the proposed EP-LBP descriptor, in each iteration, for every point $p$, a set $C_p = \{c_1, c_2, \ldots, c_N\}$ of $N$ candidate points of landmark point $p$ is determined. Every point $c_i \in C_p$ is on the line defined by $p$ normal and it is separated by $d$ pixels from $c_{i-1}$ and $c_{i+1}$. Point $p$ corresponds with candidate point $c_{\lceil N/2 \rceil}$. This is graphically explained in Figure 2.

The descriptor EP-LBP$_{(c_i)}$ associated to each candidate point $c_i \in C_p$ is calculated and the new point $p^*$, will be then the candidate point with the smallest distance between its EP-LBP descriptor and the trained EP-LBP for its corresponding landmark:

$$p^* = \arg\min_{c_i} \chi^2(\text{EP-LBP}_{(c_i)}, \overline{\text{EP-LBP}}_{(p)}), \tag{1}$$

where $\chi^2$ is the Chi-Squared histogram distance and $\overline{\text{EP-LBP}}_{(p)}$ is the mean EP-LBP for landmark point $p$ obtained in the training step.

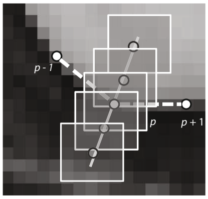


**Fig. 1.** Example of profile $\varrho(p,5,2)$

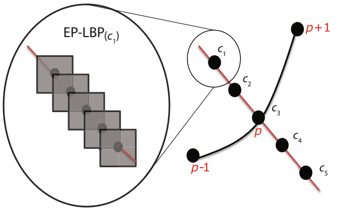


**Fig. 2.** Example of candidate points for landmark point $p$

### 3.4 EP-LBP Efficiency Improvements

In the fitting step, in each iteration, for every landmark point $p$, due to the intersection between $C_p$ and $\varrho(p, k, d)$, $O(N \cdot k)$ re-calculations of $HLBP_{\varrho i,m}$ histograms should be done when obtaining the EP-LBP descriptor for every candidate point in $C_p$. In order to avoid unnecessary processing we propose to compute a single EP-LBP$_{(p)}$ over a profile $\varrho(p, N + k - 1, d)$. Then, the process of selecting the new point $p^*$ is simplified to selecting the sub-EP-LBP that best match the extracted EP-LBP$_{(p)}$. The previously mentioned improvement reduces the time complexity of this process from $O(m \cdot N \cdot k)$ to $O(m \cdot (N + k))$.

In practice, we have realized that when dealing with face frontal images and faces with little expression variations, the algorithm converges in the first ten iterations. Based on that fact, we propose another stop criteria for the fitting stage. The number of points which displacement in the current iteration was less than 2 pixels is determined and if this number is greater than the 95% of the total landmarks, the process is stopped.

## 4 Experimental Evaluation

In this section two experiments are described. First, we determine the best parameters for our proposal and then compare it with some existing approaches. Two standard face databases were used for this purpose: the BioID and the XM2VTS. The BioID Face Database (http://www.bioid.com) consists of 1,521 gray level images from 23 different persons, captured with large variations in expression, illumination, background and face size. The XM2VTS [13] contains 2,360 images captured during four recordings of 295 subjects over a period of four months, with different variations in expression, occlusions and appearance. In both cases the manual annotations of landmark points are given, 20 landmarks in the case of BioID and 68 for images in the standard sets of XM2VTS.

### 4.1 Model Parameters

To obtain the model parameters we used BioID, as it contains large variability in illumination and facial expressions. The images were randomly divided into two equal parts, one for training and the other for testing. The three different parameters were changed in the following way: $k = \{5, 7, 9, 11, 13\}, m = \{3, 5, 7, 17, 31\}$, and $d = \{1, 2, 3, 4\}$; all combinations of these values were tested.

To measure the quality of the fit we use the mean error in landmarks localization compared to ground-truth, given by:

$$m_{e_P} = \frac{1}{P * d_{eyes}} \sum_{j=1}^{P} d_j, \tag{2}$$

where $P$ represents the number of points in the model, $d_{eyes}$ is the distance between labeled eyes, and $d_j$ is the distance between every detected point and its corresponding ground-truth position.

Figure 3 shows the results obtained for each parameter in all conducted experiments. When analyzing the graphics we found that the size of the region used, $m$, and the separation, $d$, are not as important as the number of points considered in the profile, $k$. Therefore, when computational time is a critical issue, minimum values for these parameters ($m = 3$ and $d = 1$) can be selected. Nevertheless, slightly better results were obtained for $m = 7$, and $d = 2$, so we used these values on the rest of our experiments.

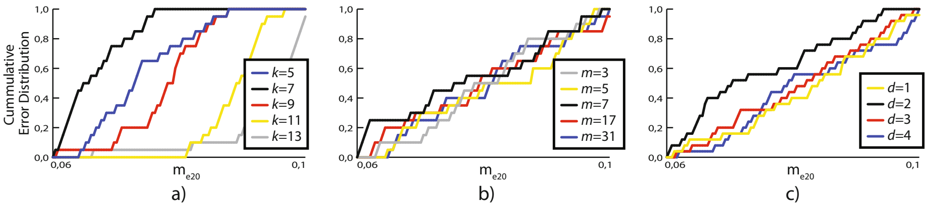


**Fig. 3.** Results for different values of EP-LBP ASM parameters: a) $k$, b) $m$, and c) $d$

## 4.2 Facial Landmarks Detection Accuracy and Timing

First, we compare our proposal with other existing ASM extensions based on LBP [4,5] and with the eyes detector used in [5]. This experiment is conducted on XM2VTS database, using the configuration I of Laussane Protocol [13]. Under this configuration, the shape model is trained with the parameters obtained in Section 4.1 using 3 images of 200 clients. The evaluation is performed on the standard test set defined in the protocol, as well as on the darkened set, where the robustness to illumination changes is evaluated. For the standard test set, the Mean Square Errors of all 68 landmarks is computed. In the case of the darkened set, similar to [5], we used the Jesorsky's measure for evaluation, that only takes into account the error in the center of the eyes points, since the annotations for all landmarks are not provided. The obtained results are shown in Table 1. As it can be seen in the table, our method obtained the best results in both, the standard and darkened sets of the XM2VTS database.

In order to compare the proposed EP-LBP ASM method with previously reported results of other ASM and AAM extensions described on Section 2.1, we used only 17 of the 68 landmarks to compute $m_{e_P}$ following Equation 2. These landmarks correspond to the inner regions of the face: eyes, nose, eyebrows and

**Table 1.** Mean and median values of Mean Square Errors for the standard test set of XM2VTS, and Jesorsky's error measure for the darkened set

| | Mean square error for the standard test set | | | | | | Jesorsky's measure (error) for the darkened set | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASM | ELBP | Profile | Square | Divided | EP-LBP | detector | ASM | Profile | Square | Divided | EP-LBP |
| mean | 73.0 | 41.0 | 71.0 | 61.0 | 46.0 | 21.0 | 0.11 | 0.11 | 0.11 | 0.09 | 0.10 | 0.06 |
| median | 32.0 | 31.0 | 43.0 | 41.0 | 28.0 | 18.0 | 0.10 | 0.09 | 0.09 | 0.08 | 0.07 | 0.05 |

mouth, which are actually the most difficult for fitting [10]. In Figure 4 a) the results are shown for our proposed EP-LBP ASM method compared to original ASM [1], Reg-ASM [3], CLM [10], SOS [11], and TST [12]. It can be seen that our method outperforms all the other methods.

Most of the above methods have been also tested on BioID database. In this case, similar to [10], we used the 20% of the images for training and the rest of them for testing, and the $m_{e_P}$ with the 17 inner points is computed. The obtained results for our method compared to ASM [1], Reg-ASM [3], CLM [10] and MKL-based method [7], are shown on Figure 4 b). It can be seen that also in this database the results achieved by the proposed EP-LBP are better than or comparable to the other methods.

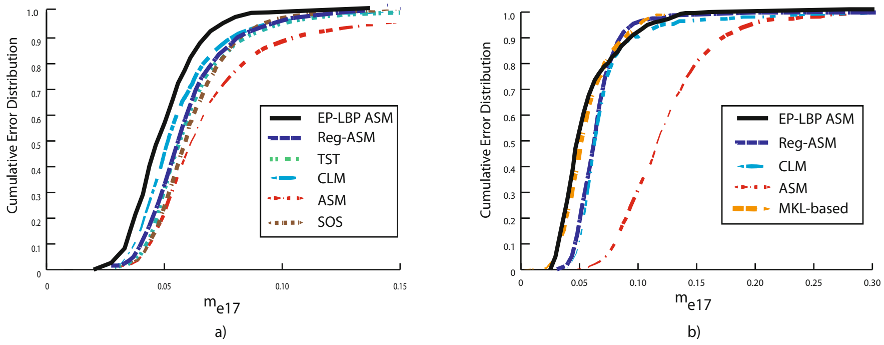


**Fig. 4.** Results in terms of $m_{e_{17}}$ on a) the standard test set of XM2VTS database and b) the BioID database

In the case of the Combined-ASM method [2], the results on BioID database are reported in terms of the Jesorsky's measure. They reported that about the 86% of the images have less than 0.10 of error, and about the 90% of them an error less than 0.15. Using this measure we have obtained a 92% of the images with less than 0.10 of error, and 98% with an error less than 0.15.

In terms of computation time, our approach, on average, needs 320 *ms* and the ASM needs 97 *ms* to detect landmarks in an image. This is comparable to other extensions [5] that use more information than the original ASM, which are also 2 or 3 times slower than ASM. The analysis was performed on a 2.5 GHz with 4 GB of RAM computer.

## 5 Conclusion and Future Work

In this paper a new Extended Profile based on LBP operator was introduced for representing the appearance of landmarks regions in the ASM method. The proposal, unlike other LBP-based approaches in the literature that describe circular or squared regions centered in the landmark point, describes regions of the face image which better define possible shape fittings in each iteration. Experimental results on two well-known databases frequently used for this purpose, showed the good performance of the method and its superiority with respect to other state-of-the-art ASM and AAM extensions.

It should be noticed that EP-LBP ASM and the other methods described in Section 2.1, extend classical ASM by proposing better landmark appearance representations. However, there are other approaches (e.g. [14]) that improve ASM by enhancing the shape constraint, i.e. the correlation between landmarks. Our future work will focus on extending our method by exploiting this idea.

## References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models - their Training and Application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
2. Zhou, D., Petrovska-Delacrétaz, D., Dorizzi, B.: Automatic landmark location with a Combined Active Shape Model. In: 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems. BTAS 2009 (2009)
3. Cristinacce, D., Cootes, T.F.: Boosted Regression Active Shape Models. In: British Machine Vision Conference, pp. 79.1–79.10. BMVA Press (2007)
4. Huang, X., Li, S.Z., Wang, Y.: Shape localization based on statistical method using extended Local Binary Pattern. In: Third International Conference on Image and Graphics. ICIG 2004, IEEE Computer Society, USA (2004)
5. Marcel, S., Keomany, J., Rodriguez, Y.: Robust-to-illumination face localisation using Active Shape Models and Local Binary Patterns. Tech Report, Idiap-RR-47-2006, IDIAP (2006)
6. 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)
7. Rapp, V., Senechal, T., Bailly, K., Prevost, L.: Multiple kernel learning svm and statistical validation for facial landmark detection. In: FG, pp. 265–271. IEEE (2011)
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 681–685 (2001)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Comparing active shape models with active appearance models. In: British Machine Vision Conference (1999)
10. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recogn. 41(10), 3054–3067 (2008)
11. Cristinacce, D., Cootes, T.F.: A comparison of shape constrained facial feature detectors. In: 6th International Conference on Automatic Face and Gesture Recognition 2004, Seoul, Korea, pp. 375–380 (2004)
12. Cristinacce, D., Cootes, T.F.: Facial feature detection and tracking with automatic template selection. In: FG, pp. 429–434. IEEE Computer Society (2006)
13. Messer, K., Matas, J., Kittler, J., Jonsson, K.: XM2VTSDB: The extended M2VTS database. In: Second International Conference on Audio and Video-based Biometric Person Authentication, pp. 72–77 (1999)
14. Sun, J., Wei, Y., Wen, F., Cao, X.: Face alignment by Explicit Shape Regression. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 2887–2894 (2012)

# Relative Spatial Weighting of Features for Localizing Parts of Faces

Jacopo Bellati and Díbio Leandro Borges

Department of Computer Science, University of Brasilia,
70910-900 Brasília, DF, Brazil
bellati@gmail.com, dibio@unb.br

**Abstract.** This paper proposes an approach for detecting important parts of faces in uncontrolled imaging settings. Regions of special interest in faces of humans are eyes and eyebrows, nose and mouth. The approach works by first extracting ORB (Oriented FAST and Rotated BRIEF) and SURF (Speeded up robust features) features, secondly a supervised learning step with a random subset of images is performed using k-means algorithm for devising the clusters' centers of the important parts of faces. For the testing set of images the normalized values of each new ORB or SURF feature is weighted positively depending on its similarity and proximity of a cluster center (a face part). Tests were performed using the BioID dataset which consists of 1521 images of 23 different subjects in a variety of situations. Results show that the use of ORB features for face parts localization is more efficient and more precise than SIFT or SURF features alone. Also, the relative spatial weighting of a combination of ORB and SURF features enhances the localization of parts of faces.

**Keywords:** Face parts localization, ORB features, face detection.

## 1 Introduction

Face detection and related applications have been at the top of approached problems by the Computer Vision and Pattern Recognition research community. A step forward to be taken is to have more detailed localization of face parts, or facial features as they are also called, since identity, sorting and editing face applications are dependent on dealing with the face parts in separate. 2D face parts can be considered as facial feature points, as specifying for example center of eye, tip of nose, mouth corners, as in [2] [6] [5], or as large scale 2D facial parts such as eyes and eyebrows, nose and mouth, as in [7] [9]. In this paper we approach the 2D face parts localization problem considering four large scale facial parts: eye and eyebrow right, eye and eyebrow left, nose, and mouth. Regarding a recent taxonomy of facial features proposed in [11] our work explores level 2 features, which are locally derived and can describe structures relevant to face recognition.

Most representative of face fiducial points works are [2], [6], and [5]. In [2] they formulate the problem of part localization as Bayesian inference combining local detectors and a prior model of face shape. They use a large collection of exemplars, and

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 415–422, 2013.

the part locations are decided on a consensus (RANSAC based) decision to disambiguate candidates. They present tests on the BioID database and a proprietary one. [6] presented a method to detect face fiducial points based on regression forests. 13000 images annotated with 10 fiducial points are trained and the ensembles of regression trees estimate the positions of the fiducial points. [5] presented a facial landmark (fiducial points) localization method based on Haar features and gradient boosted trees to predict the landmark positions, 9 landmark positions are previously defined and results are shown on subsets of BioID and their own image database.

More related to the approach proposed here are [9] and [7]. [9] investigated detection of large scale facial features by using an appearance based feature vector of gaussian derivatives of normalized face images. The facial features are defined as salient from the face images. Results were presented with 30 images detected eyes, nose, mouth and chin as important facial features, and were very dependent on scale, and invariance was not considered. [7] proposes to detail and detect facial features (eyes, nose, eyebrows, mouth, chin) by constructing appearance vectors of the features, and also of the context surrounding the features. A supervised learning discriminative algorithm is then applied to classify features and non-features samples. Results were shown for 1200 face images with uniform and controlled background with error rates below 5%. Detection of facial features on varying lighting and background conditions were not shown.

In this paper we propose a relative spatial weighting algorithm for localization of face parts. One of our motivations is to explore feature detector such as ORB [15] and SURF for face parts localization, and to investigate facial feature detection for face identification applications. Our main contributions in this work are: 1) provide a learning algorithm to select ORB and SURF features for face parts localization; 2) results on a benchmark image database for face parts localization.

## 2 Invariant Feature Descriptors

Finding correspondences between different images of the same object, considering a variety of lighting, viewing and scaling conditions is a major task in Computer Vision and applications. A large variety of feature descriptors and their respective matching algorithms have been proposed in the literature. One of the most successful is the SIFT descriptor (Scale Invariant Feature Transform)[13], although more recently promising invariant feature descriptors have been presented such as SURF (Speeded Up Robust Features)[1], and ORB (Oriented FAST and Rotated BRIEF) [15].

### 2.1 SIFT Descriptor

SIFT is a highly distinctive, scale and rotation invariant descriptor. It is computed in four main steps [13]: 1) Extract the keypoints from the image as local extrema (minima or maxima) using a Difference of Gaussians (DoG) Pyramid. Each point is compared to its 8 neighbors in the same scale, as well as to its 9 neighbors in the upper and lower scale; 2) Localize the keypoints, position and scale, by fitting a quadratic polynomial and rejecting weak keypoints by a Hessian matrix curvature test. 3) Surrounding a keypoint compute a histogram of gradient directions , and then assign the canonical orientation (single or multiple) of the patch as the peak(s) of the smoothed histogram; 4) The keypoint descriptor is formed by a 128 vector of 16 histograms with 8 orientations, considered in 16x16 windows with keypoint at center. Matching can be performed by comparing two descriptors with a distance function.

### 2.2 SURF Descriptor

SURF has been devised to be a faster and more robust descriptor, and matcher feature algorithm, than SIFT [1]. First it approximates the derivatives of Hessian matrix by box filters and uses the integral image as basis for computations. The determinant of H is also used for keypoint localization which is weighted to obtain a good approximation. Orientation assignment is done by evaluating a circular neighborhood around the keypoint and computing haar horizontal and vertical responses using the integral image also as basis. The SURF descriptor considers square regions and sum the responses (vertical and horizontal) for each subregion separately. A vector of 128 elements of those sums for keypoint regions is formed as the descriptor. It has been reported [15] that SURF is one order of magnitude faster than SIFT, however less robust to viewpoint and illumination changes.

### 2.3 ORB Descriptor

ORB descriptor [15], Oriented FAST and Rotated BRIEF, builds on good properties of FAST and BRIEF descriptors. Two main innovations are made on them, first it adds to a FAST descriptor an orientation computation by a weighted averaging of pixel intensities in the local patch. This centroid operator gives a single dominant orientation. Second it uses an ID3 machine learning algorithm for de-correlating BRIEF features under rotational invariance, and this is used for sampling point pairs to the descriptor. ORB is a binary descriptor, aimed to be an efficient alternative to SIFT or SURF descriptors [15]. Matching can be computed by Hamming distance. ORB has been reported [15] to be about 10 times faster than SURF, 100 times faster than SIFT, and less sensitive to gaussian noise than SIFT.

## 3 Face Parts Description and Localization

Face detection and face recognition are tasks with great interest from the research community. One of the most important subtasks of it is the identification and localization of important face parts, or facial features, as eyes, eyebrows, nose and mouth. Many applications besides recognition of individuals, such as autofocus, white balancing, sorting and retrieving face images, semi-automatic editing, depend on the localization of the face parts.

There are two basic different approaches for the localization of face parts: one that considers the facial feature points as relevant elements to be identified [2] [5] [6], a variation from 5 to 30 points have been reported in the literature as those facial feature points, for example eyes and mouth corners, centers, and middle points of contours and nose; and another that considers large-scale facial parts or the whole region of interest of eyes, nose and mouth as a facial part [7] [9]. The approach considered in this paper is the second one, where the face parts are four main regions: left eye and respective eyebrow, right eye and respective eyebrow, nose, and mouth. Figure.1 shows an image from the BioID database used as a training image with these four facial parts marked on it.

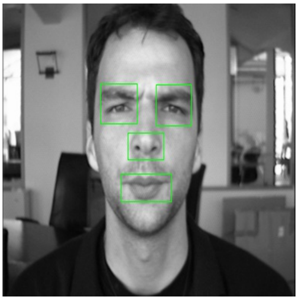


**Fig. 1.** One training image from the BioID database with the four proposed face parts shown

State of the art feature detectors such as SIFT [13], SURF [1], and more recently ORB [15] have been applied and demonstrated impressive performance on challenging recognition and tracking tasks [14]. However, only in the case of SIFT [12] [3], and SURF [8] there had been some attempts to address face identification and recognition using it as descriptor. To the best of our knowledge ORB has not been applied on the mentioned problems here. It is a hypothesis of this work that face parts localization can be addressed by one, or a combination of these feature detectors, especially the ORB since it shows top properties [15] a descriptor is aimed to demonstrate. Figure.2 shows a typical frontal face, from a benchmark image database for face identification BioID [4], with marked the 50 most salient feature points output by (a) ORB, (b) SIFT, and (c ) SURF detectors.

Since one aim is to localize face parts from typical images considering illumination variations, relevant backgrounds, and face variations by expressions such as talking, smiling, closing eyes, wearing glasses, it can be shown that these three feature detectors have different responses from each other.

This work proposes to learn from a set of images, a supervised training set of the closest features (ORB and SURF) to the face parts, and then classify other images for face parts by devising a k-means with relative weighting for the trained centroids of the parts. Initial tests had shown that SIFT, besides being the much slower detector, picked the farthest from the face parts aimed. It is a hypothesis from this work that ORB and SURF features can be used for an efficient face parts localization. A novel algorithm for performing such selection and localization is presented next.

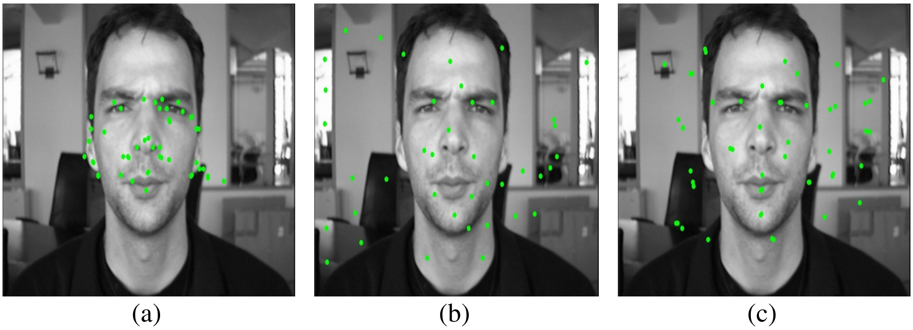


**Fig. 2.** Example images from the BioID database with (a) 50 highest ORB features shown; (b) 50 highest SIFT features shown; (c) 50 highest SURF features shown

## 4 Relative Spatial Weighting of Features

We propose to train and select a subset of features (ORB and SURF) in order to have the large scale face parts being localized and sampled for sure. For this we devise the following algorithm which positively weights the closest features to the facial parts, and negatively the farthest. First, it selects a subset of images for training, label the most salient normalized features (up to 30 ORB and SURF) belonging to the face parts; Average those feature outputs to each face part and keep its statistics (mean values and relative distances of the centroids); For a new image, compute and normalize the 30 most salient features (ORB and SURF); then for each feature, from the most to the least salient, run a k-means having as seeds the trained face parts; if a feature is decided close and similar enough (e.g. thresholded by σ deviation of the trained sets) to the face part it is selected and added for new statistics, if not it is discarded; select only up to 20 features for each image.

## 5 Results and Evaluation

BioID [4] provides a free database of face images widely used to benchmark face identification and recognition algorithms. It has 1521 images with 23 different subjects in a variety of conditions such as talking, smiling, illumination changes, and small rotations. For the evaluation it was manually partitioned by us in 7 categories regarding frontal and rotation, open and closed eyes, smiling and neutral, wearing glasses. A set of 30 images was randomly picked from all the sets and were used as a training set. Results were averaged for 10 different rounds of images for training. The category sets are not uniform since the purpose of the manual partition was to analyze the particular variations on the database. Table.1 gives the exact numbers of images in the partitions, as well as the total features detected in the face parts and their average by image.

In Table.1 it can be seen that in all image partition sets the algorithm proposed detected 6 (30%, 6 out of 20 maximum) to 10 (50%, 10 out of 20 maximum), with average of 7 (35%, 7 out of 20 maximum) facial features, or face parts. As the images from the database show relevant backgrounds, and the faces are under variations of lighting and expressions, by selecting 30% to 50% of relevant points in the face parts is a successful achievement for the task since there are only 4 face parts (eye left, eye right, nose, mouth) considered, and no training on templates or appearance models was done, but a fast feature selection and relative weighting on clustering was proposed. Similarly testes were also performed varying the size of the training set for randomly 10% (152) of the images, 30% (456) and 50% (760), and on average the number of detected features per image in the face parts (out of 20 maximum) was respectively 7, 7, and 7, keeping the relevance around 35% as shown in Table.1

**Table 1.** Results (Cross-validated) showing the number of facial features in sets of images from the BioID database. A random set of images (10%, 30% and 50%) was separated for training,. The maximum number of features that could be selected is 20.

| **Partition set** | **Number of images used for testing** | | | **Total number of detected features in the face parts (ORB + SURF)** | | | **Average number of detected features per image in the face parts (out of 20 maximum)** | | |
|---|---|---|---|---|---|---|---|---|---|
| Number of images used for training | 152<br>10% | 456<br>30% | 760<br>50% | 152<br>10% | 456<br>30% | 760<br>50% | 152<br>10% | 456<br>30% | 760<br>50% |
| expressions1 | 17 | 11 | 8 | 109 | 60 | 36 | 6 | 5 | 5 |
| expressions2 | 59 | 47 | 34 | 541 | 454 | 327 | 9 | 10 | 10 |
| expressions3 | 196 | 156 | 105 | 1199 | 949 | 627 | 6 | 6 | 6 |
| expressions4 | 432 | 342 | 247 | 3509 | 2752 | 1976 | 8 | 8 | 8 |
| normal1 | 355 | 279 | 200 | 2619 | 2065 | 1507 | 7 | 7 | 8 |
| normal2 | 234 | 174 | 118 | 1303 | 1003 | 684 | 6 | 6 | 6 |
| eyes1 | 76 | 56 | 49 | 615 | 456 | 347 | 8 | 8 | 7 |
| **Total** | **1369** | **1065** | **761** | **9895** | **7739** | **5561** | **7** | **7** | **7** |

Figure.3 shows some output images from the algorithm proposed. The images are taken from different partition sets and they show the selected (out of 20 maximum) feature points (mixed ORB and SURF) and the marked for reference face parts. Only points in the face parts would be 100% success. The variations on illumination, face rotation and expressions are illustrated. It can be seen that the selected features are concentrated on the faces mostly, and we know that these feature detectors would respond strongly to salient regions in the background. However, the concentration on the face, and especially on the face parts is the result of the proposed algorithm which positively weights points in the face parts and negatively outside. The results on the BioID database confirms that the proposed algorithm selects a subset of features from ORB and SURF localized mostly in the face, and in the face parts. Also, it has been shown that ORB detector responds much better to face features than the others SIFT and SURF, and it can be further explored for identity recognition as well.

The preference for the ORB and SURF features for this work were twofold: first, they were not explored fully yet for face parts, or face identification, as the SIFT detector [3] [12]; second their properties of robustness to noise, low computational complexity, and localization [15] [10] would favor their use instead of SIFT. Although on average the contribution of SURF features in the selected set of points in the face parts are around 5% for this database, their computation is one order of magnitude faster than SIFT [15], and SURF features would possibly bring more robustness to scale variations than ORB for different (although not tested here) databases. The presence of SURF features is constant and important on the final selected features. Matching would be done for ORB and SURF features as originally proposed [1] [15], although the results presented here open the path for exploring multiple face parts localization and hybrid matching schemes for binary and real valued features possibly.

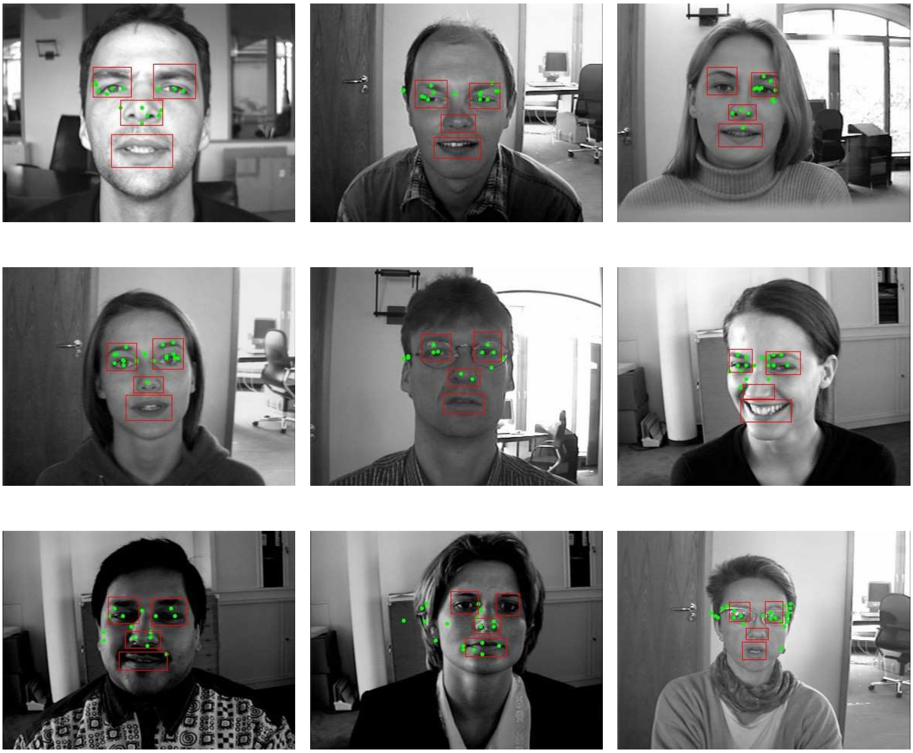


**Fig. 3.** Output images showing the selected features and their positions relative to the face parts

## 6 Conclusion and Future Work

In this paper we have proposed a new algorithm to train and select a subset of ORB and SURF features for face parts localization. It has been shown that the selected features from the algorithm concentrates on the face, and the facial features as eyes, eyebrows, nose and mouth. The selected features represent 40% on average the regions detected on BioID, a benchmark free image database. At least 30% of the points detected were in the face parts considering the most challenging partition set. ORB features have been demonstrated here to be well suited for face identification, and face parts localization. Results are interesting to investigate further the use of combined ORB and SURF features for multiple face parts identification and recognition for close to real-time applications since those features are one order of magnitude faster than state of the art feature detector as SIFT.

**Acknowledgments.** This work was partially supported by FAPDF and CIC/UnB.

## References

1. Bay, H., et al.: SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU) 110(3), 346–359 (2008)
2. Belhumeur, P., et al.: Localizing Parts of Faces Using a Consensus of Exemplars. In: Proceedings of IEEE CVPR 2011, pp. 545–552 (2011)
3. Bicego, M., et al.: On the use of SIFT features for face authentication. In: Proceedings of CVPRW 2006 (2006)
4. BioID Face Database, `http://www.bioid.com/` (visited in June 2013)
5. Chevallier, L., et al.: Facial Landmarks Localization Estimation by Cascaded Boosted Regression. In: Proceedings of VISAPP 2013 (2013)
6. Dantone, M., et al.: Real-time facial feature detection using conditional regression forests. In: Proceedings of IEEE CVPR 2012, pp. 2578–2585 (2012)
7. Ding, L., Martinez, A.M.: Features versus Context. IEEE Trans. PAMI 32(11), 2022–2038 (2010)
8. Dreuw, P., et al.: SURF-Face: Face Recognition under Viewpoint Consistency Constraints. In: Proceedings of BMVC 2009 (2009)
9. Gourier, N., Hall, D., Crowley, J.: Facial Features Detection Robust to Pose, Illumination and Identity. In: Proceedings of IEEE SMC 2004, pp. 617–622 (2004)
10. Heinly, J., Dunn, E., Frahm, J.-M.: Comparative Evaluation of Binary Features. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 759–773. Springer, Heidelberg (2012)
11. Klare, B., Jain, A.K.: On a taxonomy of facial features. In: Proceedings IEEE BTAS 2010, pp. 1–8 (2010)
12. Križaj, J., Štruc, V., Pavešić, N.: Adaptation of SIFT features for robust face recognition. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010, Part I. LNCS, vol. 6111, pp. 394–404. Springer, Heidelberg (2010)
13. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, vol 60(2), 91–110 (2004)
14. Miksik, O., Mikolajczyk, K.: Evaluation of Local Detectors and Descriptors for Fast Feature Matching. In: Proceedings of ICPR 2012, pp. 2681–2684 (2012)
15. Rublee, E., et al.: ORB: an efficient alternative to SIFT or SURF. In: Proceedings of IEEE ICCV 2011, pp. 2564–2571 (2011)

# SDALF+C: Augmenting the SDALF Descriptor by Relation-Based Information for Multi-shot Re-identification

Sylvie Jasmine Poletti[1], Vittorio Murino[2,1], and Marco Cristani[1,2]

[1] Department of Computer Science, University of Verona (IT)
[2] Pattern Analysis and Computer Vision Dept., Istituto Italiano di Tecnologia (IT)

**Abstract.** We present a novel multi-shot re-identification method, that merges together two different pattern recognition paradigms for describing objects: feature-based and relation-based. The former aims at encoding visual properties that characterize the object per se. The latter gives a relational description of the object considering how the visual properties are interdependent. The method considers SDALF as feature-based description: SDALF segregates salient body parts, exploiting symmetry and asymmetry principles. Afterwards, the parts are described by color, texture and region-based features. As relation-based description we consider the covariance of features, recently employed for re-identification: in practice, the parts found by SDALF are additionally encoded as covariance matrices, capturing structural properties otherwise missed. The resulting descriptor, dubbed SDALF+C, is superior to SDALF by about 2% and to the covariance-based description by a 53%, both in terms of average rank1 probability, considering 5 different multi-shot benchmark datasets (i-LIDS, ETHZ1,2,3 and CAVIAR4REID).

**Keywords:** re-identification, SDALF, covariance of features.

## 1 Introduction

People re-identification (re-id) has definitely become a primary module for the multi-camera video surveillance systems, allowing to recognize individuals across different locations and times. The re-id literature can be partitioned in different ways: *direct* vs. *learning-based*, and *single-shot* vs. *multi-shot* methods. Direct approaches [2,1,3] are on-line feature extractors, while learning-based techniques [11,7,12,10,4] require a training phase prior to work. Single-shot [2,11,7,10,4] and multi-shot [2,12,1,3] approaches differ for the number of images exploited to describe each probe or gallery subject: multi-shot strategies employ several shots (images) for building an individual signature.

In this paper, we present an approach for direct, multi-shot re-identification, that aims at joining two different ways to represent objects, employing *feature-based* and *relation-based* descriptions. Features serve to encode the tangible aspects of an entity, while relation-based descriptions explain how these aspects

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 423–430, 2013.

are inter-related. Both approaches have their pros and cons. Features are intuitive to understand and easy to extract, but cannot usually describe structural information. Relation-based representations are mostly suitable to encode structural information, but their effectiveness is usually limited to this purpose. In re-id, most of the descriptors are feature-based, while the sole relation-based representation is the covariance of features [1].

Our approach aims at joining both paradigms, exploiting SDALF [2] as feature-based descriptor. SDALF is a symmetry-based description of the human body, and it is inspired by the well-known principle that natural objects manifest symmetry in some form. Using symmetry and asymmetry principles, SDALF isolates three human body regions, usually corresponding to the head, the torso and the legs. After that, torso and legs regions are described by heterogeneous features, and matched by minimizing a proper distance. Our approach complements this scheme, by adding relation-based descriptions: essentially, the body regions found by SDALF are encoded as Mean Riemannian Covariances (MRCs) [1], which are semidefinite positive descriptors built by fusing multiple covariances of features, these latter encoding each shot available of an individual. MRCs are then added to the final descriptions (one for each body region). This produces a novel method, dubbed here SDALF+C.

In the experiments, we show that SDALF+C is an effective solution for direct multi-shot re-identification, allowing to get better results than their single components, on five different multi-shot benchmark datasets (i-LIDS, ETHZ1,2,3 and CAVIAR4REID).

The rest of the paper is organized as follows. In Sec. 2, SDALF and the Mean Riemannian Covariance Grid (MRCG [1], from which the MRC descriptor is extrapolated) are briefly summarized. Sec. 3 details our approach, and Sec. 4 presents the experimental results. Finally, in Sec. 5, conclusions are drawn and future perspectives are envisaged.

## 2 Fundamentals

### 2.1 Symmetry Driven Accumulation of Local Features (SDALF)

Let us suppose to have $M$ images portraying an individual: the SDALF descriptor starts by isolating the foreground (the human body) employing the STEL generative model [9]. After that, SDALF individuates three main body parts (head, torso, legs) by exploiting horizontal asymmetry principles: the rationale is that the head and the torso are horizontally asymmetric (with respect to area and color), and the same applies for the torso and the legs. On the other hand, vertical symmetry criteria allow to weight more those features which are located near the vertical axis of symmetry of the human body, thus pruning out distracting background clutter that lies on the peripheral portions (see Fig. 1 for some examples).

Given the two regions $Reg_{\text{torso}}, Reg_{\text{legs}}$ (the head is discarded as only a few pixels do not contain enough discriminative content), SDALF extracts complementary visual aspects of the human body appearance, highlighting: i) the global chromatic content by the color histogram (in the multiple-shot case,

$M$ histograms for each part are considered); ii) the per-region color displacement employing Maximally Stable Colour Regions (MSCR) [6]; iii) the presence of *Recurrent Highly Structured Patches* (RHSP), estimated by a per-patch similarity analysis. in the multiple-shot case, it is worth noting that 1) the MSCRs are opportunely distilled from the $M$ images by employing a Gaussian clustering procedure [5], which automatically selects the number of components keeping the means, and 2) the RHSP descriptors are extracted considering different frames.

This process applies for all the $M$ individuals of the probe and the gallery sets, obtaining $M$ different signatures. Each signature of the probe set is then compared with the gallery set, looking for a match. To this aim, a proper distance $d_{SDALF}$ is employed. For further details, please refer to [2].

### 2.2 Mean Riemannian Covariance Grid (MRCG)

Let $I$ be an image and $F$ be a $d$-dimensional feature image extracted from $I$,

$$F = \theta(I)$$

where function $\theta$ can be any set of $d$ mappings, such as color, intensity, gradients, filter responses, etc.. For a given rectangular region $Reg \subset F$, let $\{f_h\}_{h=1,\dots,n}$ be the $d$-dimensional feature points inside $Reg$ ($n$ is the number of feature points, e.g. the number of pixels). We represent region $Reg$ by the $d \times d$ covariance matrix of the feature points

$$C_{\text{Reg}} = \frac{1}{n-1} \sum_{h=1}^{n} (f_h - \mu)(f_h - \mu)^{\mathsf{T}} \tag{1}$$

where $\mu$ is the mean of the feature points.

In the original approach, each of the $M$ images of the subject $A$ is decomposed in $K$ patches, where each patch has a fixed location in the image plane. For each patch instance (intended as the patch content of a single image), $d$ dense features are extracted, so that a $d \times d$ covariance matrix can be built for each patch instance. To distill a single descriptor for each patch, which takes into account all the $M$ images of the same subject (i.e., all his patch instances), the Mean Riemannian Covariance (MRC) is calculated, by computing the Karcher mean [8] on all the local covariances. In practice, for patch $k$, a "mean" covariance $\mu_{A,k}$ is built, which summarizes all the correspondent patch instances. Then, in order to weight each MRC, a discriminant index is computed, which considers how different is a particular patch (i.e., its related MRC), from all the correspondent patches of all the other probe images that should be taken into account. In practice, for patch $k$, a discriminant $\sigma_{A,k}$ is created. The same approach is applied on the gallery images. At the end of the process, each patch is described by an MRC, and a discriminant index. To match the probe with a gallery subject $B$, a distance is calculated, which has the following form

$$d_{MRC}(A,B) = \sum_{k=1,\dots,K} \frac{\sigma_{A,k} + \sigma_{B,k}}{\rho(\mu_{A,k}, \mu_{B,k})} \tag{2}$$

where $\rho$ is a proper distance between covariance matrices. Minimizing such distance gives the best match. For further details, please refer to [1].

## 3 Our Approach

Our approach wants to combine SDALF and MRC, since the two descriptors are highly complementary. While SDALF extracts heterogeneous visual properties from the human appearance, MRC tells how visual properties are related with each other. For this purpose, SDALF is run in its original version, obtaining for person $A$ a given descriptor. After that, on the torso and the legs regions $Reg_{\text{torso}}, Reg_{\text{legs}}$ found by SDALF, for all the $M$ images of the same individual, $d$-dimensional covariances matrices are built, following Eq. 1. The following $d = 11$ features are taken into account:

$$\left[x, y, R_{xy}; G_{xy}; B_{xy}; \nabla^R_{xy}, \theta^R_{xy}, \nabla^G_{xy}, \theta^G_{xy}, \nabla^B_{xy}, \theta^B_{xy}\right] \tag{3}$$

where $x$ and $y$ are pixel location, $R_{xy}, G_{xy}, B_{xy}$ are RGB channel values and $\nabla$ and $\theta$ correspond to gradient magnitude and orientation in each channel, respectively. We voluntarily exploit the dense features employed in [1], in order to understand the exact added value that the two descriptors bring in the joint framework. Once the covariances are extracted, the related MRCs (one for the torso, another for the legs) and the associated discriminants $\sigma$ described in Sec. 2.2 are also computed. The two MRCs together with their discriminant indexes compose the relation-based description. After computing the descriptors on all the probe and gallery subjects into play, the matching can be performed considering two subjects $A$ and $B$. To this end, the two distances reported above for the SDALF and the COV descriptors are joined together in a weighted linear fashion, as follows:

$$d_{SDALF+C}(A, B) = \alpha d_{MRC}(A, B) + (1 - \alpha) d_{SDALF}(A, B) \tag{4}$$

where the $\alpha$ coefficient serves to weight the importance of the single description. Estimating the value of $\alpha$ giving the maximum performance will help to understand the interplay of the two components. It is important to note that the two distances are opportunely normalized to sum up to one.

## 4 Experiments

Experiments have been performed on different multi-shot datasets (i-LIDS for re-id [11], ETHZ[1] 1, 2, and 3 , and CAVIAR4REID[2]), in order to evaluate our proposal against diverse re-id problems, as explained in the following. As metrics, we adopt the standard Cumulative Matching Characteristic (CMC) curve, which represents the probability of finding the correct match in the top $n$ ranks; in practice, after calculating the distance of a probe individual with all the gallery subjects, a ranking is made, and the position of the correct match is kept. On the CMC curve, the rank 1, rank 10 and rank 20 probabilities are usually reported numerically, as so as the normalized Area Under the Curve (nAUC), which is the area under the entire CMC curve normalized over the total area of the graph. As

[1] http://www.liv.ic.unicamp.br/~wschwartz/datasets.html

[2] http://www.lorisbazzani.info/code-datasets/caviar4reid/

comparative approaches, we consider SDALF, the MRC part taken alone, and the MRCG approach [1], when the results are available.

In order to assess how the two components of the approach interact, we perform an explorative analysis by mediating the nAUC scores obtained on all the datasets (for the experimental protocol for each benchmark, see below) with different multi-shot cardinalities, i.e., number of images that compose a signature, i.e., $M = 2, 5$.

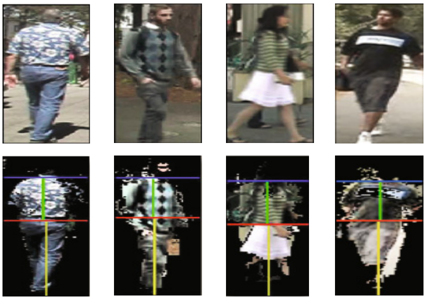


**Fig. 1.** Example of partitions obtained with the SDALF approach (best viewed in colors)

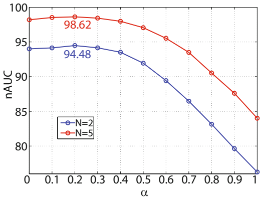


**Fig. 2.** Analysis of the influence of the $\alpha$ value on the SDALF+C performance: high $\alpha$ means high weight for the MRC part of the descriptor (best viewed in colors)

As visible in Fig. 2, we have for $M = 2$ the best nAUC for $\alpha = 0.2$ and the same happens with $M = 5$: this witnesses that SDALF plays a primary role, but MRC furnishes a complementary information which produces the best performance, independently on the cardinality of the multi-shot signature. Therefore, in all the next experiments, we report the performance of SDALF+C employing this $\alpha$ value as fixed parameter. Using this setting, we overcome in all the datasets the performances of SDALF and MRC. In addition, for each dataset, we report the performance with $\alpha_{\text{best}}$, i.e., the alpha value for which SDALF+C gives its best on that benchmark (that is, the best nAUC): this provides an upper bound of the SDALF+C performances.

In the following, we discuss the results obtained on each dataset.

***i-LIDS for Re-Identification Dataset.*** The i-LIDS Multiple-Camera Tracking Scenario dataset is a public video dataset captured at a real airport arrival hall in the busy times under a multi-camera CCTV network. In [11], i-LIDS for re-identification dataset has been built from i-LIDS Multiple-Camera Tracking Scenario. The dataset is composed by 479 images of 119 people. The images, normalized to 64 × 128 pixels, derive from non-overlapping cameras, under quite large illumination changes and subject to occlusions. This dataset a critical multi-shot scenario because the average number of images per person is 4, and thus some individuals have only two images.

The signatures are built from $M$ images of the same pedestrian, randomly selected. Due to the average number of images per pedestrian, we tested SDALF+C with $M = 2$, running 10 independent trials for each case. It is worth noting that some of the pedestrians have less than 4 images: therefore, in such a case, we simply build a multi-shot signature composed by less instances. The results show that SDALF+C gives better performances (in terms of nAUC) of all its separate components, overcoming also the MRCG approach: this happens either with the $\alpha$ value kept fixed at 0.2, and with the best value for this dataset, i.e., $\alpha = 0.1$.[1]

**Table 1.** Performances on i-LIDS for re-identification

| **i-LIDS M=2** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 45.04 | 69.13 | 78.30 | 86.55 | 93.02 |
| **MRC only** | 9.78 | 29.46 | 40.27 | 52.35 | 74.43 |
| **MRCG [1]** | 46.25 | 67.50 | 76.00 | 83.75 | - |
| **SDALF + C** ($\alpha = 0.20$) | **47.40** | **72.55** | **80.43** | **87.66** | 93.36 |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.10$) | 47.14 | 72.24 | 80.13 | 87.26 | **93.41** |

***ETHZ Dataset.*** The data are captured from moving cameras in a crowded street. The challenges covered by this dataset are illumination changes, occlusions and low resolution (32×64 pixels). This dataset contains three sub-datasets: ETHZ1 with 83 people (4.857 images), ETHZ2 with 35 people (1.936 images), and ETHZ3 contains 28 with (1.762 images). Even if this dataset does not mirror a genuine re-identification scenario (a single camera is employed), it still carries important challenges not exhibited by other public dataset, as the high number of images per person. The protocol is the same as the one employed for i-LIDS, but here we also include $M = 5$ (as more per-person images are available). As visible in Table 2, in all the cases the nAUC performances of SDALF+C, both choosing the best $\alpha$, or keeping it fixed at $\alpha = 0.2$, are better that the SDALF and the MRC ones. Please note that here, being the nAUC scores near 100%, it is more difficult to get a strong improvement.

***CAVIAR for Re-Identification Dataset.*** CAVIAR4REID dataset contains images of pedestrians extracted from the CAVIAR repository, and consists of several images captured in a shopping centre in Lisbon. A total of 72 unique pedestrians have been identified: 50 with both the camera views (20 images per pedestrian) and 22 with one camera view (10 images per pedestrian). The challenging features of this dataset are a broad change in the image resolution, with a minimum and maximum size of $17 \times 39$ and $72 \times 144$, respectively; pose variations are severe, as so as the illumination changes and the occlusions.

In this case, we took only the 50 individuals for which 20 images are available, 10 per camera: images taken from one camera form the probe set, the other

[1] We remember here that as best performance for an approach we consider that one which gives the best nAUC, irrespective of the other figure of merits.

**Table 2.** Performances on ETHZ1 (a), ETHZ2 (b), ETHZ3 (c)

(a) ETHZ1

| **ETHZ1 M=2** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | **74.12** | **89.20** | 92.24 | **95.08** | 96.68 |
| **MRC only** | 18.48 | 33.61 | 44.41 | 59.13 | 75.20 |
| **SDALF + C** ($\alpha = 0.20$) | 73.86 | 88.39 | **92.72** | 95.04 | **96.71** |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.20$) | 73.86 | 88.39 | **92.72** | 95.04 | **96.71** |

| **ETHZ1 M=5** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 86.36 | 94.07 | 95.81 | 96.60 | 97.80 |
| **MRC only** | 23.47 | 43.47 | 54.87 | 70.12 | 81.53 |
| **SDALF + C** ($\alpha = 0.20$) | **86.70** | **94.36** | **95.93** | **96.80** | **97.99** |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.20$) | **86.70** | **94.36** | **95.93** | **96.80** | **97.99** |

(b) ETHZ2

| **ETHZ2 M=2** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 83.71 | 95.77 | **98.69** | 99.43 | 98.11 |
| **MRC only** | 20.00 | 51.14 | 72.57 | 91.77 | 80.54 |
| **SDALF + C** ($\alpha = 0.20$) | 84.51 | 96.17 | 98.63 | **99.83** | 98.36 |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.10$) | **84.91** | **96.46** | 98.51 | 99.71 | **98.57** |

| **ETHZ2 M=5** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 90.97 | 97.71 | **99.26** | 99.43 | 98.94 |
| **MRC only** | 29.89 | 66.00 | 84.29 | 96.11 | 86.93 |
| **SDALF + C** ($\alpha = 0.20$) | **92.57** | **98.69** | **99.26** | **99.83** | **99.26** |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.20$) | **92.57** | **98.69** | **99.26** | **99.83** | **99.26** |

(c) ETHZ3

| **ETHZ3 M=2** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 88.79 | 97.86 | 99.64 | **100.00** | 98.86 |
| **MRC only** | 33.29 | 74.54 | 86.21 | 96.07 | 86.33 |
| **SDALF + C** ($\alpha = 0.20$) | **92.79** | **99.50** | 99.71 | **100.00** | 99.40 |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.30$) | 92.14 | 99.43 | **100.00** | **100.00** | **99.46** |

| **ETHZ3 M=5** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 95.14 | 99.21 | **100.00** | **100.00** | 99.30 |
| **MRC only** | 42.50 | 82.43 | 90.79 | 98.07 | 90.28 |
| **SDALF + C** ($\alpha = 0.20$) | 96.43 | **100.00** | **100.00** | **100.00** | 99.76 |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.30$) | **97.50** | **100.00** | **100.00** | **100.00** | **99.87** |

camera individuates the gallery. This way, chromatic dissimilarity between probe and gallery images is maximized. All the images are resampled at $64 \times 32$ pixels, and ten independent trials have been run. Results are reported in Table 3. In this case, the best performances of SDALF+C are obtained exploiting the standard $\alpha = 0.2$ (so $\alpha$ and $\alpha_{\text{best}}$ do coincide), overcoming SDALF and the MRC ones.

## 5 Conclusions

In this paper, we provide a novel hybrid descriptor for re-id, SDALF+C, which joins together a feature-based and a relation-based description of the human appearance. The former focuses on characterizing visual properties of the human

**Table 3.** Performances on CAVIAR4REID

| **CAVIAR4REID M=2** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
|---|---|---|---|---|---|
| **SDALF** | 32.16 | 57.20 | 70.64 | 84.12 | 83.34 |
| **MRC only** | 8.32 | 24.88 | 38.60 | 58.72 | 64.75 |
| **SDALF + C** ($\alpha = 0.20$) | **34.96** | **60.80** | **72.68** | **85.24** | **84.55** |
| **SDALF + C** ($\alpha_{\mathbf{best}} = 0.20$) | **34.96** | **60.80** | **72.68** | **85.24** | **84.55** |
| **CAVIAR4REID M=5** | **rank1** | **rank5** | **rank10** | **rank20** | **nAUC** |
| **SDALF** | 72.04 | 89.20 | 94.28 | **98.08** | 96.52 |
| **MRC only** | 15.72 | 42.04 | 58.60 | 78.08 | 77.30 |
| **SDALF + C** ($\alpha = \alpha_{\mathbf{best}} = 0.20$) | **74.40** | **91.32** | **96.16** | 96.68 | **97.42** |

body, the latter captures how visual properties are interrelated. The experimental results show that, in terms of nAUC, SDALF+C overcomes the single parts (visual-based and relation-based) of which it is composed, in a systematic way. Therefore, our proposal paves the way for further studies, aimed at providing hybrid solutions for the single-shot re-identification case. In addition, we plan to embed SDALF+C in a learning framework, in order to automatically infer the best value for alpha for a given scenario.

## References

1. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: AVSS (2011)
2. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. CVIU 117(2), 130–144 (2013)
3. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. PRL 33(7), 898–903 (2012)
4. Figueira, D., Bazzani, L., Quang, M.H., Cristani, M., Bernardino, A., Murino, V.: Semi-supervised multi-feature learning for person re-identification. In: AVSS (2013)
5. Figueiredo, M., Jain, A.K.: Unsupervised learning of finite mixture models. TPAMI 24(3), 381–396 (2002)
6. Forssen, P.E.: Maximally stable colour regions for recognition and matching. In: CVPR (2007)
7. Hirzer, M., Roth, P.M., Kostinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: ECCV (2012)
8. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. IJCV 66(1), 41–66 (2006)
9. Perina, A., Jojic, N., Cristani, M., Murino, V.: Stel component analysis: Joint segmentation, modeling and recognition of objects classes. IJCV 100(3), 241–260 (2012)
10. Satta, R., Fumera, G., Roli, F., Cristani, M., Murino, V.: A multiple component matching framework for person re-identification. In: Maino, G., Foresti, G.L. (eds.) ICIAP 2011, Part II. LNCS, vol. 6979, pp. 140–149. Springer, Heidelberg (2011)
11. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: BMVC (2009)
12. Zheng, W., Gong, S., Xiang, T.: Re-identification by relative distance comparison. TPAMI (99) (2012)

# Multi-sensor Fusion Using Dempster's Theory of Evidence for Video Segmentation

Björn Scheuermann, Sotirios Gkoutelitsas, and Bodo Rosenhahn

Institut für Informationsverarbeitung (TNT)
Leibniz Universität Hannover, Germany
{last name}@tnt.uni-hannover.de

**Abstract.** Segmentation of image sequences is a challenging task in computer vision. Time-of-Flight cameras provide additional information, namely depth, that can be integrated as an additional feature in a segmentation approach. Typically, the depth information is less sensitive to environment changes. Combined with appearance, this yields a more robust segmentation method. Motivated by the fact that a simple combination of two information sources might not be the best solution, we propose a novel scheme based on Dempster's theory of evidence. In contrast to existing methods, the use of Dempster's theory of evidence allows to model inaccuracy and uncertainty. The inaccuracy of the information is influenced by an adaptive weight, that provides a measurement of how reliable a certain information might be. We compare our method with others on a publicly available set of image sequences. We show that the use of our proposed fusion scheme improves the segmentation.

## 1 Introduction

Segmentation of foreground objects in video sequences is a fundamental step in many computer vision applications and has been widely studied in the last years. A popular application in movie production is the integration of virtual objects into a sequence [1]. Because of many aspects in real-world scenarios video segmentation is a very challenging task. Illumination changes or background appearance changes, caused by people walking around, are typical problems that need to be treated.

The segmentation problem can be formulated using probabilistic models like Markov or conditional random fields. It has been shown, that the maximum a posteriori solution for these models corresponds to the discrete minimization of an appropriate energy function [2–4].

Time-of-Flight (ToF) cameras are perfect candidates to simplify the problem of binary video segmentation. ToF cameras use active sensors to measure the time taken by infrared light to travel to the object and back to the camera. The travel time corresponds to a certain depth value. Thus, ToF cameras are able to determine the depth value for the pixels in an image, which can be seen as additional information for each pixel.

The proposed algorithm is related to many recent works on binary image or video segmentation [2–7]. In [2–4], the authors use a discrete energy minimizing framework for interactive image segmentation. The problem of segmentation is transferred on a graph, where the minimum cut corresponds to the minimum energy state. In [5] and

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 431–438, 2013.

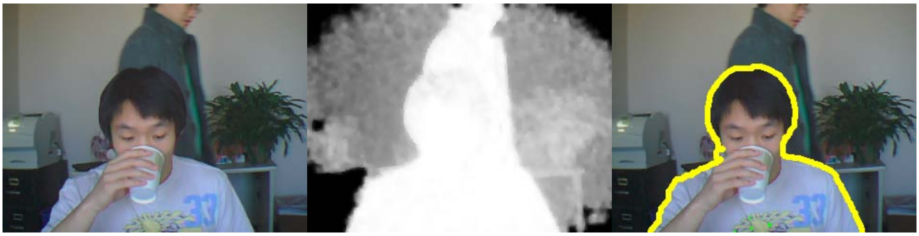


**Fig. 1.** Example segmentation result by fusing color and depth information using Dempster's theory of evidence. The explicit modeling of uncertainty forces the algorithm to segment the person in the foreground even if the depth information of the person in the background is similar. Input data taken from [9].

in [7], stereo images where used to estimate the scene depth. They showed that the combination of estimated depth and color improves the segmentation result. However, the estimation of the scene depth is a non trivial problem that is prone to errors in real-world scenarios.

The two most related methods are [8, 9]. In [8], Scheuermann and Rosenhahn proposed to use Dempster's theory of evidence for energy minimizing segmentation. They proposed a variational energy functional, including mass functions to fuse color and texture information, and solved it using level sets. In [9], Wang et al. proposed a very similar method, the so-called ToFCut algorithm. They combine depth and color cues in a discrete energy function and weight the information adaptively.

In this paper, we propose a novel method to fuse color and depth information in a discrete energy function. Therefore we use Dempster's theory of evidence to combine the different information. Using the proposed feature fusion allows us to explicitly model inaccuracy and uncertainty. This modeling provides an elegant way to incorporate the reliability of a feature channel. The information about how reliable a feature channel might be, can be either defined manually, based on prior information, or using our proposed adaptive weighting function. The adaptive weighting uses the symmetric Kulback-Leibler divergence as a measure of reliability. Therefore we compute distances of foreground and background histograms based on the segmentation result of the previous frame.

In summary, our main contributions are:

- A novel discrete energy function including Dempster's theory of evidence for feature fusion.
- An adjustable mass function, that can either use prior information or an adaptive weighting function based on the symmetric Kullback-Leibler divergence.
- Improved color and depth models, that are more robust.

In contrast to [9], we propose to use Dempster's theory of evidence to fuse color and depth information. We show that the proposed discrete energy function is more intuitive then the ToFCut functional. Furthermore, we propose stable functions, based on the Kulback-Leibler divergence, to adaptively compute the confidence of each sensor.

The experimental validation on the data set used in [9] shows that the proposed method outperforms ToFCut.

## 2 Segmentation by Discrete Energy Minimization

The problem of binary segmenting an image or image sequence can be formalized by minimization of a discrete energy function $E : \mathcal{L}^n \to \mathbb{R}$. Usually the energy function is written as the sum of unary $\varphi_i$ and pairwise $\varphi_{i,j}$ potentials.

$$E(x) = \sum_{i \in \mathcal{V}} \varphi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \varphi_{i,j}(x_i, x_j)\,, \tag{1}$$

where $x \in \mathcal{L}^n$ is a labeling, $\mathcal{V}$ corresponds to the set of all image pixels and $\mathcal{E}$ is the set of all neighboring pixels. In case of binary segmentation, the label set $\mathcal{L}$ consists of foreground (FG) and background (BG) labels. The unary potential $\varphi_i$ is given as the negative log-likelihood of a probability measure, e.g. a standard Gaussian mixture model (GMM) [4]:

$$\varphi_i(x_i) = -\log p(I_i \mid x_i = L)\,, \tag{2}$$

where $I_i$ is the feature vector of pixel $i$, e.g. RGB values. $L$ is either FG or BG and $p$ is the likelihood. The pairwise potential is usually given by a contrast sensitive Ising model, defined by

$$\varphi_{i,j}(x_i, x_j) = \gamma \cdot \mathrm{dist}(i,j)^{-1} \cdot [x_i \neq x_j] \cdot \exp(-\beta ||I_i - I_j||^2)\,. \tag{3}$$

Here $\gamma$ specifies the impact of the pairwise potential, $[\cdot]$ is the indicator function and $\mathrm{dist}(\cdot)$ is the Euclidean distance between neighboring pixels. The parameter $\beta$ is defined as $\beta = (2\langle ||I_i - I_j||^2 \rangle)^{-1}$, where $\langle \cdot \rangle$ indicates expectation [10].

In [9], the energy function is extended by means of additional depth information. Therefore, the unary potential takes the form:

$$\varphi_i(x_i) = -\lambda_c \cdot \log p_c(I_i \mid x_i = L) - \lambda_d \cdot \log p_d(D_i \mid x_i = L)\,, \tag{4}$$

where $D_i$ is the depth of pixel $i$. The likelihood $p_c$ is a GMM learned using 3D histograms with $8^3$ bins in the RGB color space and the likelihood for depth $p_d$ is modeled by two Gaussian distributions. The parameters $\lambda_c$ and $\lambda_d$ are used to adaptively weight the impact of both cues. They are based on the discriminative capabilities of the two likelihoods. The color confidence is computed using the Kulback-Leibler divergence (KL) between the grayscale histograms of frames $I^{t-1}$ and $I^t$ (denoted by $\delta_{lum}^{KL}$) and the KL divegence between foreground and background color histograms of frame $I^{t-1}$ ($\delta_{rgb}^{KL}$) . This yields the confidence of the color term

$$\mathcal{R}_c = \exp\left(-\frac{\delta_{lum}^{KL}}{\eta_{lum}}\right) \cdot \left(1 - \exp\left(-\frac{\delta_{rgb}^{KL}}{\eta_{rgb}}\right)\right)\,, \tag{5}$$

with parameters $\eta_{lum}$ and $\eta_{rgb}$. The depth confidence $\mathcal{R}_d$ is computed using the distance between the average depth values for foreground and background in frame $I^{t-1}$

($\Delta\chi = |(\chi^f + \chi'^f) - (\chi^b + \chi'^b)|/2$). Here, $\chi^f, \chi'^f, \chi^b$ and $\chi'^b$ are the mean values of the Gaussian distributions $p_d$. This yields

$$\mathcal{R}_d = 1 - \exp\left(-\frac{\Delta\chi}{\eta_d}\right) , \tag{6}$$

with the additional parameter $\eta_d$. Finally the adaptive weights are defined as $\lambda_c = \mathcal{R}_c/(\mathcal{R}_c + \mathcal{R}_d)$ and $\lambda_d = \mathcal{R}_d/(\mathcal{R}_c + \mathcal{R}_d)$. For more details on the likelihood terms and the adaptive weighting the reader is referred to [9].

In contrast to ToFCut, we propose to use the symmetric Kulback-Leibler divergence, since the symmetric distance does not depend on the order of the feature channels. We also use the symmetric KL divergence to measure the distance between FG and BG depth histograms in frame $I^{t-1}$, since the given definition using $\Delta\chi$ lacks in precision.

It has been shown that, using the defined unary and pairwise potentials, the energy (1) is submodular and can hence be represented by a graph $G$ [10]. In this form, the global minimum of the energy function corresponds to the minimum cut of graph $G$ that can be computed using standard maximum flow algorithms [11].

### 2.1 Dempster's Theory of Evidence

We continue with a brief review of Dempster's theory of evidence [12, 13], which is later used to fuse color and depth cues. Several works [8, 14, 15] applied the theory to image segmentation and showed that it can be superior to classical probability theory.

Dempster's theory of evidence is a generalization of classical probability theory, with the ability to jointly represent inaccuracy and uncertainty information. The theory is build on so-called basic probability assignments (also known as mass functions), that are defined on a hypotheses set $\Omega$. In our case, the hypotheses set is composed by the labels FG and BG. The mass function $m(A) : \wp(\Omega) \rightarrow [0, 1]$ is defined on the power set of $\Omega$.

The quantity $m(A)$ is interpreted as the belief strictly placed on hypothesis $A$. In contrast to classical probability theory, this belief is distributed on both simple and composed classes and models the impossibility to separate several hypotheses. This characterizes the principal advantage of the evidence theory.

Another particular characteristic of Dempster's theory, one which differs from classical probability theory, is: if $m(A) < 1$, then the remaining mass $1 - m(A)$ does not need necessarily refute $A$ (i.e. support its negation). Thus we do not have the so-called additivity rule $p(A) + p(\overline{A}) = 1$.

To fuse mass functions from different feature channels we use Dempster's rule of combination, denoted by $m = m_1 \otimes m_2$. This rule combines two independent bodies of evidence, defined on the same hypotheses set $\Omega$, into one body of evidence. Since Dempster's rule of combination has shown to be associative, we can combine information arising from more than two channels.

## 3 Feature Fusion Using Dempster's Theory of Evidence

In this Section we describe the details of our proposed segmentation scheme and show similarities and differences to existing approaches.

The unary potential used by ToFCut is defined as a weighted sum of negative log likelihoods, see Equation (4), and can be reformulated as:

$$\varphi_i(x_i) = -\log\left[p_c(I_i|x_i = L)^{\lambda_c} \cdot p_d(D_i|x_i = L)^{\lambda_d}\right] \,, \tag{7}$$

which can be interpreted as follows: if the confidence for a channel is near zero, the likelihood is near one. That means, to ignore a channel we push the corresponding likelihoods near one. This is a neither intuitive nor elegant solution. Furthermore, this non-linear solution heavily depends on a good adaptive weighting function.

In contrast to ToFCut our unary potential is defined using Dempster's basic probability assignment:

$$\varphi_i^{DS}(x_i) = -\log m(x_i = L) \,, \tag{8}$$

where the mass function $m = m_c \otimes m_d$ fuses the information of color and depth according to Dempster's rule of combination. Thus the complete energy function reads:

$$E(x) = \sum_{i\in\mathcal{V}} \varphi_i^{DS}(x_i) + \sum_{(i,j)\in\mathcal{E}} \varphi_{i,j}(x_i, x_j) \,, \tag{9}$$

Using the proposed unary potential $\varphi_i^{DS}$, we can elegantly model the uncertainty of a channel by defining the corresponding mass functions appropriately. Since we use Dempster's rule of combination, that is associative, we can also include additional information e.g. texture and motion.

### 3.1 Mass Functions

The most important difference between the proposed method and ToFCut is the feature fusion using Dempster's theory of evidence instead of summing up weighted log-likelihoods. Therefore the main contribution is the definition of appropriate mass functions, that model inaccuracy and uncertainty in an elegant way. The mass functions modeling color and depth information are defined by:

$$\begin{aligned} m_c(\varOmega) &= \frac{\lambda_d(1 - (p_c(I_i|x_i = \text{FG}) + p_c(I_i|x_i = \text{BG})))}{K} \,, \\ m_c(L) &= (1 - m_c(\varOmega))\frac{p_c(I_i|x_i = L)}{p_c(I_i|x_i = \text{FG}) + p_c(I_i|x_i = \text{BG})} \end{aligned} \tag{10}$$

for the color term and

$$\begin{aligned} m_d(\varOmega) &= \frac{\lambda_c(1 - (p_d(I_i|x_i = \text{FG}) + p_d(I_i|x_i = \text{BG})))}{K} \,, \\ m_d(L) &= (1 - m_d(\varOmega))\frac{p_d(D_i|x_i = L)}{p_d(D_i|x_i = \text{FG}) + p_d(D_i|x_i = \text{BG})} \end{aligned} \tag{11}$$

for the depth term, where $L$ is either FG or BG. The uncertainty $m_c(\varOmega)$ and $m_d(\varOmega)$ of the models is defined by summing up the likelihoods. This means that the uncertainty of a model is high, if FG and BG likelihoods are small. The normalization factor $K$ is chosen so that $m_c(\varOmega) + m_d(\varOmega) = 1$, which means that the sum of modeled uncertainty is one. The parameters $\lambda_d$ and $\lambda_c$ are the adaptive weights coming from the histogram analysis. They can be used to further increase or decrease the importance of a feature channel.

**Table 1.** Comparison between the proposed method DS and ToFCut obtained on four video sequences. The mean percentage error, computed across the whole sequence, is provided. The results obtained by ToFCut are taken from [9]. The proposed method clearly outperforms ToFCut.

| Seq. ID | WL | | MS | | MC | | CW | |
|---|---|---|---|---|---|---|---|---|
| No. Frames | 200 | | 400 | | 300 | | 300 | |
| Alg. | ToFCut | DS | ToFCut | DS | ToFCut | DS | ToFCut | DS |
| % Error (Equal Weight Fusion) | 1.37 | 0.54 | 0.51 | 0.23 | 0.16 | 0.06 | 11.68 | 2.21 |
| % Error (Adaptive Weight Fusion) | 1.35 | 0.51 | 0.51 | 0.23 | 0.15 | 0.06 | 0.38 | 0.26 |

### 3.2 Color and Depth Likelihoods

We also use an improved color model, since the one proposed in [9] is sensitive to small bins and lacks in precision, leading to suboptimal segmentation results. Similarly to [9], we use two 3D histogram with $H = 8^3$ bins in the RGB space for FG and BG. For each bin we learn a 3D-Gaussian with mean $\mu_k^j$, covariance matrix $\Sigma_k^j$ and weight $w_k^j$, for $k \in 1 \dots H$ and $j \in \{\text{FG, BG}\}$. The conditional probability is now given by:

$$p(I_i \mid x_i = L) = \sum_{i \in \mathcal{N}} w_i^L G(I_i | \mu_i^L, \Sigma_i^L) \, . \tag{12}$$

In contrast to ToFCut we omit the normalization term, to make the model more robust.

To model the depth likelihoods we use the conditional probability proposed by Wang et al. [9], where two Gaussian's are used for foreground and background. Furthermore we define a threshold $T$ on the depth map, to exclude pixels from the training of the Gaussians. This threshold forces pixels with a depth value smaller than $T$ to be segmented as background and improves our FG and BG models. Thus, the single parameter $T$ is intuitive and easy to adjust.

## 4 Experimental Results

In this Section, the evaluation of the proposed method is presented. For qualitative and quantitative analysis we use the ToFCut data set with the corresponding ground truth data [1]. In Table 1 we present the obtained results and compare them to ToFCut by means of mean percentage error of misclassified pixels [5, 9]. In the experiments we use an equal weight fusion of color and depth information by setting $\lambda_c = \lambda_d = 0.5$ and an adaptive weight fusion based on histogram analysis. The quantitative results show that for both systems, equal weight fusion and adaptive weight fusion, the proposed fusion with Dempster's theory outperforms ToFCut. Important to notice is, that we only need to adjust two intuitive parameters: $\gamma$, the weighting of neighboring discontinuities and $T$, the threshold of the depth map. The parameters $\eta_{lum}, \eta_{rgb}$ and $\eta_d$, controlling the adaptive weighting, remain constant in all our experiments, while in [9] they have to be adjusted for each sequence manually. Furthermore, the results show that the proposed

[1] `http://vis.uky.edu/`

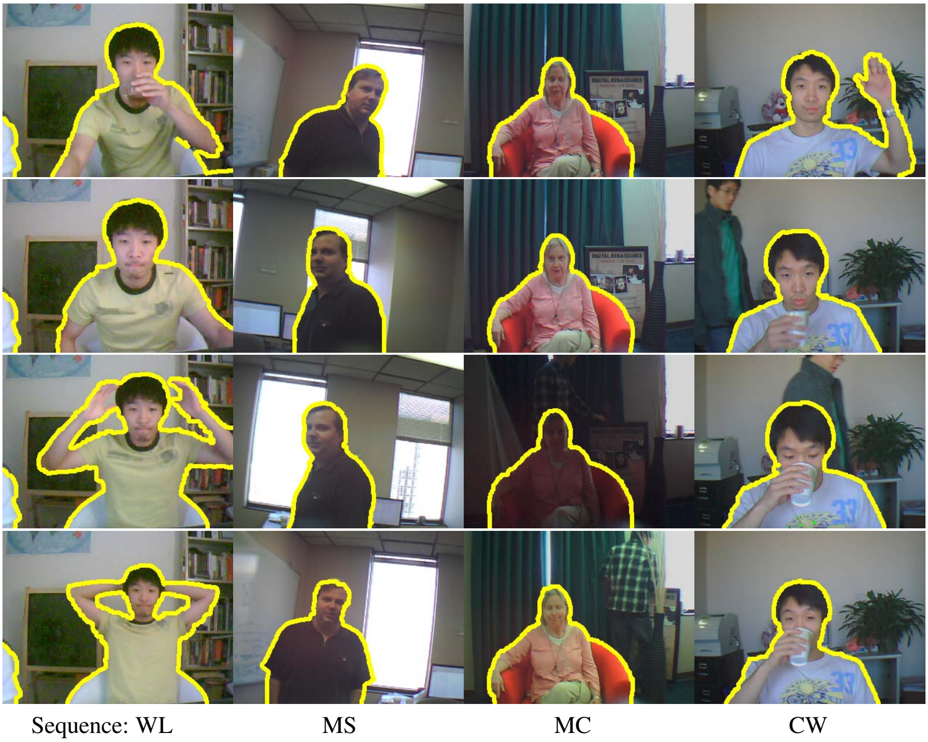


**Fig. 2.** Example segmentation results, on four sample frames from each of the video sequences

fusion works well on many sequences without an adaptive weighting. Qualitative results for all sequences are presented in Figure 2. They show that the small segmentation error corresponds to a high-quality segmentation.

Besides video segmentation, interactive image segmentation is a challenging task. Since there exists no benchmark including depth images, we use the same data set. Qualitative results are presented in Figure 3. Since color and depth models are learned from rough user strokes, the models are likely to be incomplete. By using the proposed fusion based on Dempster's theory of evidence, this is elegantly modeled in our mass functions and the segmentation result outperforms ToFCut.

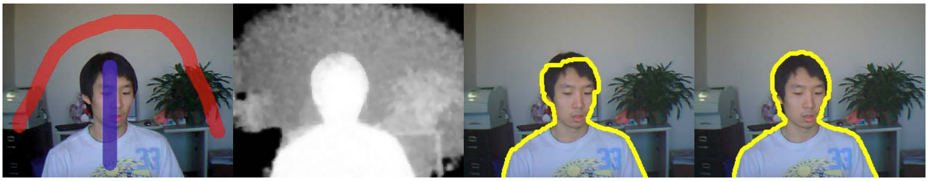


**Fig. 3.** Example interactive segmentation result. From left to right: Color image with initialization (FG in blue/BG in red), corresponding depth image, segmentation result using ToFCut with equal weights, proposed DS fusion with equal weights.

## 5 Conclusion

The paper presents a novel video segmentation scheme. It uses Dempster's theory of evidence to fuse color and depth information. With Dempster's theory of evidence we are able to define the uncertainty of a feature in an elegant way using prior information or an adaptive weight based on the symmetric Kullback-Leibler divergence. Furthermore, we propose adjusted color and depth models to improve the segmentation results. The quantitative evaluation shows that the proposed method outperforms ToFCut. In contrast to ToFCut, the proposed method has less parameters that are more intuitive and easy to adjust. An additional property of the proposed fusion scheme is the naturally given possibility to include further information like motion or user priors.

## References

1. Cordes, K., Scheuermann, B., Rosenhahn, B., Ostermann, J.: Learning object appearance from occlusions using structure and motion recovery. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 611–623. Springer, Heidelberg (2013)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. TPAMI 23(11), 1222–1239 (2001)
3. Blake, A., Rother, C., Brown, M., Perez, P., Torr, P.: Interactive image segmentation using an adaptive GMMRF model. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 428–441. Springer, Heidelberg (2004)
4. Rother, C., Kolmogorov, V., Blake, A.: Grab cut: Interactive foreground extraction using iterated graph cuts. In: SIGGRAPH, vol. 23, pp. 309–314 (2004)
5. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Bi-layer segmentation of binocular stereo video. In: CVPR, vol. 2, pp. 407–414. IEEE (2005)
6. Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: CVPR, vol. 1, pp. 53–60. IEEE (2006)
7. Harville, M., Gordon, G., Woodfill, J.: Foreground segmentation using adaptive mixture models in color and depth. In: EVENT, pp. 3–11 (2001)
8. Scheuermann, B., Rosenhahn, B.: Feature quarrels: The dempster-shafer evidence theory for image segmentation using a variational framework. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 426–439. Springer, Heidelberg (2011)
9. Wang, L., Zhang, C., Yang, R., Zhang, C.: Tofcut: Towards robust real-time foreground extraction using a time-of-flight camera. In: 3DPVT (2010)
10. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient nd image segmentation. IJCV 70(2), 109–131 (2006)
11. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? TPAMI 26(2), 147–159 (2004)
12. Dempster, A.P.: A generalization of Bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological) 30(2), 205–247 (1968)
13. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
14. Adamek, T., O'Connor, N.E.: Using Dempster-Shafer theory to fuse multiple information sources in region-based segmentation. In: ICIP, pp. 269–272 (2007)
15. Chaabane, S.B., Sayadi, M., Fnaiech, F., Brassart, E.: Dempster-Shafer evidence theory for image segmentation: application in cells images. IJSP (2009)

# A One-Shot DTW-Based Method for Early Gesture Recognition

Yared Sabinas, Eduardo F. Morales, and Hugo Jair Escalante

Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro # 1, Tonantzintla, Puebla, México
{y.sabinas,emorales,hugojair}@inaoep.mx

**Abstract.** Early gesture recognition consists of recognizing gestures at their beginning, using incomplete information. Among other applications, these methods can be used to compensate for the delay of gesture-based interactive systems. We propose a new approach for early recognition of full-body gestures based on dynamic time warping (DTW) that uses a single example from each category. Our method is based on the comparison between time sequences obtained from known and unknown gestures. The classifier provides a response before the unknown gesture finishes. We performed experiments in the MSR-Actions3D benchmark and another data set we built. Results show that, in average, the classifier is capable of recognizing gestures with 60% of the information, losing only 7.29% of accuracy with respect to using all of the information.

**Keywords:** Early gesture recognition, DTW, one-shot learning, Kinect.

## 1 Introduction

The automated recognition of gestures has many applications in diverse fields, including video games, sign-language recognition and medical-monitoring systems, among others [5]. Very effective methods for gesture recognition are available nowadays, some of which require of specialized and expensive devices to capture gestures features. The Kinect sensor emerged recently and since then it has boosted the number of applications that make use of gesture recognition technology. This is due to the fact that this sensor is cheaper than similar devices, and provides useful data like RGB-D video and position of body joints (skeleton) in real time [11]. Most of the available methods for gesture recognition provide an answer once the gesture has finished. However, there are certain applications where the delay in gesture recognition is critical, e.g. in interactive and security systems. Despite the importance of this problem, called early gesture recognition, it has been scarcely explored [1,3,6,9].

This paper proposes a new method for early gesture recognition based on DTW using the Kinect sensor. Input sequences are compared with stored ones by using DTW, a prediction criterion is proposed to determine when the method is confident of the identity of the gesture depicted in input sequences. The proposed method can work under the one-shot learning framework [2], that is, using a

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 439–446, 2013.

single example of each gesture category to be recognized. This is advantageous for personalized and dynamic applications, where labeled data is scarce. Our method is easy to implement, it has no training phase and it is very efficient. We report results in a data set we built and in the MSR-Actions3D benchmark. Results show that the method can recognize gestures with 60% of the information, losing only 7.29% of accuracy with respect to using all of the information.

Early classification of gestures is a relatively young field; the first results were published in 2006 by Mori et al. [6]. They wanted to use the anticipated time to compensate the response delay of a robot that imitated their movements. This method was based on dynamic programming, and their gesture dictionary was composed by 18 different gestures that involved only the upper body. With these specifications, they reported up to 1 second anticipation. M. Kawashima et al. [3] and A. Shimada et al. [9] proposed early classification based on self organized maps (SOM) where each neuron learns one different posture of the possible gestures. In [9] the sparse code is extracted from the SOM and then the classification is done. In [3], while the incoming gesture is performed, initial parts from the gestures in the dictionary are chosen, with the intention of comparing similar duration gestures. The comparison of gestures is performed by Hausdorff distance, the gesture with the smallest distance is selected as the answer. Very recently, Ellis et al. proposed a method for early recognition that compares canonical poses (learnt from training data) to test gestures [1]. The authors report acceptable recognition rates, but it is difficult to assess the anticipation performance. In all of these works full body gestures are used, nevertheless, in none of these gestures more than two limbs are moved at the same time.

Differently from previous work, in this paper we recognize no only upper or lower body-movements but full-body movements. Also, our method is based on a DTW cumulative algorithm instead of SOM [3,9] or learned poses [1], thus no training phase is needed as in these alternative works. Furthermore, the proposed approach can work with only one example of each gesture to be recognized, no other early gesture-recognition approach can work under this setting.

## 2 One-Shot Early Recognition of Gestures with DTW

We want to classify full-body gestures made by one person regardless of his/her weight, height or speed of execution of the gestures, More importantly, we want to recognize a gesture before the user finishes its execution. This is a very complex problem because we have to classify the gesture with incomplete information and we do not know its duration beforehand. The problem is further complicated because of the similarity of gestures in the vocabulary, mainly at their beginning parts. Additionally we have to deal with noise incorporated by the considered sensor in the data acquisition process. Therefore, it is complicated to trigger a timely and correct response. We approach the problem with a DTW-based classifier and a novel criterion for early recognition. The proposed method comprises 3 main components: feature extraction, generating partial predictions, and triggering the final decision, which are described in detail in the rest of this section.

### 2.1 Data Representation

While the user is performing a gesture, a virtual skeleton is generated using Kinect [11] and OpenNI/NITE libraries[1]. The skeleton consists of 15 - 3D coordinates corresponding to body joints. The left plot in Figure 1 shows these points. Data is recorded at a speed of 30 frames per second (fps). For each gesture the user performs, we create a $15 \times 3\times$ *number-of-frames* matrix to save the raw data, where *number-of-frames* varies depending on the gesture.

Raw-data collected with the Kinect sensor is represented using a simplified version of the method presented in [8]. This representation reduces dimensionality and makes the data invariant to rotation, translation, and scale. Instead of using principal component analysis as in [8] to get the torso frame, we propose the following: (1) obtain a normalized vector $\overrightarrow{r}$ from the segment $(\overrightarrow{N,\ T})$, where $N$ and $T$ are the neck and torso joints respectively. (2) obtain a normalized vector$\overrightarrow{u}$ from the segment $(\overrightarrow{N,\ RS})$, where $RS$ is the right-shoulder joint, adjusting$\overrightarrow{r}$ in order to $\overrightarrow{u} \cdot \overrightarrow{r} = 1$ and still be a normalized vector, (3) calculate $\overrightarrow{t} = \overrightarrow{u} \times \overrightarrow{r}$. Then, we describe the **first-degree joints**, as in [8], (c.f. Figure 1, left) with two angles ($\theta$ and $\varphi$) which are calculated relative to the *torso frame* and the **second-degree joints** represented by two angles calculated relative to the limb to which they are connected. The result of the transformation is a 16-dim. vector per-frame instead of the initial $3 \times 15$ matrix.

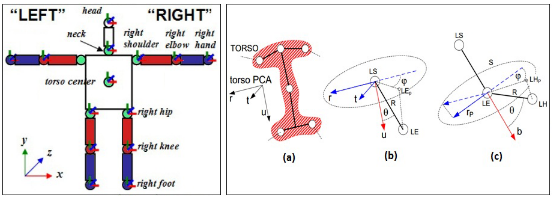


**Fig. 1.** Left: the 15 points of the skeleton we use: first-joints (in red), second-joints (in blue), green circles are used to calculate the *torso frame*. Right: (taken from [8]) shows the *torso frame*(a), and the angles representing first (b) and second (c) degree joints.

### 2.2 Early Classification

Let $\mathcal{D} = \{G_1, \ldots, G_R\}$ be the dictionary of gestures after the data transformation representation process and $G_r = \{f_r(1), \ldots, f_r(T_r)\}$ be one of the gestures in the vocabulary for $r \in \{1, \ldots, R\}$. Each $G_r$ is composed of a sequence of $T_r-$frames, where each frame is represented by 16 angles as described in the previous section: $f_r(t_r) = \{\theta_1, \ldots, \theta_8, \varphi_1, \ldots, \varphi_8, \}$. One should note that we assume we have a single gesture of each particular class, that is, a one-shot learning scenario [2], thus we have $R$ different classes of gestures.

[1] `http://www.openni.org/`

The *new gesture* we want to recognize is denoted by $G_T = \{f_T(1), \ldots, f_T(T_T)\}$, thus $G_T$ has $T_T$ frames. The classifier receives sequentially the frames of the *new gesture* at a 30fps rate. In order to avoid having to make predictions every time a frame is received, the classifier waits until $w$-frames are accumulated and then it makes a partial prediction by comparing *known* gestures with the *new* one. If this is not possible, the method waits again to receive another $w$-frames and it performs another comparison. This iterative process is repeated several times until either the gesture is recognized or the end of the *new gesture* is reached.

For comparing sequences, we estimate the distance between the partial information of the *new gesture* and the partial information of all the *known gestures* in $\mathcal{D}$. For this, we considered each of the 16 angles in a gesture up to time $t_{it}$ as follows: $G_r(t_{it}) = \{A_{r,1}(t_{it}), \ldots, A_{r,16}(t_{it})\}$, where $A_{r,i}(t_{it}) = \{\theta_{r,i}(1), \ldots, \theta_{r,i}(t_{it}), \varphi_{r,i}(1), \ldots, \varphi_{r,i}(t_{it})\}$ for $1 \leq i \leq 16$, are the 16 time sequences of the gesture $G_r$ and $\theta_{r,i}(t_{it})$, $\varphi_{r,i}(t_{it})$ are, respectively, the first and second degree angles of gesture $G_r$ until time $t_{it}$. We used dynamic time warping (DTW) to compute the distance between sequences because it is one of the most used methods to compare sequences that may vary in time or speed. To avoid recalculating the similarity between the partial information of *known* and *new gestures* that was already calculated in previous iterations, we modified DTW to be accumulative (DTWacc): in each iteration DTWacc receives a new part of two time sequences to be compared, calculates the similarity between these parts and adds it to the results of the comparisons of previous iterations, this is shown in Figure 2. The comparison of two time sequences with DTWacc yields a distance value. To calculate the distance between the partial information of a *known* ($G_r$) and the *new gesture* ($G_T$) within DTW we proceed as follows:

$$D(G_r, G_T, t_{it}) = \textstyle\sum_{i=1}^{16} dist(G_{r,i}, G_{T,i}, t_{it})$$

where $dist(G_{r,i}, G_{T,i}, t_{it}) = DTWacc\,(A_{r,i}(t_{it}), A_{T,i}(t_{it}))$, and $A_{r,i}(t_{it})$, $A_{T,i}(t_{it})$ are the sequences of angles of the $i^{th}$-joint up to time $t_{it}$ for the known ($G_r$) and test gestures ($G_T$). We also incorporated a motion threshold $\gamma_m$ to eliminate those limbs that the user hardly moves, and therefore are useless for recognition; thus, only those time sequences that move more than $\gamma m$, are taken

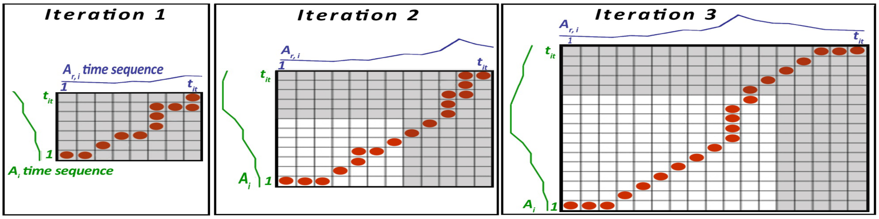


**Fig. 2.** Three iterations of the DTWacc method. In gray we show the part of each sequence that DTWacc compares; in white are shown the results of previous iterations; orange circles show how DTWacc aligns the two time sequences.

into account. For each distance $D(G_r, G_T, t_{it})$ we calculate the normalized score $S(G_r, G_T, T_{it}) = \frac{(D(G_r,G_T,t_{it}))^{-1}}{\sum_r (D(G_r,G_T,t_{it}))^{-1}}$. $S(G_r, G_T, T_{it})$ can be considered the probability that the gesture $r$ is the one depicted in the test gesture $G_T$ up to time $t_{it}$. The gesture with the highest probability will be chosen as a partial prediction for iteration $it$. We propose two ways to take a final decision on the identity of the gesture (i.e., triggering a flag indicating that a gesture has been recognized): **By separation** where one of the *known gestures* is noticeably more similar to the *new gesture.* **By forced classification** where the *new gesture* is about to end, according to an estimate on the duration of the gesture.

For **decision by separation** we consider two aspects: (1) the number of standard deviations $n_\sigma$ that fit in the difference between the best gesture probability and the average of the next $L$ best gesture probabilities, and (2) verify that a certain percentage of the estimated duration of the *new gesture* has been already executed. We defined the constant $L$ to discard the $R - L + 1$ *known gestures* with the lower probabilities. With the remaining gestures we calculate the standard deviation $dev$ and the average $avg$ to calculate $n_\sigma$. If $n_\sigma$ exceeds a certain threshold $\mu$, then the classifier throws a final decision.

On **forced decision** the classifier provides an answer because it is estimated that more than $maxPer$ (a defined limit percentage very close to 100%) of the *new gesture* has been already performed and there was no decision by separation. We do not know how much the *new gesture* will last, so we need to do an estimate to prevent the new gesture of finishing without a prediction from the classifier or prevent hasty decisions. We consider that the total length of the *new gesture* is the minimum duration obtained from the two *known gestures* with the greatest probability on the most recent iteration, therefore, this duration is recalculated in each iteration.

## 3 Experimental Results

For our experiments we used two data sets. The first one is our Dance data set that consist of four dancing gestures: *up an down arm (A), pointing to the sky (B), moving arms and feet (C),* and *cow boy dance (D)* (see Figure 3, left), the gestures were performed by one person ten times each. This data set was captured with a Kinect at a 30fps rate and a resolution of 640x480. The second data set is MSR-Action3D [4], it comprises 20 gestures associated to interactive games (e.g., *side-boxing, tennis serve,* etc). Each gesture was performed by ten subjects for at most three times. The data were captured with Kinect at a 15fps and a resolution of 640x480. For this data set, the skeleton is represented with 20 points, but we only used the 15 available with the OpenNI skeleton. The MSR-Action3D data set has not been previously used for early gesture recognition, but we used here due to the lack of a benchmark for this task. Besides this is one of the most used data sets for action recognition using Kinect data. The parameters of our method: $\mu$, $\gamma_m$, $L$ and $maxPer$ were fixed empirically in preliminary experimentation.

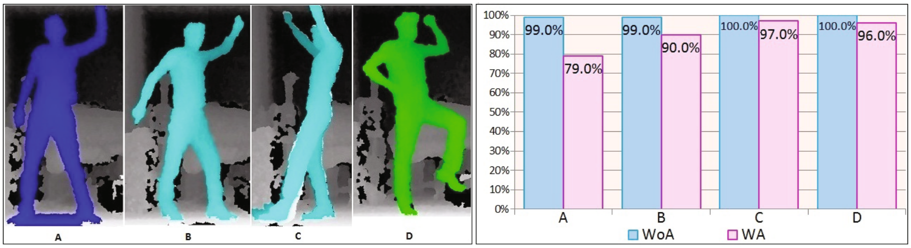


**Fig. 3.** Depth map generated with Kinect device of the four gestures of the data set Dance (left). Precision achieved per gesture WA and WoA for the Dance data set(right).

For the experiments with the dance dataset our approach obtained 99% recognition rate without anticipation (WoA, i.e., using 100% of the information for gestures) and 90% recognition rate with anticipation (WA, i.e., applying our early recognition technique), see Figure 3. On average, the proposed method was able to recognize a gesture using only ≈ 60% of the total duration of gestures and the response time for early recognition was below the 33.5ms.

For the MSR-Action3D data set, in a first experiment, we compared between randomly choosing one example of each gesture and using the half of the gestures to choose the best example of each category to form the training set and the rest of the examples for testing. The results are shown in Table 1 (a), where the column MSR-R shows the results with a random selection and the column MSR-S shows the results with the best selection of half of gestures. It can be seen that very similar results are obtained when using a randomly selected example for each category (MSR-R) and when the best example from the training set is obtained (MSR-S). This result evidences the robustness of our method to the selection of good training examples. For the random selection, we gained 2% of accuracy with anticipation and only needed 55% on average of the total duration of the *new gestures*. Although the accuracy is lower than that in the Dance data set, one must consider that the number of gestures in MSR-Action3D is 5 times larger than in the Dance data set and that gestures were performed by several subjects. The best recognition result for this collection is 88% [10], however, we emphasize that our method works under one-shot learning and it is intended to run with the gestures of a single subject, as in [2]. Another method based on DTW obtained 54% of accuracy in this collection [7], which is slightly better than our proposal, but that method is neither one-shot nor early recognition. Finally, the anticipation method in [1] achieved rates of up to 65.7% in the MSR-Action3D data set, but anticipation performance is not reported.

For the rest of the experiments we considered the MSR-Action3D data set and used the half of the gestures to choose the best example of each category to form the training set and the rest of the examples for testing.

In order to further evaluate the performance of our method when using gestures from a single subject we performed experiments dividing the examples of the MSR-Action3D data set by subjects c.f. Table 2. It can be seen that DTW

**Table 1.** Results of our method in the MSR-Action3D data set

| MSR-R | MSR-S |
|---|---|
| 45 | 50 |
| 47 | 48 |

(a)

| | All subjects | | | Sep. subjects | | | Ref. [4] | | |
|---|---|---|---|---|---|---|---|---|---|
| | AS1 | AS2 | AS3 | AS1 | AS2 | AS3 | AS1 | AS2 | AS3 |
| **WoA** | 42 | 47 | 52 | 97 | 93 | 96 | 72 | 71 | 79 |
| **WA** | 46 | 44 | 50 | 95 | 89 | 93 | - | - | - |

(b)

is very effective for recognizing gestures when a single subject is considered. Also, we can see that the proposed method is very effective at anticipating the recognition of gestures, as accuracy only decreases by 6.5%. Also, the average per-subject performance under WoA and WA (93.2% and 86.7%, respectively) is comparable with the best performance reported so far for the MSR-Action3D data set. As further comparison with other approaches, we divided the gestures by complexity, as reported in [4], where they form three groups: AS1, AS2 and AS3. AS1 and AS2 are intended to group gestures with similar movement (difficult to classify), while AS3 is intended to group very dissimilar actions together. Table 1 (b) shows the results of these sub-groups, considering all the subjects c.f. Table 1.b (columns 1-3); considering subjects and groups, c.f. Table 1.b (columns 4-6), and the results reported in [4], c.f. Table 1.b (columns 7-9).

**Table 2.** Results of the classification by subject

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| WoA | 95.0 | 94.4 | 86.1 | 87.5 | 94.7 | 97.1 | 97.4 | 97.5 | 86.8 | 95.0 | 93.2 |
| WA | 85.0 | 69.4 | 80.6 | 87.5 | 94.7 | 88.2 | 100.0 | 90.0 | 81.6 | 90.0 | 86.7 |

From Table 2 (columns *All subjects*), we can see that the average performances (over groups) when considering all of the subjects are of 47% WA and 48% WoA; thus losing 1% in accuracy but using only the 47% of the duration of gestures. However, when we evaluate the performance over groups by separating users we obtained average performances (over groups) of 92.3% and 95.3% for WA and WoA, respectively (column Sep. subjects in Table 2); for these results only $\approx$ 50% of the gestures were needed for recognition. When compared with the 74% average accuracy obtained in [4], our method has higher precision using half of the information. Therefore, the proposed method is very effective for the classification of gestures when a single-user is considered, even when a single example is used for training the model.

The time required for the classification depends on the number of gestures. For 20 known gestures it takes 50.8ms WA and 450.5ms WoA on average to classify the new gesture. Using only 10 known gestures, it takes 38.9ms WA and 186.8ms WoA. Besides, our method can be parallelized so these response times can be improved.

## 4 Conclusions and Future Work

We proposed a DTW-based method for one-shot early-recognition of gestures. The proposed method is able to recognize gestures before the user finishes of executing it. The highest drop in accuracy when making early recognition was of 7.29% in terms of accuracy, but our savings in recognition response were between $40\%-50\%$. The features of DTW allowed us to design a method for one-shot learning, eliminating the training phase and thereof, the number of labeled gestures needed to generate a model. Also, the method proved to be robust to the selection of examples for the dictionary, and we show that it produces better results when a single-subject performs the gestures. The response time of our method depends directly on the number of *known gestures* used, in our experiments the time needed for classification WoA is up to 8 times larger in average than that required to make classification WA. For future work, we want to include a segmentation method to the classifier to detect the beginning and the end of the gestures in order to achieve online classification. Also we want to parallelize our method to reduce even more the response time. Finally, we want to automate the learning of the parameters $\mu$, $\gamma_m$, $L$ and $maxPer$ to avoid setting them empirically.

## References

1. Ellis, C., Masood, S.Z., Tappen, M.F., LaViola Jr, J., Sukthankar, R.: Exploring the trade-off between accuracy and observational latency in action recognition. International Journal of Computer Vision 101(3), 420–436 (2013)
2. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J., Hamner, B.: Results and analysis of the chaLearn gesture challenge 2012. In: Jiang, X., Bellon, O.R.P., Goldgof, D., Oishi, T. (eds.) WDIA 2012. LNCS, vol. 7854, pp. 186–204. Springer, Heidelberg (2013)
3. Kawashima, M., Shimada, A., Nagahara, H., Taniguchi, R.I.: Adaptive template method for early recognition of gestures. In: 17th WFCV, pp. 1–6. IEEE (2011)
4. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: CVPRW, pp. 9–14. IEEE (2010)
5. Mitra, S.: Gesture recognition: A survey. Trans. on Syst. Man and Cyb. - C 37, 311–324 (2007)
6. Mori, A., Uchida, S., Kurazume, R., Taniguchi, R.-I., Hasegawa, T., Sakoe, H.: Early recognition and prediction of gestures. In: ICPR, pp. 560–563 (2006)
7. Muller, M., Roder, T.: Motion templates for automatic classification and retrieval of motion capture data. In: Proc. SIGGRAPH-SAC, pp. 137–146 (2006)
8. Raptis, M., Kirovski, D., Hoppe, H.: Real-time classification of dance gestures from skeleton animation. In: SoCA, pp. 147–156. ACM (2011)
9. Shimada, A., Kawashima, M., Taniguchi, R.-I.: Early recognition based on co-occurrence of gesture patterns. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part II. LNCS, vol. 6444, pp. 431–438. Springer, Heidelberg (2010)
10. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR, pp. 1290–1297. IEEE (2012)
11. Zhengyou, Z.: Microsoft kinect and its effect. IEEE MultiMedia 19, 4–10 (2012)

# Occlusion Handling in Video-Based Augmented Reality Using the Kinect Sensor for Indoor Registration

Jesus Adrián Leal-Meléndrez, Leopoldo Altamirano-Robles, and Jesus A. Gonzalez

National Institute for Astrophysics, Optics and Electronics, Computer Science Department,
Luis Enrique Erro No. 1, Tonantzintla, Puebla, Mexico
{jalm,robles,jagonzalez}@ccc.inaoep.mx
http://ccc.inaoep.mx

**Abstract.** Video-based Augmented Reality (VAR) aims to add 3D virtual objects (3D VOs) to a real world video sequence, in order to provide additional and useful information to facilitate some tasks, like computer aided surgery, simulation in a real environment, satellite positioning, interior design, among others. To achieve a consistent and convincing augmented scene, it is necessary that the VOs are properly occluded by real objects (Occlusion Problem in VAR); in this paper, we present a strategy based on the use of the *Kinect* sensor to solve this problem. In the occlusion stage we evaluate distances between real and VOs. Then, the parts of the VO occluded by a real object are calculated and removed. We found that the *Kinect* sensor is appropriate to be used for handling occlusions in indoor environments, dynamic scenarios and real-time applications. Experiments showed comparable results with the state of the art in both issues: occlusion handling and processing time.

**Keywords:** occlusion handling, video based augmented reality, hidden surface removal, kinect.

## 1 Introduction

Augmented Reality (AR) could be the answer for the growing demand of new user interfaces, in which space is not restricted to a screen and controls become unnecessary. AR adds 3D virtual objects (3D VOs) to a real scene, allowing the superposition of computer-generated graphics on real world scenes, in such a way that both look as a part of the same 3D scene [6]. In this way, a user can receive useful information in real time and in the most adequate place (real environment) and be guided in a determined task. Nowadays several applications in areas such as medicine, entertainment, education, architecture, among others, use AR; soon, even more areas will benefith from it.

An important task in order to create a synthetic *realistic* scene, is to align virtual and real objects in two ways: geometrical (spatial precision) and semantical (graphic credibility) [4]. Spatial precision requires the *3D VOs* to be appropriately registered in the real world, which means that they always must be in the right position and orientation with respect to the world. On the other hand, graphic credibility refers to the scene realism, i.e., the illusion of both elements, virtual and real, coexisting at the same spatiotemporal place. Graphic credibility has two main branches: the photo-realism,

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 447–454, 2013.

wich deals with illumination effects such as shadows and reflections, and occlusion handling, which requires that the *3D VOs* are correctly occluded by real-world elements.

The occlusion problem consists of determining which objects, real or virtual, are visible from a given vision angle and, based on that, hiding certain elements from the user view, all of this considering a 3D environment. Occlusion occurs when an object close to the user hides a further object on the same vision line.

According to some of the most recent works in the literature [3], [1], the use of depth information about the real world has lead to better results in the occlusion handling in AR. In this approach several stereo vision systems and 3D cameras have been used to calculate the distances in the real world. Despite leading to better results, the use of these technologies brings some problems: the intensity image and the depth map are not aligned and have low resolution, some stereo vision systems require excesive processing time and are inadequate to be used in real time, equipment is expensive and unaccesible to most users and, finally, some systems rely on big hardware and are not adequate to mobile configurations.

In this work we propose a strategy based on depth information and visual markers tracking. Our method combines the well known framework *ARToolKit* with the *Kinect* sensor to deal with, respectively, the positioning and AR occlusion issues. Moreover, we add a processing stage to correct the depth map and present our related conclusions.

This work improves the existent related works in the following aspects: 1) the use of the *Kinect* sensor allows us to work in real time environments and mobile configurations with resolutions above $640 \times 480$ pxs; 2) the tracking of visual markers a-llows correct registration of virtual objects (position and orientation); and 3) the parallel implementation (tracking and depth improvement) makes possible to work in real-time applications.

The rest of this work is organized as follows: section 2 shows a summary of related work, section 3 introduces our method and each of its parts, section 4 decribes the methodology we used to perform experiments and the obtained results; finally, section 5 presents our conclusions and future work.

## 2 Related Work

The first efforts to solve the occlusion problem in *AR* are focused on the segmentation of images. The main idea of this approach is to segment the real object that must occlude the *VO*; then, the *VO* is drawn on the real scene and, finally, the previously segmented region is put on top, in such a way that the real object occludes the *VO*. Some works that use this approach are [6] and [7].

In recent years, with the emergence of stereo vision systems and TOF cameras, the use of depth information has become the dominant approach to occlusion handling. A method to solve the occlusion problem in *VAR* is proposed in [2]. The authors use stereo vision and contour matching to calculate the depth of the objects in the foreground (user hands). Due to the high processing cost, this work focuses in the particular case in which the user hands must occlude the *VO* and viceversa. In addition, as a result of the approach used to segment the user hands, the method is not appropiate to work with occlusive objects with different color and texture.

Zhu et al. [3] propose a probabilistic approach to handle occlusion in *AR* using depth information obtained from a stereo vision system. Instead of using only the estimated depth, their method combines depth, color and neighborhood information, therefore reducing the noise inherent to the stereo pair. In order to accelerate the matching process between images, the authors incorporate a color quantization method; they also introduce Mixed Gaussian Kernels to describe objects of interest and to background subtraction. Finally, the estimated depth is used, together with a color addition method and neighborhood information, to establish occlusion relations between the objects of interest (only in the forefront). Due to the high computational cost of this approach, the authors focus on handling occlusions by certain pre-defined objects of interest, thus disabling the proposed method to work on dynamic scenarios.

Dong et al. [1] propose an algorithm for occlusion handling using depth obtained by a high-resolution TOF camera (*PMD CamCube 3.0*) and technology based on hardware to supress the background illumination. The authors add a second camera in order to obtain a RGB image; the first task performed is the alignment of both images. Then, to handle the occlusion, they use the principle of hidden surfaces removal to draw on the scene only the parts of the *VO* that are not occluded by a real object. The main drawback of this work is the alignment stage between the RGB image and the depth map, which produces over-occluding VOs leaving blank gaps around the occlusor real objects. Furthermore, the use of a specialized high-cost camera makes this work inaccessible to most users.

## 3 Occlusion Handling

Considering the drawbacks of the works described in the previous section, we focus on a method that covers differents cases: the ability of handling occlusion relationships between several objects despite their shape, size or color; dynamic cases in which the scene changes over time; and the use of technologies accessible to the majority of users. Furthermore, we consider both, geometrical and semantic, aspects. In the former, we use visual marker tracking to align the *VO*; in the later, we handle occlusion in real time.

In this section we describe the three main stages of our proposed method. It is important to point out that the first two stages take place in a parallel way, speeding up the processing time and making our method suitable for real-time applications.

### 3.1 Markers Recognition

The stage of markers tracking makes use of the framework *ARToolKit*[1], with modifications that include the integration of *Kinect* as the video input device, and the implementation of the function *Automatic thresholding* based on *ARToolKitPlus* [2]. The *tracker* is initialized through a calibration file for *Kinect*, obtained in previous offline calibration stage and feeded with the RGB image delivered by *Kinect*. In the final recognition stage, the view model matrix is obtained and applied to the VO when this is rendered.

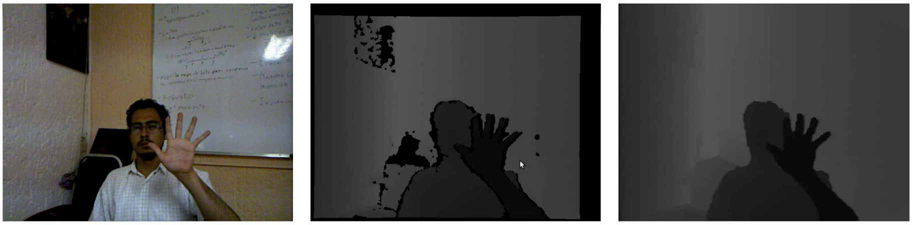
[1] http://www.hitl.washington.edu/artoolkit/

[2] http://handheldar.icg.tugraz.at/artoolkitplus.php

### 3.2 Depth Improvement Stage

The method we propose to handle occlusion is based on the evaluation of distances between real and virtual objects; for this reason, a precise depth map is required. Fig. 1(a) y (b) shows the original images delivered by the *Kinect* sensor. As we can see in 1(b), the depth map is contamined with noise. These black pixels represent blind spots to the sensor, in which it was not able to estimate the distance to the real objects; later, in the experiments section, we show the negative impact of this issue on the occlusion handling.

During the correction stage we evaluated different methods based on an inpaiting technique to correct the depth map. Fig. 1(c) shows the results obtained with the *inpainting telea* algorithm [5], which was the one with the best results when the corrected depth map was used for occlusion handling. In general, the methods based on inpainting techniques estimate the missing data using the neighborhood's information by expanding regions and they do not take into account the RGB image. Therefore, as shown in Fig. 1(c), the expanded regions in the depth map do not correspond to the original RGB image and there is a gap if both images are superimposed (we show this in the first experiment, in the section 4).

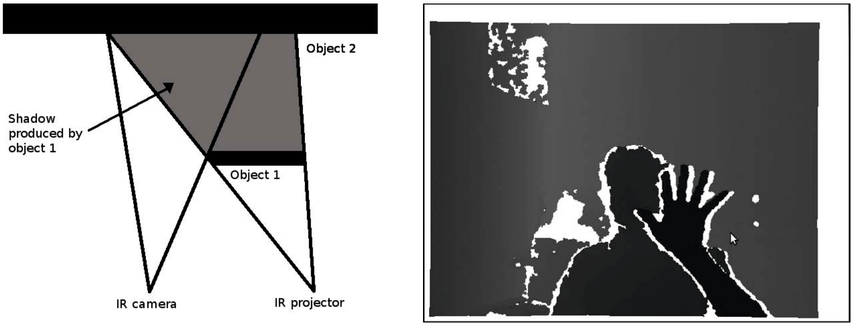


(a) RGB image (b) Original depth map (c) Corrected depth map

**Fig. 1.** Correction of the depth map: this figure shows the original images delivered by the sensor (Fig. 1(a) and 1(b)) and the depth map corrected by using the *inpainting telea* algorithm (Fig. 1(c)). Note that the black holes in (b), are calculated in (c) by using neighborhood information (with radius equal to 5).

After the black holes in the depth map are calculated, we can still appreciate a lack of alignment between the *RGB* image (Fig. 1(a)) and the depth map (Fig. 1(c)). In Fig. 3(b) the result in occlusion handling is shown when using the corrected map. We can see that the lack of alignment between the images is more evident when the scene is augmented. Analyzing the previous images we conclude that the main problem in the depth map is the effect we call *shadow effect*, which can be seen in Fig. 2(a). In this image there is a separation between the projector and the *IR* camera, and this results in a blind point between what is projected and what is seen. As a consequence, a black hole appears in object 2 that looks like a shadow of the object 1.

Taking into account this situation, we want to supress the shadow effect, since it belongs to object 2 (the furthermost object) and, therefore, it should not affect in an occlusion by object 1 (object in the foreground). Fig. 2(b) shows the depth map once



(a) Shadow effect explanation

(b) Shadow effect suppression

**Fig. 2.** Shadow effect generated by the *Kinect* sensor

the black holes have been removed (they go from being in the foreground to being in the background). Fig. 3(c) shows the occlusion handling method using this idea.

### 3.3 Depth Buffering

Occlusion handling is performed applying the technique known as hidden-surface removal, which consists on removing object parts that are obscured by a closer object.

The distances obtained through the *Kinect* are in the range of $[0 - N]$ mm. Considering that the optimal range for the correct *Kinect* operation is between $50$ and $4000$mm, all the values that exceed $4000$mm are scalated. Before they are written in the depth buffer, the vertices are transformed to the clip coordinates through the equation

$$cc = \frac{rd * (f + n)}{f - n} - \frac{2 * f * n}{f - n}, \tag{1}$$

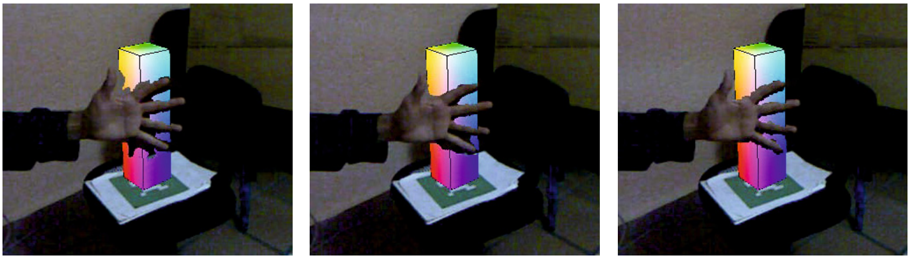
then, they are normalized $ndc = \frac{cc}{rd}$ and forced to be in the range $[-1, 1]$. Finally, the values are transformed to the range $[0 - 1]$ by $fd = \frac{ndc+1}{2}$. Where $rd$ is the distance obtained using the *Kinect* sensor (*raw data*), $n$ and $f$ are, respectively, the near and far plane projections, $cc$ is the clip coordinates after the projection matrix, $ndc$ represents the normalized device coordinates and $fd$ is the final depth written on the depth buffer.

These distances are stored in the depth buffer and represent the distances of the real scene. When a new object is drawn on the screen, its depth is compared against the depth previously stored in the depth buffer, and only if the new object's depth is less, the VO is actually drawn.

## 4 Experiments

To perform our experiments, we built a system that integrates the acquisition, processing and display of an image. The experiments were performed inside a room (indoor configuration), where a user moved freely across the room and interacted with the virtual content. All the experiments have the purpose of evaluating the robustness of the proposed strategy to solve occlusion relationships in *VAR*, and the impact of the quality of the depth map on the occlusion handling.

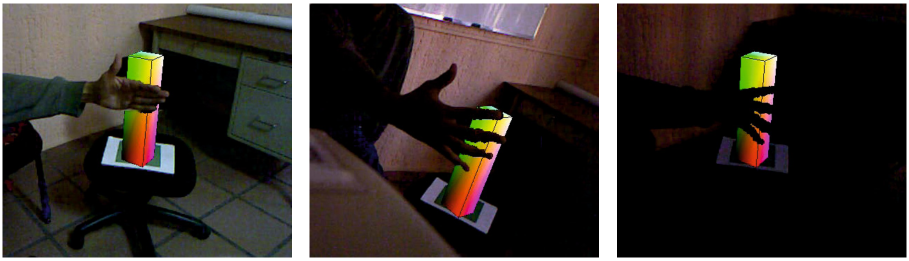
**Experiment 1: Impact of the Depth Map.** Fig. 3 shows the results of the stage of the depth map correction. Despite correcting the depth map, when VO occlusion is handled there is a lack of alignment with the RGB image; this translates into a poor occlusion (Fig. 3(b)). Fig. 3(c) shows the results of removing the shadow effect from the depth map; this technique gave better results.



(a) Original depth map (b) Depth map improved by *inpainting* (c) ***Shadow effect* removed**

**Fig. 3.** Impact of the depth map on occlusion handling: (a) occlusion handling using the depth map delivered by the sensor, without processing; (b) occlusion handling using the depth map corrected by the *inpainting telea* algorithm; (c) occlusion handling with supression of the shadow effect

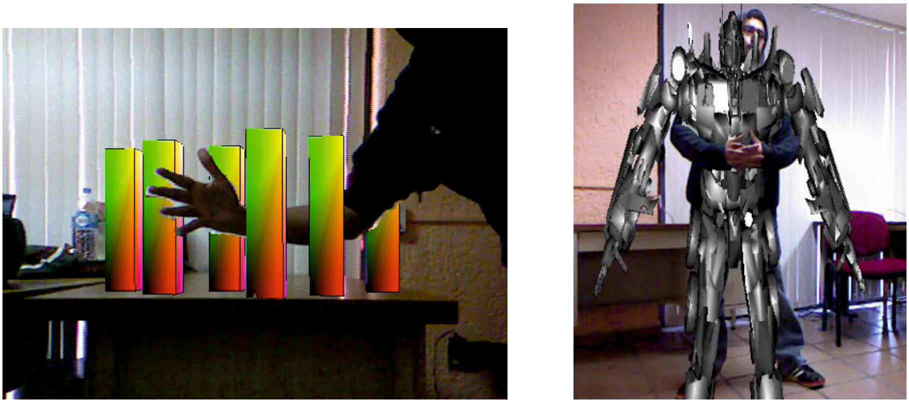
**Experiment 2: Variant Lighting Conditions.** In Fig. 4 we observe the handling occlusion under three scenaries with different lighting. In the three scenes we can see that the marker was recognized and the *VO* was correctly drawn over the marker and occluded by the user hand.



(a) Ideal lighting (b) Moderate lighting (c) Poor lighting

**Fig. 4.** Occlusion handling under variant lighting conditions

**Experiment 3: Multiple and Deformable Objects.** In this experiment, we worked with multiple and deformable objects. Fig. 5(a) shows the results of the proposed method when working with multiple VOs. In the image we can see the user interacting with virtual content and how the user's arm is occluded by VOs that are closer and occluding *VOs* elements located further away. Also, in this experiment we used more realistic *VOs*, with a bigger size and undefined shapes. Fig. 5(b) shows how the virtual elements are correctly (partially) occluded by the user.



(a) Multiple VOs

(b) Deformable VOs

**Fig. 5.** Occlusion of multiple and deformable virtual objects. Image on the left shows that multiple virtual objects can be added to the scene and the proposed method is able to solve the occlusion relationships between them and the real objects. Image on the right shows partial occlusion of deformable virtual objects.

### 4.1 Discussion of the Results

The experiments showed that our proposed method can handle occlusion of deformable objects, multiple objects (occlusive and occluded) and work under different lighting conditions. It was also shown that the method is appropiate to work in environments demanding real-time response; the performed experiments reached a processing rate over 30 f/s.

Some of the drawbacks found in the use of the *Kinect* sensor are (1) the inability to handle occlusion with transparent or refracting objects, due to the fact that the sensor is not able to solve the object distance; and (2) a great sensibility to sun light, therefore the method is only appropiate for indoor configurations.

## 5 Conclusions

We have explored the use of a motion sensing input device, the *Kinect* sensor, in an Augmented Reality task. The experiments showed the feasibility of this method to build augmented scenes properly occluded under indoor configurations and real-time; this could lead to new tasks for mobile robots, for example, by including AR in their navigation tasks. Furthermore, the obtained results are comparable to those of other works in the state of the art that use stereo vision systems and depth cameras with high economical and computational cost.

In our future work we are interested in exploring the use of algorithms that take into account the RGB image to correct the depth map, in such a way that the map can attain higher precision while maintaining the real-time requirement, and the removal of visual markers by calculating, instead, flat surfaces like tables, floors, walls, etc. Moreover, we would like to investigate the construction of a 3D model of the environment, so that we can keep virtual objects registered even when the camera angle changes.

Considering that we obtained good quality results, in different scenarios, with a low computational cost, we can say that the *Kinect* sensor is suitable for handling occlusions in AR applications.

**Acknowledgments.** This work was done under partial support of CONACyT-Mexico (scholarship 301754).

## References

1. Dong, S., Feng, C., Kamat, V.R.: Occlusion Handling Method for Ubiquitous Augmented Reality Using Reality Capture Technology and GLSL. In: Proceedings of the 2011 ASCE International Workshop on Computing in Civil Engineering, Reston, VA, pp. 494–503 (2011)
2. Li, L., Guan, T., Ren, B.: Resolving occlusion between virtual and real scenes for augmented reality applications. In: Proceedings of the 12th International Conference on Human-Computer Interaction: Interaction Platforms and Techniques, Beijing, China, pp. 634–642 (2007)
3. Zhu, J., Pan, Z., Sun, C., Chen, W.: Handling Occlusions in Video-Based Augmented Reality Using Depth Information. Journal of Animation and Virtual Worlds: Wiley Online Library Computer 21, 509–521 (2010)
4. Zhu, J., Pan, Z.: Occlusion registration in video-based augmented reality. In: Proceedings of The 7th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, vol. 10, pp. 1–6 (2008)
5. Telea, A.: An Image Inpainting Technique Based on the Fast Marching Method. Journal of Graphics, Gpu, and Game Tools 9, 23–34 (2004)
6. Lepetit, V., Berger, M.-O.: Handling occlusion in augmented reality systems: a semi-automatic method. In: Proceedings EEE and ACM International Symposium on Augmented Reality (ISAR 2000), pp. 137–146 (2000)
7. Fischer, J., Bartz, D., StraBer, W.: Occlusion Handling for Medical Augmented Reality using a Volumetric Phantom Model. In: Journal of VRST 2004 Proceedings of the ACM Symposium on Virtual Reality Software and Technology, pp. 174–177 (2004)

# Object Tracking in Nonuniform Illumination Using Space-Variant Correlation Filters

Víctor Hugo Díaz-Ramírez[1], Kenia Picos[1], and Vitaly Kober[2]

[1] Instituto Politécnico Nacional - CITEDI, Ave. del Parque 1310, Mesa de Otay,
Tijuana B.C. 22510, Mexico
[2] Department of Computer Science, Division of Applied Physics, CICESE, Carretera
Ensenada-Tijuana 3918, Zona Playitas, Ensenada B.C. 22860, Mexico
vdiazr@ipn.mx, kpicos@citedi.mx, vkober@cicese.mx

**Abstract.** A reliable system for recognition and tracking of a moving target in nonuniformly illuminated scenes is presented. The system employs a filter bank of space-variant correlation filters adapted to local statistical parameters of the observed scene in each frame. When a scene frame is captured, a fragment of interest is constructed in the frame around predicted location of the target based on a kinematic model. The fragment is firstly pointwise processed to correct the illumination. Afterwards, the state of the target is estimated from the restored fragment by employing a bank of space-variant correlation filters. The performance of the proposed system in terms of object recognition and tracking is tested with nonuniformly illuminated and noisy scenes. The results are compared with those of common techniques based on correlation filtering.

## 1 Introduction

Nowadays, object recognition attracts research interests due to the need of developing imaging systems to improve activities such as video surveillance, vehicle navigation, object tracking, among others. Object recognition consists in identification of the target within a observed scene and in estimation of the target's exact coordinates. When a target moves across an environment, the appearance of the target with respect to the observer varies with time. Actually, target tracking consists in estimation of the target trajectory in the observed scene while the object moves. Target tracking can be solved by detecting the object in successive frames and by finding the correspondence between object states across scene frames. Commonly target tracking is performed by employing feature-based methods and state-space models. This approach yields good results when the target suffers from occlusions and geometrical modifications such as rotation and scaling. However, when the target exits and reenters to the observed scene and when the scene is degraded by additive and high cluttering background noise feature-based methods face some difficulties. A detailed review of tracking algorithms can be found in [1]. An attractive option for target recognition is given by correlation filtering. Correlation filters possess a good

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 455–462, 2013.

mathematical background, and they can be implemented by exploiting massive parallelism either in hybrid opto-digital correlators [2,3] or in digital hardware such as graphic processing units (GPU) [4]. A correlation filter is a linear system where the coordinates of the system output maximum are estimates of the target coordinates in the observed scene [5,6]. Correlation filters can recognize objects in cluttered and noisy environments, even when targets suffer from geometrical distortions [7]. In this sense, the problem of target tracking can be addressed by applying correlation filters to multiple frames. Recently, several proposals have been suggested to perform target tracking with the help of correlation filters [3]. In this work, we propose a reliable system for recognition and tracking of a moving target in nonuniformly illuminated scenes using a filter bank of space-variant correlation filters. The frequency response of the filters are varied accordingly to local statistical parameters of the input signal in each frame. First, the proposed system performs a pointwise illumination correction to the input frame. Next, the target is detected from the restored frame by analyzing the correlation peaks obtained at the outputs of the filter bank. Then, the system predicts the state of the target for subsequent frame, and based on the prediction creates a fragment of interest in the input frame and modifies the number of filters in the bank using predicted orientation of the target in the current frame. Both location and orientation predictions are calculated by analyzing current and past state estimates and by taking into account a two-dimensional motion model. The resultant system is able to track a moving target in nonuniform illumination with reduced false alarms probabilities by focusing the processing only on a small fragment. The paper is organized as follows. Section 2 presents the approach used for target recognition in nonuniformly illuminated and noisy scenes. Section 3, explains the system proposed for object recognition and tracking. Section 4 presents the results obtained with the proposed system by testing its performance in nonuniformly illuminated scenes. Finally, section 5 summarizes our conclusions.

## 2 Recognition of a Target in Nonuniformly Illuminated and Noisy Scenes

Let $f(x,y)$ be an input scene composed by a target $t(x,y)$ located at unknown coordinates $(\alpha,\beta)$ and embedded into a disjoint background $b(x,y)$. The scene is assumed to be corrupted by a nonuniform illumination function $d(x,y)$ and with zero-mean additive noise $n(x,y)$. The input scene is expressed by

$$f(x,y) = [t(x-\alpha,y-\beta) + \bar{w}(x-\alpha,y-\beta)b(x,y)]\,d(x,y) + n(x,y), \qquad (1)$$

where $\bar{w}(x)$ is the inverse region of support of the target given by unity outside the target area and zero elsewhere. We assume that $d(x,y)$ is a slow varying function, which is approximately constant in a small region (for instance, the region of support of the target $w(x,y)$). Note that this is the case of Lambertian surfaces. In order to correct the illumination of the input frame we perform the following pointwise processing

$$\hat{f}(x,y) = r_{x,y}f(x,y) + s_{x,y}, \qquad (2)$$

where $r_{x,y}$ and $s_{x,y}$ represent unknown restoration coefficients. The mean squared error (MSE) between the restored frame $\hat{f}(x,y)$ and the reference image of the target $t(x,y)$, is given by

$$MSE_{\alpha,\beta} = \sum_{x,y \in w}\sum \left(r_{\alpha,\beta} f(x+\alpha, y+\beta) + s_{\alpha,\beta} - t(x,y)\right)^2 . \tag{3}$$

By minimization of the $MSE_{\alpha,\beta}$, we get

$$s_{\alpha,\beta} = \mu_t - r_{\alpha,\beta}\mu_f(\alpha,\beta), \tag{4}$$

and

$$r_{\alpha,\beta} = \frac{\frac{1}{N_w}\sum_{x,y \in w}\sum t(x,y) f(x+\alpha,y+\beta) - \mu_t \mu_f(\alpha,\beta)}{\mu_{f^2}(\alpha,\beta) - \mu_f^2(\alpha,\beta)}, \tag{5}$$

where $N_w$ is the number of signal elements inside $w(x,y)$, $\mu_t = \frac{1}{N_w}\sum_{x,y \in w}\sum t(x,y)$, $\mu_f(\alpha,\beta) = \frac{1}{N_w}\sum_{x,y \in w}\sum f(x+\alpha,y+\beta)$ and $\mu_{f^2}(\alpha,\beta) = \frac{1}{N_w}\sum_{x,y \in w}\sum f^2(x+\alpha,y+\beta)$. Note that $\hat{f}(x,y)$ in Eq. (2) represents the input frame with approximately uniform illumination, i.e.,

$$\hat{f}(x,y) \approx t(x-\alpha,y-\beta) + \bar{w}(x-\alpha,y-\beta) b(x,y) + \tilde{n}(x,y), \tag{6}$$

where $\tilde{n}(x,y)$ is a nonstationary noise process.

### Recognition of a Target in Additive and Nonoverlapping Noise

Here the goal is to recognize and to precisely estimate the location of the target from the nonoverlapping signal model of Eq. (6). In this case, the optimum filter with respect to the signal to noise ratio (SNR) and to the minimum variance of target's location error (LE) is the Generalized Matched Filter (GMF), whose frequency response is given by [8,6]

$$H^*(\mu,\nu) = \frac{T(\mu,\nu) + \mu_b \bar{W}(\mu,\nu)}{S_{b0}(\mu,\nu) \otimes \left|\bar{W}(\mu,\nu)\right|^2 + S_n(\mu,\nu)}. \tag{7}$$

In Eq. (7), $\mu_b$ and $S_{b0}(\mu,\nu)$ represent the mean value of the background $b(x,y)$ and the power spectral density of $b_0(x,y) = b(x,y) - \mu_b$, respectively. The terms $T(\mu,\nu)$, $\bar{W}(\mu,\nu)$ and $S_n(\mu,\nu)$ are the Fourier transform of the target, the Fourier transform of $\bar{w}(x,y)$ and the power spectral density of $n(x,y)$, respectively. It is important to realize that for real applications the terms $T(\mu,\nu)$ and $\bar{W}(\mu,\nu)$ are a-priori known. Nevertheless, the terms $\mu_b$, $S_{b0}(\mu,\nu)$ and $S_n(\mu,\nu)$ are generally unknown and must be estimated.

### Estimation of Nonoverlapping Noise Parameters

Assume that the target $t(x,y)$ is located inside a small fragment $\hat{f}_r(x,y)$ of the input frame and is embedded into the background $b_r(x,y)$. The mean value of the scene fragment can be computed as $\mu_{\hat{f}_r} = \frac{1}{N_{\hat{f}_r}}\sum_{x,y \in \hat{f}_r}\sum \hat{f}_r(x,y)$, where $N_{\hat{f}_r}$ is the number of pixels in the fragment. Since the target is known and is contained in the fragment, the mean value of the background $b_r(x,y)$ can be estimated as

$$\hat{\mu}_{b_r} = \frac{\mu_{\hat{f}_r} N_{\hat{f}_r} - \mu_t N_w}{N_{\hat{f}_r} - N_w}. \tag{8}$$

The sample variance of $b_r(x,y)$ can be computed by $\sigma^2_{b_r} = \frac{1}{N_{b_r}} \sum\sum_{x,y\in b_r} b^2_r(x,y) - \hat{\mu}^2_{b_r}$. By noticing that for the disjoint model $\sum\sum_{x,y\in b_r} b^2_r(x,y) = \sum\sum_{x,y\in \hat{f}_r} \hat{f}^2_r(x,y) - \sum\sum_{x,y\in w} t^2(x,y)$, and with the help of Eq. (8) the local variance of the background is estimated by

$$\hat{\sigma}^2_{b_r} = \frac{1}{N_{\hat{f}_r} - N_w} \left( \sum\sum_{x,y\in \hat{f}_r} \hat{f}^2_r(x,y) - \sum\sum_{x,y\in w} t^2(x,y) \right) - \left( \frac{\mu_{\hat{f}_r} N_{\hat{f}_r} - \mu_t N_w}{N_{\hat{f}_r} - N_w} \right)^2. \tag{9}$$

The parameters $\hat{\mu}_{b_r}$ and $\hat{\sigma}^2_{b_r}$ are used to estimate the power spectral density $S_{b0}(\mu,\nu)$ required in Eq. (7). This is done by a separable exponential model of the covariance function, as follows:

$$\hat{S}_{b0}(\mu,\nu) = \sum\sum_{x,y\in \hat{f}_r} \hat{\sigma}^2_{b_r} \rho_x^{|x|} \rho_y^{|y|} \exp\left[ -i2\pi\left(\mu x + \nu y\right)/N_{f_r} \right], \tag{10}$$

where $\rho_x$ and $\rho_y$ are the correlation coefficients of the background function in $x$ and $y$ directions.

### Estimation of Additive Noise Parameters

Consider that the fragment $f_r(x,y)$ is corrupted by zero-mean additive white Gaussian noise $n_r(x,y)$. Assume that $f_r(x,y)$ and $n_r(x,y)$ are independent. The autocorrelation function of the noisy fragment can be approximated by $R(x,y) \approx R_{f_r}(x,y) + R_n(x,y)$, where $R_n(x,y) = \sigma^2_n \delta(x,y)$ is the autocorrelation function of the noise, and $R_{f_r}(x,y)$ is the autocorrelation function of the fragment. Note that the noise variance $\sigma^2_n$ can be estimated by $\hat{\sigma}^2_n = R(0,0) - R_{f_r}(0,0)$, where $R_{f_r}(0,0)$ is unknown. We see that $R_n(x,y) = 0$, $\forall$ $(x,y) \neq 0$. Thus, the values $\{R(x,y) | (x,y) \neq 0\}$ can be used as estimates of $R_{f_r}(x,y)$. To estimate the noise variance $\sigma^2_n$ we can calculate $R_{f_r}(0,0)$ by means of a extrapolation of the values of $R(x,y)$ close to the origin. In this manner, the power spectral density of the noise can be approximated by $S_n(\mu,\nu) = \hat{\sigma}^2_n$.

## 3 Proposed System for Object Recognition and Tracking in Nonuniform Illumination

Now we describe the proposed system for object recognition and tracking in nonuniform illumination. Let us consider the optical setup shown in Fig. 1 in where a target moves in horizontal direction of the two-dimensional plane, and it is under the influence of a illumination source. At time $t_k$, the camera captures

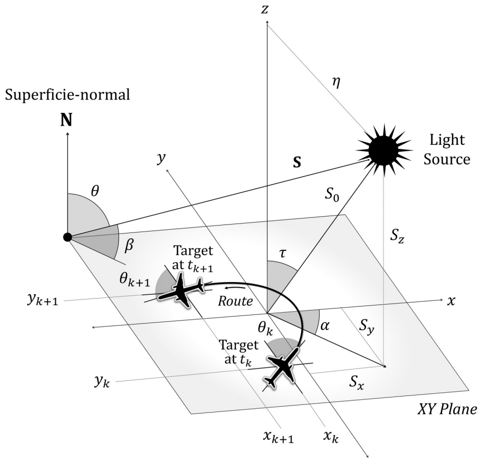


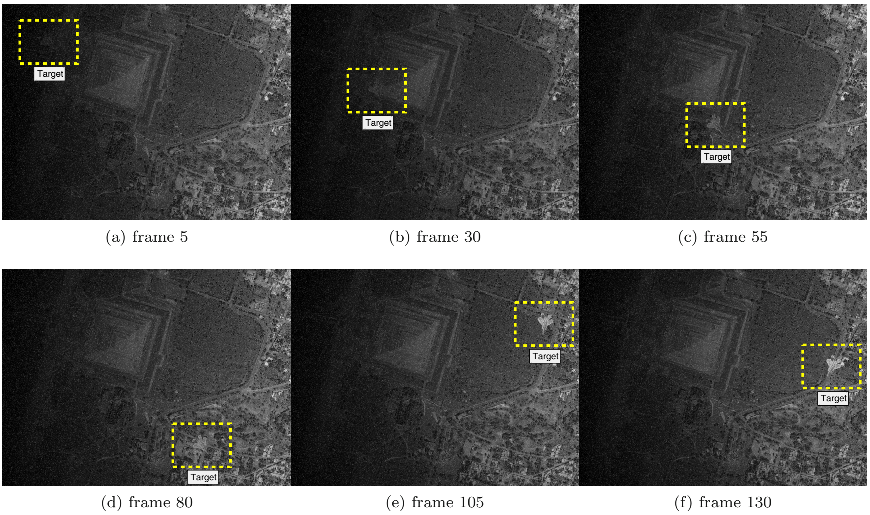
**Fig. 1.** Optical setup for target tracking in nonuniform illumination

a scene-frame containing the target with a orientation angle $\theta_k$. The target is embedded into the background at the coordinates $(\alpha_k, \beta_k)$. Next at time $t_{k+1}$ the target moves to a new position with coordinates $(\alpha_{k+1}, \beta_{k+1})$ and the orientation angle has changed from $\theta_k$ to $\theta_{k+1}$. We want to estimate the sequence of target positions $(\alpha_k, \beta_k)$ and orientation angles $\theta_k$ (sequence of states) as a function of time $\{t_k = k\Delta_t | k = 1, 2, \dots\}$, where $\Delta_t$ is the sampling interval. The state of the target in time $t_k$ is represented by a state vector $\mathbf{s}_k = [\alpha_k, \beta_k, \theta_k]^T$. The operation steps of the proposed tracking system are summarized below.

- Step 1: Read a scene frame $f_k(x, y)$ from the input observed sequence.
- Step 2: Correct illumination of the frame using Eqs. (2), (4) and (5).
- Step 3: Estimate noise parameters (see Eqs. (8) and (10)) and design a filter bank using Eq. (7).
- Step 4: Process the corrected frame with the filter bank and find the correlation plane $c(x, y)$ with the highest discrimination capability (DC). The DC is the ability of a filter to distinguish among a target and unwanted objects; it is defined by [6]

$$DC = 1 - \frac{\left|c^b\right|^2}{\left|c^t\right|^2}, \tag{11}$$

  where $c^b$ is the value of the maximum correlation sidelobe in the background area and $c^t$ is the value of the correlation peak due to the target.
- Step 5: Estimate the target coordinates as $\left(\hat{\alpha}_k, \hat{\beta}_k\right) = \arg\max_{x,y}\left\{|c(x,y)|^2\right\}$, and the orientation angle as $\hat{\theta_k} = 2\,(r-1)$, where $r$ is the index of the filter in the bank which detects the target with the highest DC. Set the current state of the target as $\hat{\mathbf{s}}_k = \left[\hat{\alpha}_k, \hat{\beta}_k, \hat{\theta}\right]^T$.
- Step 6: Predict subsequent state vector $\hat{\mathbf{s}}_{k+1} = \left[\hat{\alpha}_p, \hat{\beta}_p, \hat{\theta}_p\right]^T$ from $\hat{\mathbf{s}}_k$ by characterization of the motion behavior of the target. This behavior is characterized by the following kinematic equations [9]:



(a) frame 5 (b) frame 30 (c) frame 55

(d) frame 80 (e) frame 105 (f) frame 130

**Fig. 2.** Examples of nonuniformly illuminated scene frames corrupted with 20 dB SNR additive noise

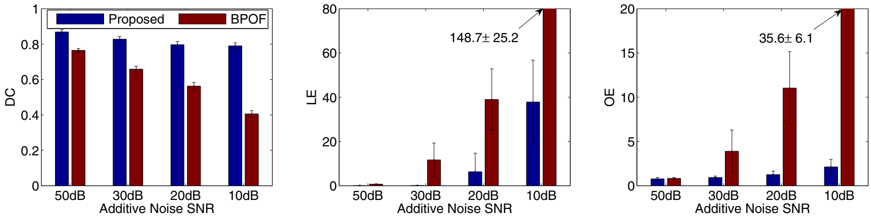
$$
\begin{aligned}
\alpha_{k+1} &= \alpha_k + \frac{\sin\left(\omega_k \mathit{\Delta}_t\right)}{\omega_k}\dot{\alpha}_k - \frac{1-\cos\left(\omega_k \mathit{\Delta}_t\right)}{\omega_k}\dot{\beta}_k + a_{\alpha,k}\frac{\mathit{\Delta}_t^2}{2},\\
\beta_{k+1} &= \beta_k + \frac{1-\cos\left(\omega_k \mathit{\Delta}_t\right)}{\omega_k}\dot{\alpha}_k + \frac{\sin\left(\omega_k \mathit{\Delta}_t\right)}{\omega_k}\dot{\beta}_k + a_{\beta,k}\frac{\mathit{\Delta}_t^2}{2},\\
\dot{\alpha}_{k+1} &= \cos\left(\omega_k \mathit{\Delta}_t\right)\dot{\alpha}_k - \sin\left(\omega_k \mathit{\Delta}_t\right)\dot{\beta}_k + a_{\alpha,k}\mathit{\Delta}_t,\\
\dot{\beta}_{k+1} &= \sin\left(\omega_k \mathit{\Delta}_t\right)\dot{\alpha}_k - \cos\left(\omega_k \mathit{\Delta}_t\right)\dot{\beta}_k + a_{\beta,k}\mathit{\Delta}_t,\\
\omega_{k+1} &= \omega_k + a_{\omega,k}.
\end{aligned}
\tag{12}
$$

The variables $\alpha_k$ and $\beta_k$ denote the position of the target in Cartesian coordinates, $\dot{\alpha}_k$ and $\dot{\beta}_k$ are velocity components in $\alpha$ and $\beta$ directions, and $\omega_k$ is the target's angular rate. Furthermore, $a_{\alpha,k}$ and $a_{\beta,k}$ are random variables representing acceleration components (due to turbulence) in $\alpha$ and $\beta$ directions and $a_{\omega,k}$ is the angular acceleration. The position of the target in a subsequent time, is predicted by substitution of the estimated position $\left(\hat{\alpha}_k, \hat{\beta}_k\right)$, the estimated velocity components $\left(\hat{\dot{\alpha}}_k, \hat{\dot{\beta}}_k\right)$, and the estimated turn rate $\hat{\omega}_k$ (calculated from current and past frames) into the state space model in Eq. (12), and then by taking the expected value.

- Step 7: Read a new frame $f_k(x, y)$ from the scene and create a fragment of interest according to state prediction. Go to STEP 2.

## 4 Results

Here, results obtained with the proposed system for target tracking in nonuniformly illuminated scenes are presented. The results are given in terms of



(a) (b) (c)

**Fig. 3.** Performance of tracking systems with 95% confidence in terms of: (a) DC, (b) LE, and (c) OE

recognition performance and tracking accuracy. The obtained results are compared with those obtained with a system based on binary phase-only (BPO) filters [3]. The recognition performance is measured in terms of the DC, whereas tracking accuracy is characterized by the precision of estimates carried out for the target state across scene frames. The accuracy in location estimation is characterized by the LE, which is given by [6]

$$LE = \left[ \left( x^q - \hat{x^q} \right)^2 + \left( y^q - \hat{y^q} \right)^2 \right]^{1/2}, \tag{13}$$

were $(x^q, y^q)$ and $\left(\hat{x^q}, \hat{y^q}\right)$ are the exact and estimated coordinates of location of the target, respectively. The accuracy of estimation of the orientation angle is characterized by the orientation error (OE), defined by

$$OE = \left| \phi^q - \hat{\phi^q} \right|, \tag{14}$$

where $\phi^q$ and $\hat{\phi^q}$ are the true and estimated orientation angles, respectively. The units for LE and OE metrics are pixles and degrees, respectively. In our experiments we use a sequence of 200 nonuniformly illuminated scene frames with 800×600 pixels. Figure 2 shows examples of various scene frames in environment of 20 dB SNR additive noise. The target is an airplane which can move and rotate freely in the horizontal plane. With 95% confidence, the results for 200 scene frames obtained with proposed and BPO systems are presented in Fig. 3. We can see that the proposed system yields the best results in all the cases. Observe from Fig. 3(a) that the proposed system yields DC values close to unity even in highly noisy conditions of 10 dB SNR. Furthermore, we see that the

**Table 1.** Detection performance of tracking systems in 200 scene frames

| | Additive Noise SNR | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Decision** | 50dB | 30dB | 20dB | 10dB | 50dB | 30dB | 20dB | 10dB |
| detected | 200 | 200 | 198 | 187 | 200 | 180 | 136 | 104 |
| not detected | 0 | 0 | 2 | 13 | 0 | 20 | 64 | 96 |
| % of error | 0% | 0% | 1% | 6.5% | 0% | 10% | 32% | 48% |
| | Proposed system | | | | BPO system | | | |

proposed system can estimate with a good precision the location of the target in noisy conditions of 50 dB, 30 dB, and 20 dB SNR (see Fig. 3(b)). Additionally, from Fig. 3(c) we see that the proposed system estimates with a good accuracy the orientation angle of the target even in highly noisy conditions. The BPO system, yields good results in terms of the DC for 50 dB, 30 dB, and 20 dB SNR. Furthermore, this system yields high LE and OE values for 20 dB and 10 dB SNR. Table 1 shows the recognition performance of the tracking systems in 200 scene frames. The proposed system yields no detection errors for 50 dB and 30 dB SNR, and yields only two false detections for 20 dB SNR, and yields thirteen false detections for highly noisy conditions of 10 dB SNR. The BPO system yields good results for 50 dB SNR, however when the SNR decreases the number of false detections increases.

## 5 Conclusions

A tracking system for nonuniformly illuminated scenes was presented. The system employs a filter bank of time variant correlation filters to estimate the state trajectory of a moving target in a sequence of images. By incorporation of a prediction stage the system creates a fragment of interest in the observed frame and modify the number of required filters in the bank to estimate the state of the target in current frame. By means of computer simulations we showed that the proposed system yields a superior performance in terms of tracking accuracy comparing with recent state of the art tracking systems based on correlation filtering.

## References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. 38(4) (2006)
2. Diaz-Ramirez, V.H., Kober, V.: Adaptive phase-input joint transform correlator. Appl. Opt. 46(26), 6543–6551 (2007)
3. Manzurv, T., Zeller, J., Serati, S.: Optical correlator based target detection, recognition, classification, and tracking. Appl. Opt. 51, 4976–4983 (2012)
4. Ouerhani, Y., Jridi, M., Alfalou, A., Brosseau, C.: Optimized pre-processing input plane GPU implementation of an optical face recognition technique using a segmented phase only composite filter. Opt. Comm. 289, 33–44 (2013)
5. Yaroslavsky, L.P.: The theory of optical method for localization of objects in pictures. In: Wolf, E. (ed.) Progress in Optics, vol. XXXII, pp. 145–201. Elsevier North-Holland (1993)
6. Kober, V., Campos, J.: Accuracy of location measurement of a noisy target in a nonoverlapping background. J. Opt. Soc. Am. A 13(8), 1653–1666 (1996)
7. Aguilar-Gonzalez, P.M., Kober, V.: Design of correlation filters for pattern recognition using a noisy reference. Opt. Comm. 285(5), 574–583 (2012)
8. Javidi, B., Wang, J.: Design of filters to detect a noisy target in nonoverlapping background noise. J. Opt. Soc. Am. A 11, 2604–2612 (1994)
9. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 34(3), 334–352 (2004)

# GPU Based Implementation of Film Flicker Reduction Algorithms*

Martn Piñeyro, Julieta Keldjian, and Alvaro Pardo

Department of Electrical Engineering, School of Engineering and Technologies,
Universidad Catolica del Uruguay
{mpineyro,apardo,jukeldji}@ucu.edu.uy

**Abstract.** In this work we propose an algorithm for film restoration aimed at reducing the flicker effect while preserving the original overall illumination of the film. We also present a comparative study of the performance of this algorithm implemented following a sequential approach on a CPU and following a parallel approach on a GPU using OpenCL.

## 1 Introduction

Visual flicker is one of the most common consequences of degradation in old films. It is the result of global intensity fluctuations between consecutive frames. In some cases flicker can be a local effect too, with regions of the frame experiencing local intensity variations between consecutive frames. Although it may seem a simple problem that could be addressed with traditional intensity normalization techniques, usually this approach is not able to remove the distortion completely.

When film is digitalized by capturing its projection with a digital camera flicker is introduced as a consequence of temporal mismatch between the capture and projection system (the acquisition of frames is unsynchronized with the projectors shutter). Aging is also an important cause of flicker. Not all the frames of a film suffer the same degradation along the time and therefore unintended variations in mean illumination may be encountered. When removing flicker through digital image processing the effects due to digitalization and aging have to be removed while preserving the flicker caused by the mechanical limitations of the equipment originally used to capture the content.

In order to avoid the introduction of new structures as a consequence of the restoration process usually the flicker reduction algorithms use histogram corrections that preserve the geometry of the frames. Also, since flicker is a temporal distortion affecting a sequence of frames, restoration processes aimed at reducing it need to consider multiple frames in order to capture the temporal intensity variations.

## 2 Previous Work

The works in the literature of flicker reduction can be categorized in two types: the ones that apply local methods and the ones that apply global methods.

* ANII FMV200913042 and SticAmsud MMVPSCV.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 463–470, 2013.

In this work we concentrate in algorithms that apply global methods to modify the histogram of each frame of the film in order to smooth the intensity variations across time. The first approach for global intensity modification is based in the use of affine transformations. The corrected frame at time $t$, $\hat{I}(x;t)$, is transformed with the equation $\hat{I}(x;t) = a(t)I(x;t) + b(t)$ where parameters $a(t)$ and $b(t)$ are selected to match the input dynamic range to a desired one. Since flicker varies with time both parameters are time dependent. The values of $a(t)$ and $b(t)$ are selected to reduce the mean intensity variations within the dynamic range. The main difficulty with this solution is that the desired output dynamic range has to be manually given to the algorithm.

Another option for applying global changes to the frames is using histogram modifications (not only affine as in the first case). This allows more general intensity modifications between frames. In [5] the authors proposed matching the histogram of a given frame $I(x;t)$ to the mean histogram in a window centered at time $t$. In [3,2] Delon analyzed the algorithm from [5] and proposed an improvement. The main observation made by Delon is the following: if the histograms of two frames that differ only in a constant are averaged, two unimodal histograms may produce a bimodal one. This simple observation shows the limitations of the algorithm proposed in [5]. To deal with this problem Delon proposed to average the inverse of the cumulative histograms.

For trasforming two images $I_1$ and $I_2$ to have the same histogram both images are transformed with continuous and strictly increasing functions $g_i : [0, 255] \rightarrow [0, N]$. Assuming continuous images the cumulative histograms are defined as: $H_i(q) = \int_0^q h_i(\lambda)d\lambda$, where $h_i(q)$ is the histogram of the image $I_i$. In [2] Delon shows that the following transformations produce images with the same cumulative histogram: $g_1 = \frac{H_1^{-1}+H_1^{-1}}{2} \circ H_2,\ \ g_2 = \frac{H_1^{-1}+H_1^{-1}}{2} \circ H_1$ In this way the final cumulative histogram of each image is: $H_i \circ g_i^{-1} = \left(\frac{H_1^{-1}+H_1^{-1}}{2}\right)^{-1}$. This method was also presented in the discrete case by Cox in [1].

In this work we use this idea but using a weighted average of frames within a temporal window following the ideas discussed in [3]. Given a set of $N$ frames in a given window the inverse average is defined as:

$$H_a = \left( \sum_{i=1}^{N} w_i H_i^{-1} \right)^{-1} .$$

To remove the flicker each frame is transformed with a function $g_i = H_a \circ H_i$.

Using a weighted inverse average inside a window permits to smoothy correct the flicker while allowing variations (in the temporal axis). In that way global brightness variations that could be part of the original material and not the result of any degradation process are preserved. Using the average of all the frames of the scene is not recommended because it will destroy the original content of the scene forcing all frames to have equal histograms. With the proposed approach the temporal variations of the intensity are smoothed and discontinuities avoided while respecting the original variations of the film.

## 3 GPUs and OpenCL

Modern GPUs are very efficient in parallel processing of computer graphics data but also with any other type of data that can take advantage of the parallel nature of the GPUs. At the same time they are very competitive in terms of price. NVIDIA was the first company to introduce a general-purpose programming model with the release of CUDA and recently other companies joined efforts around the OpenCL standard. Although CUDA is widely extended OpenCL has the attractiveness of being a standard supported by many companies. In fact NVIDIA also supports OpenCL just like ATI which is another big player in the field.

One of the goals of this work was to explore the use of OpenCL for image processing. The research group has been working with CUDA so we took this project as a test case to evaluate the suitability of OpenCL for this kind of problem. The development was done using Windows 7, Visual Studio and OpenCL and tested in a ATI Radeon 5650 GPU.

## 4 Algorithm Implementation

**Histogram Computation:** The first stage of the algorithm computes the histogram of the $N$ frames within the temporal window that will be considered in the flicker reduction process. For comparison purposes we implemented a CPU routine using C++ and another one based on the GPU using OpenCL. Since sequential computation of histograms presents no difficulties we will focus on the GPU-based routine.

An OpenCL application consists of two main parts: the host program and the kernels. Initially the host program defines a context (in this work the context consists of a CPU and a GPU). Then it defines a command queue in which commands issued by the CPU are scheduled for execution on the GPU. Subsequently, the host program loads a kernel file and creates a kernel object from the code in that file (in this case the set of instructions that calculate the histogram of an image). Finally, it transfers the arguments to the GPU (input image and empty array in which to return the calculated histogram) and enqueues the execution of the kernel. Once the execution finishes the host program reads the results back into a result buffer.

Each kernel execution processes in parallel every pixel within a subregion of the input image called work-group which dimensions are defined by the user before queuing the execution. The kernel gets the gray level of every pixel within a work-group and increases the value of the corresponding bins of the histogram. Notice that with a parallel approach as the one being described when two pixels of the same work-group have the same gray level, two threads will try to write to the same memory location simultaneously (the same bin of the histogram). To preserve data integrity in such cases it is necessary to use atomic operations, which implies a decrease in performance. See the following code:

```
if(x < image_width && y < image_height){
  color = read_imagef(inputImage, sampler, coordinates);
  atomic_inc(&histogram[color]);
}
```

**Cumulative Histogram:** Computing the cumulative histogram of an image consists in adding the values of all the bins of its histogram, hence it is inherently a sequential operation which isn't likely to be parallelized. For this reason this stage of the algorithm was implemented on the CPU. This routine takes as argument a structure containing the histograms of the N frames of the window being considered for restoration and returns another structure containing the $N$ corresponding cumulative histograms.

**Weighted Average of Cumulative Histograms:** We define a symmetric time window of $N$ frames centered at the frame that is being processed (input frame) and compute the weighted average of the cumulative histograms within this window. The weight assigned to each frame of the window is determined through a normal distribution centered in the input frame. This way the coefficient corresponding to each cumulative histogram considered in the weighted average is calculated in terms of its proximity in time respect to the input frame using weights: $w_i = \exp(-\frac{(i-N/2)^2}{2\sigma^2})$

The standard deviation must be defined in terms of the number of frames that comprise time window. During the tests performed in this study it was found that using $\sigma^2 = 2N$ allows flicker reduction while still respecting the intensity variations that are part of the film's content.

When the first frame of the sequence is being processed the system does not have information from previous frames, therefore the initial weighted average considers only subsequent frames. Once the first frame has been restored previous frames become available and the routine progressively incorporates their cumulative histograms to the computation of the weighted average. The opposite case is presented while reaching the last frame of the sequence.

**Midway Equalization:** Once the average cumulative histogram is computed the algorithm proceeds to equalize the histogram of the input frame. For this we implemented the method proposed in [4]. The routine defines a transformation that assigns to each pixel of the input image the gray level of the element of the average cumulative histogram which has the same rank as the considered pixel.

First the routine determines the rank of each gray level in the input image using its cumulative histogram ($H_{input}$): $H_{input}(\lambda_i) = k_i(rank)$

Then it calculates the gray level that corresponds to that rank through the inverse of the weighted average cumulative histogram: $H_a^{-1}(k_i) = \lambda_i'$

```
for(i=0; i<256; i++){
      //find the rank of each gray level of the input image
      rank = input_cummulative_histogram[i];
      //find the gray level that has the same rank
      r = 0;
      while(input_cummulative_histogram [r] < rank){
           r++;
```

```
        }
        //write the corespondent gray level to the transformation array
        transform[i] = r;
}
```

**Image Transformation:** The routine that applies the transformation to the input image was implemented on an OpenCL kernel. The routine takes as arguments the input image, an array describing the transformation and a blank image to which the result will be written to. Each execution of the kernel reads in parallel the gray level of a section of 16 x 16 pixels (one work-group) and writes the new gray level to the output image (according to the transformation described by the array).

```
if(x < image_width && y < image_height){
  input_color = read_imagef(input_image, sampler, coordinates);
  output_color = transform[input_color];
}
write_imageuf(output_image, coordinates, output_color);
```

## 5 Results and Discussion

In order to quantify the effectiveness of the flicker reduction algorithm implemented in this work several films with different degrees of flicker were processed and their mean gray level was compared before and after processing. The performance of CPU based and GPU based implementations of the algorithm was also measured during these tests.

**Performace Results:** Three films with different frame sizes were processed in order to compare the time it takes to compute the histograms and to apply the transformation using a sequential routine implemented in the CPU and a routine implemented in the GPU.

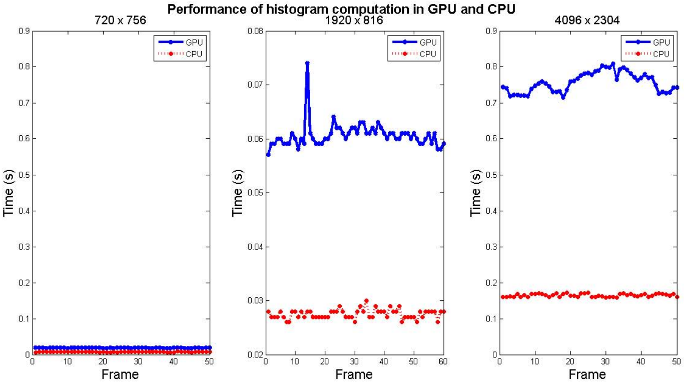


**Fig. 1.** Histogram computation performance

The results presented in Figure 1 shows that in all three cases the computation of histograms runs faster on the CPU than on the GPU. Two possible reasons for the GPUs poor performance on the computation of image histograms are the following:

- In order to compute the histogram of an image on the GPU the input image and a buffer (where to write the histogram to) must be transferred from the CPU to the GPU. After the computation the result must be read from the GPU and this involves transferring the buffer back to the CPU.
- To preserve the integrity of data during simultaneous attempts to writing to the same memory location atomic operations must be used (preventing the execution of parallel instructions).

Figure 2 presents a comparison between the time it takes to apply the transformation to each frame of the sequence using a routine running on the GPU and a routine running on the CPU. These results show that for an image of 720 × 756 pixels both routines achieve similar performances. When processing larger frames (1920 × 816 and 4096 × 2304) the routine runs faster on the GPU than on the CPU.

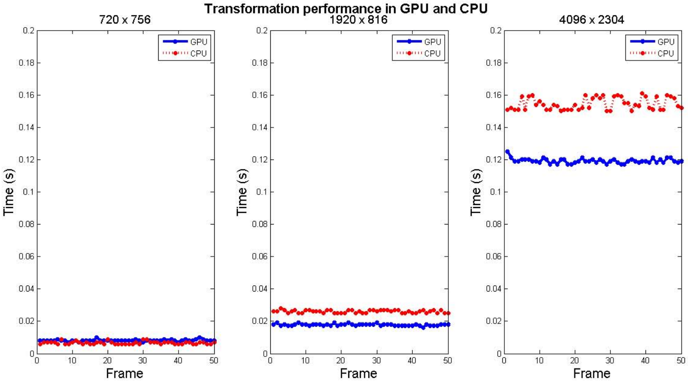


**Fig. 2.** Transformation computation performance

As in the computation of histograms before applying the transformation in the GPU it is necessary to transfer the input data from the CPU to the CPU. In this case the parameters that must be transferred are the input image, an array in which the transformation is defined and a blank image on which the result will be written to. For small frames (less than 1 MP) the time it takes to the CPU and the GPU to apply the transformation is similar, thus the additional delay of the data transfer is not justified. For frames larger than 1 MP pixel the transformation runs faster enough on the GPU over the CPU to justify the computational cost of the data transfer and therefore the total processing time of the GPU is less than the total processing time of the CPU.

**Flicker Reduction:** This section presents the results of three tests aimed at measuring the ability of the algorithm for the reduction of film flicker. The results arise from the comparison of the mean gray level of each frame of a flickering film before and after being processed. Figure 3 shows in blue the mean

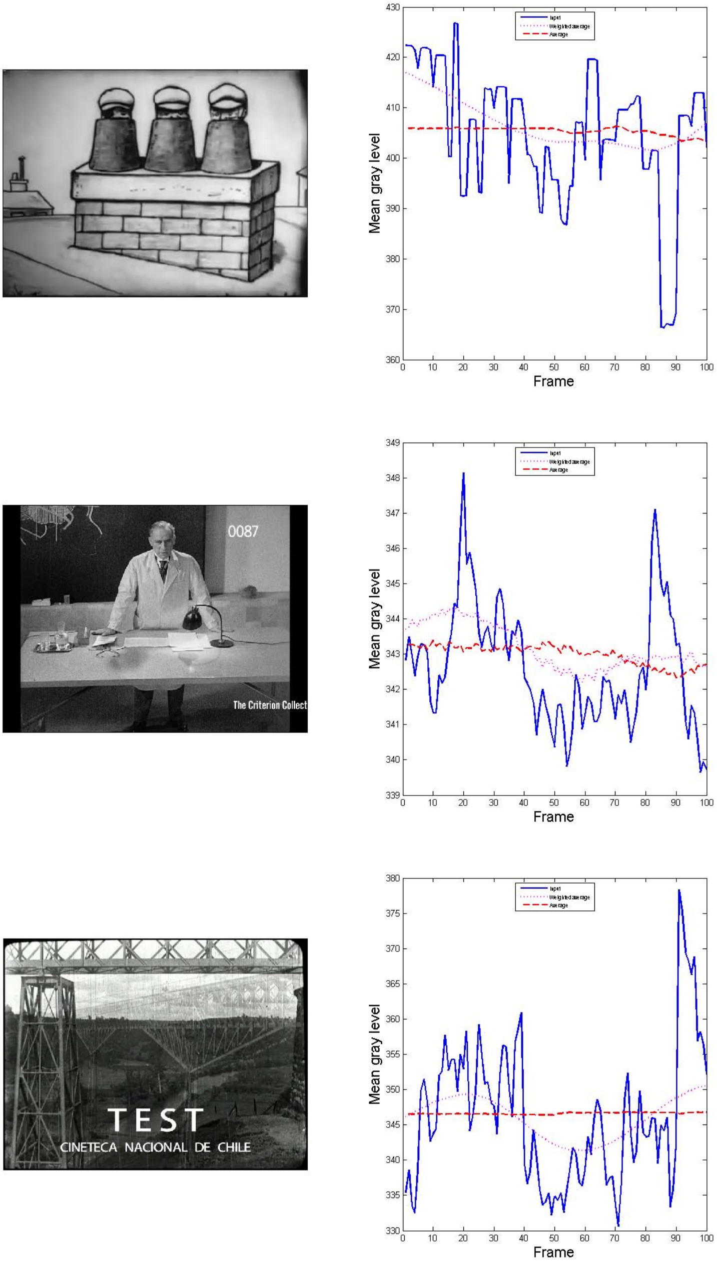


**Fig. 3.** Film flicker reduction. Left: Sample Frame. Right: original mean luminance (blue), restored mean luminance using average (red) and restored mean luminance usong wiegthed average (magenta).

gray level of the first 100 frames of the films. In all three films the mean gray level of the original frames experience significant variations between consecutive frames and several discontinuities along the temporal axis indicating that the films have flicker.

In order to preserve the original content of the scene, it is desirable that the flicker reduction algorithm removes the discontinuities that cause the flicker effect while preserving the intensity variations that are part of the content of the film and not part of any degradation process.

Figure 3 shows that using an average of the cumulative histograms results in an almost constant mean gray level along the stream. This indicates that while this approach will reduce the flicker it will not preserve the original intensity variations of the original content of the film. On the other hand, the mean gray level values that result from using a weighted average of the cumulative histograms within a temporal window show that although rapid transitions and discontinuities were removed, slow transitions remain. This suggests that with this approach the global brightness variations of the original content are preserved.

## 6 Conclusions

This work proposes a way of preserving the original global intensity variations of a film when applying a flicker reduction algorithm based on the computation of the inverse weighted average of the cumulative histograms within a time window. The performance results obtained in this study lead to the conclusion that using OpenCL for computing image histograms on a GPU fails to achieve better performance than the computation of image histograms on a CPU. Regarding the transformation process that reduces the flicker it was observed that the extra time required to transfer data to and from GPU is not justified unless the frames being processed are larger than 1 MP. In such cases running histogram calculation on the CPU and applying the transformation on the GPU can achieve a better performance of the flicker reduction algorithm than if it runs entirely on the CPU.

## References

1. Cox, I., Roy, S., Hingorani, S.: Dynamic histogram warping of image pairs for constant image brightness. In: ICIP, pp. 366–369 (1995)
2. Delon, J.: Midway image equalization. Journal of Mathematical Imaging and Vision 21(2), 119–134 (2004)
3. Delon, J.: Movie and video scale-time equalization; application to flicker reduction. IEEE Trans. Image Proc. 15(1), 241–248 (2006)
4. Delon, J., Desolneux, A.: Stabilization of flicker-like effects in image sequences through local contrast correction. SIAM Journal on Img. Science 3(4), 703–734
5. Naranjo, V., Albiol, A.: Flicker reduction in old films. In: ICIP, pp. 657–659 (2000)

# Motion Silhouette-Based Real Time Action Recognition

Marlon F. de Alcântara, Thierry P. Moreira, and Helio Pedrini

Institute of Computing - University of Campinas
Campinas, SP, Brazil, 13083-852

**Abstract.** Most of the action recognition methods presented in the literature cannot be applied to real life situations. Some of them demand expensive feature extraction or classification processes, some require previous knowledge about starting and ending action times, others are just not scalable. In this paper, we present a real time action recognition method that uses information about the variation of the silhouette shape, which can be extracted and processed with little computational effort, and we apply a fast configuration of lightweight classifiers. The experiments are conducted on the Weizmann dataset and show that our method achieves the state-of-the-art accuracy in real time and can be scaled to work on different conditions and be applied several times simultaneously.

## 1 Introduction

The recent advances in technology have made computers faster, data storage cheaper and video capture more available. This provided extensive usage of applications of automatic human action recognition in video. They can be seen in surveillance systems, cell phones, cars and video games for various purposes. However, researchers face the efficiency-speed trade-off dilemma, seen in many computational problems, which hampers the implementation of real time solutions.

Over the last decade, numerous works have addressed video action recognition, aiming to achieve better classification rates. Eventually, the rates on some datasets have already reached around 100% – some examples are [5, 10, 13]. Hence, more recently, some works have consisted of making the techniques applicable to real life situations, even at the cost of reducing the classification rate. Some researches have addressed real time recognition [2, 8] and studied recognition of multiple actions simultaneously or in sequence [18]. The classification rate reduction in some recent works can be seen in Table 1.

Several methods presented in the literature, such as [10, 14], extract interest points from the video and describe them using solely appearance information. A Bag-of-Word approach is often used to unite all the local features, thus loosing their spatio-temporal distribution. These methods usually have a slow training phase and have limited application. Bregonzio et al. [1] developed a method in which the geometric information is preserved, obtaining better accuracy results,

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 471–478, 2013.

but still not solving these limitations. Another common approach is to use silhouettes; there are simple and fast ways to obtain them. One challenge of these approaches is to find a suitable form of representation; Yi and Krim [17] used an homotopy function to describe a space-time volume formed by silhouettes over time.

The work proposed by Guo et al. [6] have achieved an impressive success rate, however, it uses a dense set of feature vectors and covariance matrices. With some optimization, it can operate in real time, but is not scalable. Other methods, such as [15], work in real time, but have room for improvement in the correct classification rate and in the capability of recognizing multiple actions in space and time. Frequently, the reason why a method cannot be applied in real time is that the used features represent the entire action, therefore, the sequence must be acquired in order to call the classifier. Table 1 summarizes some related works, their accuracy and a short description of their techniques.

This paper proposes a lightweight action recognition method that is capable of identifying actions using only a small number of frames. The method is based on the shape variation of the motion silhouette, thus the features can be extracted on-the-fly and quickly be used to classify the action. For these characteristics, it can be readily applied to work with multiple simultaneous actions and actions in sequence.

**Table 1.** A summary of related works for the Weizmann dataset

| **Work** | **Weizmann rate (%)** | **Techniques** |
|---|---|---|
| Fathi and Mori (2008) [5] | 100 | Combination of low-level features with AdaBoost |
| Niebles, Wang and Fei-fei (2008) [10] | 90 | Bag-of-Words + pLSA |
| Sun, Bhen and Hauptmann (2009) [13] | 97.8 | Zernike moments + Bag-of-Words + SVM |
| Ta et al. (2010) [14] | 94.5 | Bag-of-Words + SVM |
| Hsieh, Huang and Tang (2011) [7] | 98.3 | Silhouette histogram in polar coordinates + Nearest Neighbor |
| Wang, Huang and Tan (2009) [15] | 93.3 | Optical Flow + AdaBoost |
| Bregonzio, Xiang and Gong (2012) [1] | 96.7 | Bag-of-Words + Clouds of Points + Multiple Kernel Learning |
| Junejo and Aghbari (2012) [8] | 88.6 | SAX + Nearest Neighbor |
| Zhang and Tao (2012) [18] | 93.9 | Slow Feature Analysis |
| Chaaroui and Climent-Pérez (2013) [2] | 90.3 | Silhouette clustering into key poses + Nearest Neighbor |
| Guo, Ishwar and Konrad (2013) [6] | 100 | Covariance matrix of spatio-temporal descriptors |

This paper is organized as follows. Section 2 defines the proposed methodology for this work. Section 3 presents and discusses some of the results obtained with the proposed method. Section 4 concludes the paper and includes some future work suggestions for improving the proposed method.

## 2 Methodology

The proposed method for identifying different actions is initialized with a video stream that contains an action, according to Figure 1. The first step is to acquire the motion silhouette by using the difference between consecutive frames; this step is fast to be applied and is responsible for the real time application.

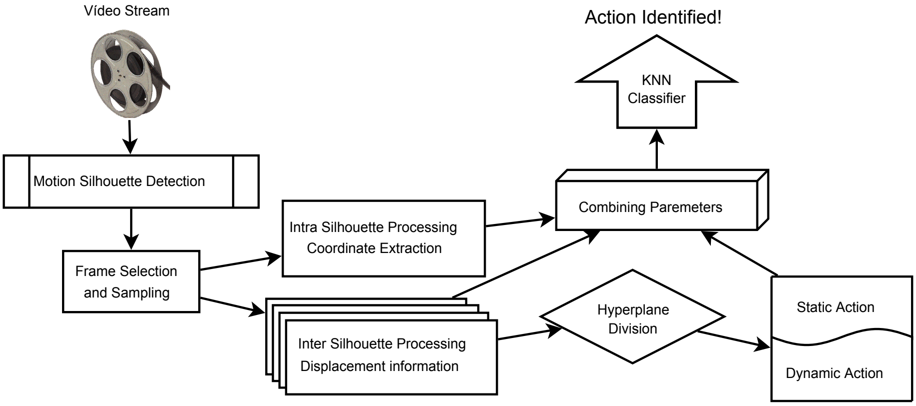

Vídeo Stream
Action Identified!
KNN Classifier
Motion Silhouette Detection
Combining Paremeters
Intra Silhouette Processing Coordinate Extraction
Frame Selection and Sampling
Hyperplane Division
Static Action
Inter Silhouette Processing Displacement information
Dynamic Action

**Fig. 1.** Diagram with the main stages of the proposed method

An action can be represented in the video stream with distinct number of frames. Some of these frames do not represent significant information to classify and can contain some outliers, due to the low robustness and fast speed method for acquiring the points. To overcome this weakness, the algorithm select frames to acquire the points with relevant information. Among the frames selected, a fixed number is sampled and will be used in the subsequent steps.

The extracted silhouettes are used in two distinct processes. The first one is the usage of a bounding box containing the entire silhouette. The bounding box contains some control points; the number of control points is parametrically defined and is equally divided into the four bounding box sides and equally spaced. The control points are used to choose the silhouette interest points; the selected points are those which the distance to each control point is the shortest; Figure 2 illustrates some control points and the selected points in a bounding box.

While the first step is applied to each silhouette separately, the second considers the silhouette point displacement in the same stream (Figure 3). The

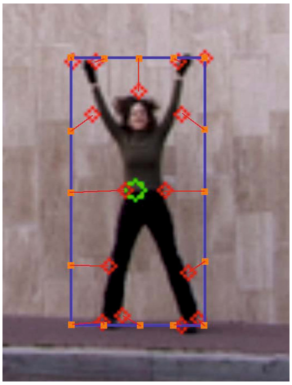


**Fig. 2.** Points of interest are chosen by their distance to the grid points

displacement is measured by using the Euclidean distance on the same points in different frames. Particularly, the displacement in the horizontal axis is used to differentiate static action from dynamic actions. While static actions start and end in the same place, dynamic actions start and end in different places. This information is used to create two hyperplanes, the first for static actions and the second for dynamic actions. Thus, any static action can be identified as a dynamic action and vice versa.

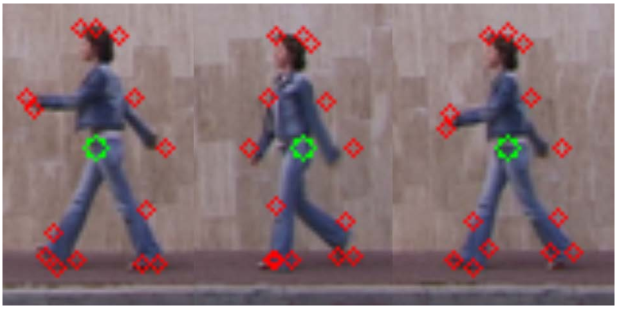


**Fig. 3.** Displacement of interest points in an action sequence

After the intra and inter silhouette process, the resulting parameters are combined into a unique descriptor and submitted to a classifier, which identifies the actions performed on the video stream. There are two viable classifier options: Support Vector Machine (SVM) [3] and $k$-Nearest Neighbors (KNN) [4].

SVM is originally a binary classifier. The training consists of finding a high-dimensional hyperplane that optimally separates the features of two classes. Multiclass classification is usually achieved by the use of several binary SVMs. Two common approaches are the *one-versus-all* and *one-versus-one*. In the first, each class is trained against all others together, and the classifier with the best output function gives the result. In the second, every class is trained against all others, one by one, and the result is given by a voting strategy.

KNN is a multiclass classifier. No training step is required, since the classification step uses the training vectors directly. It works by searching the space for $k$ nearest vectors from the testing instance. If $k$ is 1, it becomes a nearest neighbor classification.

The KNN classifier was chosen since it properly handles multiclasses and works very well in the coordinate system used in the proposed method, once the action classes tend to be organized into clusters. Also, uncorrelated classes usually do not weight in the classification, because only the surroundings of the test vector are considered.

## 3 Experimental Results

In this section, we evaluate our method on public datasets and present the results, as well as details of each method step, such as the tools and the parameters employed.

The experiments presented were conducted on the Weizmann human action dataset [16]. It consists of 10 classes of actions: *walking*, *bending*, *jumping jack*, *jumping*, *jumping in place*, *running*, *side walking*, *skipping*, *waving one hand* and *waving two hands*. Each action class is performed by 9 people, three of those classes have one person with two sequences recorded. Figure 4 shows some examples of actions from the dataset.

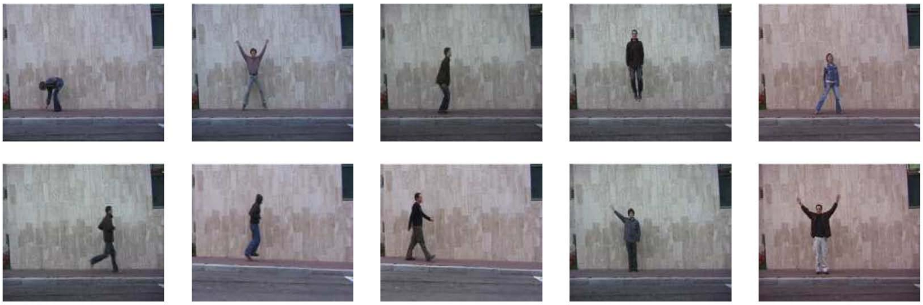


**Fig. 4.** Frames extracted from the Weizmann dataset [16]

To demonstrate the robustness of our method, we performed leave-one-out cross-reference tests. Table 2 shows the results obtained with the Weizmann dataset. It can be seen that 60% of the misclassification happens to the *skipping* class, as also reported by [2, 12], because it has a large intra class variation and is normally confused with *side walking* and *jumping*. The overall correct classification rate is 94.62%. This result shows that it is possible to achieve high rates of classification using simple descriptors, such as the proposed motion silhouette that allowed to perform real time data extraction and classification.

The processing time for a frame sequence is smaller than the video duration. The Weizmann dataset has 93 video sequences with an average duration of 2.45 seconds. The extraction of the features of each video took, in average, 0.135 seconds, and the average time to classify the videos is less than 1 millisecond. The ratio of the *average video duration* to the *average processing time* is *18.14*. The experiments were conducted in a 2.4GHz Intel i7 processor using no parallelism. It shows that our method works in real time with room for inserting improvements, without making it slow.

**Table 2.** Confusion matrix of the results for the Weizmann dataset

| | walk | bend | jack | jump | pjump | run | side | skip | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| walk | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bend | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jack | 0 | 0 | 0.77 | 0 | 0.22 | 0 | 0 | 0 | 0 | 0 |
| jump | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| pjump | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| run | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| side | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| skip | 0 | 0 | 0 | 0.2 | 0 | 0 | 0.1 | 0.7 | 0 | 0 |
| wave1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| wave2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

The classification results also show that not all video frames are necessary to identify the action correctly. After some tests using different number of frames, fifteen frames were used in the final algorithm. For tests using more than fifteen frames, no significant gain was observed.

Sixteen interest points were selected in the silhouettes (the number must be a power of two), resulting in a large number of final descriptor dimensions. The tests using more than sixteen points did not improve the identification process and the tests using less than sixteen showed a weak representation. To reduce the number of dimensions in the final descriptor, the PCA algorithm [11] was applied. Several dimensions were tested and the best value acquired was 30 to classify the video stream.

The KNN classifier allows a parameter to set a number of neighbors to be considered in the classification process. The best parameter observed in this case was one. It is because the used data set contains a short number of videos to train the classifier. In larger databases, this parameter could possibly be increased for a best classification.

Tests were also conducted on the KTH dataset [9], however, the method turned out to be ineffective on it since some videos have constant zooming and camera motion, which causes the detection of untrue displacement. This makes the displacement detection step that separates static from dynamic actions more difficult. The best recognition rate, reached by tunning the parameters, was 58.34%.

## 4 Conclusions

This paper introduced and discussed a new real time motion silhouette-based method for human action recognition. The most onerous part for any action recognition system is usually the descriptor extraction. In this work, we used a simple point selection that considers the relative point position for the control points fixed in a bounding box. This allows a fast silhouette representation and it is possible to recognize an action performed in a video stream correctly through a movement measure. The action sequence is described by the displacements of these points in time.

When using a silhouette-based algorithm, its robustness depends on the sampling adopted. In this step, the amount of points and which of them must be used are critical decisions for achieving high classification rates. To improve the results, our algorithm is capable of sampling a number of points based on the video resolution. For the Weizmann dataset, only sixteen control points were used, corresponding to sixteen interest point coordinates.

Unlike [1, 5, 6], the performance of our algorithm is more than necessary for real time requirements. Nevertheless, our classifier proved to be accurate and even better than other works of literature (Table 1) on the Weizmann dataset with an accuracy of 94.62%.

A motion-based algorithm is not indicated for videos containing camera motion, for instance the KTH dataset, since the method interprets a camera motion as movement, not segmenting correctly the action performed. A possible strategy is to use sophisticated tracking techniques such as [2], which, on the other hand, slow down the method performance.

The solution employed the proposed method is lightweight and easily scalable to work with multiple action instances on a single video sequence, because it is applied on each movement instance separately. Our approach achieves state-of-the-art 94.62% accuracy on the Weizmann dataset.

As future directions, we suggest to the proposed method for extracting and tracking silhouettes in the presence of more complex background, and also apply our method to different types of datasets, such as surveillance videos, videos with camera movements, or videos with other camera angles. It is possible that the descriptor developed in this work is suitable with other classifiers besides KNN.

**Acknowledgements.** The authors are grateful to FAPESP, CNPq and CAPES for the financial support.

## References

1. Bregonzio, M., Xiang, T., Gong, S.: Fusing Appearance and Distribution Information of Interest Points for Action Recognition. Pattern Recognition 45(3), 1220–1234 (2012)

2. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based Human Action Recognition using Sequences of Key Poses. Pattern Recognition Letters (2013)
3. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)
4. Cover, T., Hart, P.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
5. Fathi, A., Mori, G.: Action Recognition by Learning Mid-Level Motion Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
6. Guo, K., Ishwar, P., Konrad, J.: Action Recognition From Video Using Feature Covariance Matrices. IEEE Transactions on Image Processing 22(6), 2479–2494 (2013)
7. Hsieh, C., Huang, P., Tang, M.: The Recognition of Human Action Using Silhouette Histogram. In: Reynolds, M. (ed.) Proceedings of Australasian Computer Science Conference, vol. 113, pp. 11–16. ACS, Perth (2011)
8. Junejo, I.N., Aghbari, Z.A.: Using SAX Representation for Human Action Recognition. Journal of Visual Communication and Image Representation 23(6), 853–861 (2012)
9. KTH Royal Institute of Technology: KTH Action Dataset (2004), http://www.nada.kth.se/cvap/actions/
10. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. International Journal of Computer Vision 79(3), 299–318 (2008)
11. Pearson, K.: Principal Component Analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6(2), 559 (1901)
12. Saghafi, B., Rajan, D.: Human Action Recognition using Pose-based Discriminant Embedding. Image Communications 27(1), 96–111 (2012)
13. Sun, X., Chen, M., Hauptmann, A.: Action Recognition via Local Descriptors and Holistic Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 58–65 (2009)
14. Ta, A.P., Wolf, C., Lavoue, G., Baskurt, A., Jolion, J.M.: Pairwise Features for Human Action Recognition. In: Proceedings of the 20th International Conference on Pattern Recognition, pp. 3224–3227. IEEE Computer Society, Washington, DC (2010)
15. Wang, S., Huang, K., Tan, T.: A Compact Optical Flowbased Motion Representation for Real-Time Action Recognition in Surveillance Scenes. In: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt, pp. 1121–1124 (November 2009)
16. Weizmann Institute of Science: Weizmann Classification Database (2005), http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
17. Yi, S., Krim, H.: Capturing Human Activity by a Curve. In: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt, pp. 3561–3564 (November 2009)
18. Zhang, Z., Tao, D.: Slow feature analysis for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 436–450 (2012)

# A Video Summarization Method Based on Spectral Clustering

Marcos Vinicius Mussel Cirne and Helio Pedrini

Institute of Computing - University of Campinas
Campinas, SP, Brazil, 13083-852

**Abstract.** The constant increase in the availability of digital videos has demanded the development of techniques capable of managing these data in a faster and more efficient way, especially concerning the content analysis. One of the research areas that have recently evolved significantly at this point is video summarization, which consists of generating a short version of a certain video, such that the users can grasp the central message transmitted by the original video. Many of the video summarization approaches make use of clustering algorithms, with the goal of extracting the most important frames of the videos to compose the final summary. However, special clustering algorithms based on a spectral approach have obtained superior results than those obtained with classical clustering algorithms, not only in video summarization techniques but also in other fields, such as machine learning, pattern recognition, and data mining. This work proposes a method for summarization of videos, regardless of their genre, using spectral clustering algorithms. Possibilities of algorithm parallelization for the purpose of optimizing the general performance of the proposed methodology are also discussed.

## 1 Introduction

Due to the great increase in the generation of digital videos in the last years, there is an increasingly need to develop techniques that are capable of manipulating these data in an automatic, efficient and accurate way, concerning the issues of searching, browsing, retrieval and content analysis. Among these techniques is the video summarization, which consists of deriving a short version from a given video, preserving as much relevant information as possible, such that the users can grasp the message transmitted by the original video. The generated summaries can then be integrated into many applications, such as interactive searching and browsing systems, making both management and access to video content more accurate [14].

Nevertheless, defining what is important or not in video summarization is an open problem, especially because there is a variety of video genres, such as sports, movies, news programs, documentaries, and home movies in general. Even to humans, it is hard to reach a consensus to know how good a summary is, since what is relevant to ones may not be to others. Thus, the main challenge of the video summarization field is in how to make a system to take the best decisions to choose the most important parts of a video. This is usually done by analyzing

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 479–486, 2013.

high level features (e.g. semantic content, time, space) or low level features (e.g. color histograms, texture, subtitles, audio, shape and motion descriptors).

Among the various approaches to the video summarization problem are those which make use of clustering algorithms, that are also objects of study in fields such as data mining, machine learning, and statistics. The idea beyond these approaches is to split the frames of a given video into different groups such that frames that belong to the same group are more similar among themselves. Then, a set of keyframes is extracted from these groups, i.e., the frames that best represent both the belonging groups and the essence of the original video content. Later, the final summary is generated from these keyframes.

A clustering technique that has been increasingly growing recently is the spectral clustering [13], due to the fact that it can generate more satisfactory results than those obtained by classic clustering algorithms. In the case of video summarization, even though there are many approaches that use clustering algorithms, little has been produced with spectral clustering algorithms so far.

The objective of this work is to propose and analyze a new method for video summarization of any genre using spectral clustering algorithms. A qualitative analysis of the generated summaries with different feature descriptors is conducted, comparing the results with a specific database, which includes summaries from other approaches.

The main contributions of this work include the creation and implementation of a method that can be integrated into many video processing environments and a performance and accuracy analysis of the proposed method, considering the variety of existing video genres.

This paper is organized as follows: Section 2 describes the main concepts about video summarization and spectral clustering, as well as works related to both topics; Section 3 defines the proposed methodology for this work; Section 4 presents and discusses some of the obtained results with the proposed method; Section 5 includes the general conclusions about the discussed topic and some future work suggestions in order to improve the proposed method.

## 2 Concepts and Related Work

This section describes general concepts about video summarization, together with related works and spectral clustering algorithms, and how they are applied to the video summarization context.

### 2.1 Video Summarization

A digital video can be defined as a collection of images that have the same dimensions, grouped according to a temporal sequence. Each of these images is known as *frame*, which corresponds to the smallest structural unit of a video, representing a picture captured by a camera in a given time instant of the video. The frames can be grouped into *shots*, which are sequences of frames, captured in a contiguous way, and that represent a continuous action in time or space. Finally, a group of shots that are semantically correlated constitutes a *scene*.

Video summarization techniques can be divided into *static* and *dynamic*. In the first category, the summary is generated as a collection of still images denominated *keyframes* [16], that represent the content of a video in the form of a storyboard. The advantage of this approach is in its simplicity and efficiency, usually being free of redundancies, but it may not preserve the temporal order of the selected keyframes. In the second category, many segments of the video are chosen, which are then organized such that the temporal order of the video is preserved [21]. Dynamic summarization has the main advantage of generating summaries which a higher richness of details, but it is more expensive than static summarization approaches, besides the possible generation of redundancies.

Another challenge in the video summarization field is the definition of standard metrics to evaluate the quality of the results. At the moment, there is no consistent platform to evaluate summaries. Thus, each work has its own evaluation method and, in most cases, it does not compare the results with other existing methods [20].

### 2.2 Spectral Clustering

Spectral clustering [13] has become one of the most popular clustering techniques lately, being an important research object in fields such as pattern recognition, machine learning, and signal processing. It provides better results than those from classic clustering algorithms (such as $K$-means) and it can be easily implemented by means of numeric computation platforms. In the video summarization context, spectral clustering can be used in tasks such as keyframe extraction [7] and shot boundary detection [8].

Given a set of $n$ points, located at an $l$-dimensional space, to be divided into $k$ distinct subsets, where $n$, $l$ and $k$ are positive integers, an affinity matrix $A_{n \times n}$ is constructed such that each element $A(i,j)$ corresponds to a similarity measure $s_{ij} \geq 0$ that represents the likelihood degree between a pair of points $i$ and $j$ of the set, with $A(i,i) = 0$. Thus, the bigger the value of $A(i,j)$, the higher is the similarity between the points $i$ and $j$ and vice-versa.

Later, the diagonal matrix $D_{n \times n}$ is defined, where $D(i,i) = \sum_{j=1}^{n} A(i,j)$. From $A$ and $D$, the Laplacian matrix $L = I - (D^{-1/2}AD^{-1/2})$ is constructed, where $I_{n \times n}$ is the identity matrix. In the next step, the $k$ largest eigenvectors of $L$ are calculated, forming the matrix $X = [x_1 x_2 ... x_k]$ by stacking these eigenvectors in $k$ columns. After that, the matrix $Y$ is created from $X$ by normalizing the rows of $X$ such that each one has unitary length. Finally, the rows of $Y$ are separated into $k$ groups by the $K$-means algorithm (or any other clustering algorithm, such as the ones described in [9]), assigning the point $i$ of the initial set to group $j$ if, and only if, the row $i$ of matrix $Y$ is assigned to cluster $j$.

The choices of the similarity measure to be used and the number of clusters in which the dataset is split are not trivial tasks, once that they are subject to the application domain of this set. First of all, it must be assured that the data considered as "very similar" by the chosen similarity measure have a very close relationship in the application domain as well [13]. Furthermore, in most of the cases, there is not a "correct" number of groups. In this situation, it is common to use strategies that find this number in an automatic way [18].

Usually, the matrices computed by the spectral clustering algorithms are very large, demanding a large storage space, especially when working with digital videos, composed of a considerable number of frames. In order to guarantee the efficiency on the implementation of these algorithms, it is necessary that the Laplacian matrix related to the similarity graph be sparse, simplifying the task of calculating the $k$ largest eigenvectors. To do this, graphs such as $\epsilon$-neighborhood and $k$-nearest neighbors are used, eliminating the computation of the similarity measures between every single pair of points.

## 3 Methodology

The methodology of this work will be focused on a new method for digital video summarization of any genre using spectral clustering to obtain summaries with a better quality than those found in the state-of-the-art. A comparative analysis of the generated summaries of some methods of the literature with the ones generated by the proposed method is conducted. A general flowchart of the methodology stages is shown in Figure 1.

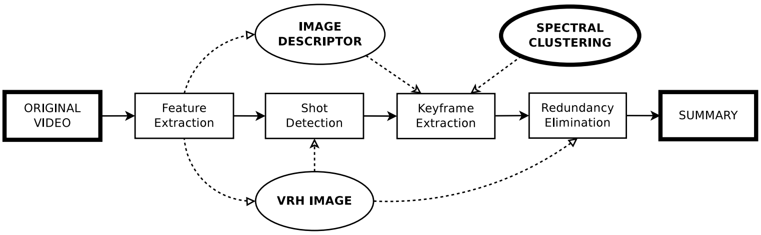


**Fig. 1.** Flowchart of the main stages of the proposed method for video summarization

From a given digital video, the feature extraction stage will primarily make a sampling of this video in frames. To optimize the performance of the application, only 5 frames per second are used in this stage. From these frames, both the visual rhythm by histogram (VRH) [11] and image descriptors for each frame are calculated. In this process, many image descriptors that encompass spatial and temporal features are evaluated, such as SIFT (Scale-Invariant Feature Transform) [12], SURF (Speeded Up Robust Features) [5] and ORB (Oriented FAST and Rotated BRIEF[1]) [17].

In the shot detection stage, the estimation of the number of video shots is started, which will be the number $k$ of clusters used in the next stage. From the VRH image, the shot boundaries are detected by using the local adaptive threshold technique described in [19], which produces more accurate results rather than

[1] Acronym that stands for Binary Robust Independent Elementary Features [6].

using a fixed threshold to detect the boundaries. Starting with $k = 1$, every time a shot boundary is detected, $k$ is incremented by 1.

After estimating $k$, a spectral clustering algorithm is executed for the keyframe extraction stage. Using descriptor feature vectors, an affinity matrix $A$ is constructed, as defined in Section 2.2, where the element $A(i, j)$ corresponds to the distance between the feature vectors of frame $i$ and the one of frame $j$. After the calculation of the normalized eigenvectors, the $K$-means algorithm is run to cluster the frames according to the shots to which they are associated, where the number of clusters corresponds to $k$.

Finally, the keyframes of each cluster are extracted based on the centroids calculated by $K$-means (one keyframe per cluster) and preserving their temporal order. The selected keyframes correspond to the ones that are closest to their respective cluster centroids. Before the summary generation process, a post-processing is performed to eliminate redundant frames. This is done by computing the sums of pairwise pixel distances between the columns of the VRH image (generated in the feature extraction stage) related to two consecutive keyframes. After that, these values are compared to a distance threshold $T_d$. If the distance between keyframe $i$ and keyframe $i + 1$, where $1 \leq i \leq k - 1$, is less than $T_d$, the keyframe $i$ will be considered as redundant and, therefore, will not be included in the final summary. The threshold value was empirically defined as $T_d = (\mu_d + \sigma_d)/4$, where $\mu_d$ and $\sigma_d$ are the mean and the standard deviation of all distances, respectively. This approach performs well with most of the generated redundant frames from the videos used in the tests, but it may fail at detecting redundant frames with high luminosity differences (brightness and contrast), since their columns in the VRH image are very distant from each other.

From the remaining keyframes, the final summary is then created, which can be done in a static way, generating a storyboard, or in a dynamic way, taking a certain amount of frames around each keyframe in the original video, such that the total number of selected keyframes correspond to a percentage of the total number of frames of the original video.

The advantage of this method is that every stage is executed in an unsupervised way, such that the number of shots does not need to be known *a priori*. However, the whole summarization process is still expensive, because of the spectral clustering, even though it leads to more accurate results than standard clustering approaches.

## 4 Experimental Results

The tests were done using an AMD Phenom II X6 3.2 GHz processor and 4 GB of memory. The methodology described in Section 3 was implemented with OpenCV platform [1]. A collection of 50 videos of several genres from Open Video Project (OVP) [2], available at the VSUMM database [3] (provided by the authors of the approach described in [4]), were used in the tests, together with the respective summaries produced by different video summarization methods, which include Delaunay Triangulation (DT) [15], STIMO (STIll and MOving Video Storyboards) [10], as well as the OVP summaries and the ones provided by VSUMM. All of the videos have, together, approximately a total duration of 75 minutes and 150,000 frames (352 × 240 pixels). After the execution of the implementation of the proposed method for each descriptor and using all videos, it was observed that SIFT provided the fastest execution time, with a total execution time of 1.10 hours, followed by ORB (4.04 hours) and SURF (7.59 hours).

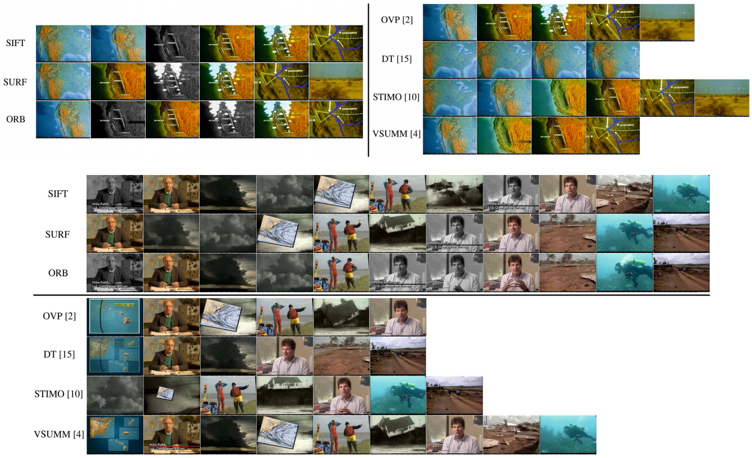


**Fig. 2.** Summarization results for *The Great Web of Water, Segment 02*) video (upper image) and *Hurricane Force - A Coastal Perspective, Segment 03* (lower image). For each descriptor, redundant frames are represented as greyscale images.

To evaluate the quality of the summaries, only two videos are analyzed due to space limitation in the paper: *The Great Web of Water, Segment 02*, which has 5 shots, and *Hurricane Force - A Coastal Perspective, Segment 03*, with 12 shots. Figure 2 shows the respective results, together with the summaries generated by different approaches, as well as the one provided by the OVP database. For the first video, it can be seen that the proposed method generated summaries with 6 keyframes, one more than the number of shots, which means that the shot boundary detection process performed very well for this video. Also, the redundant frames (represented as greyscale images) were properly detected and eliminated for the final summary, once that the respective contents of the detected redundant frames are similar to their consecutive frames, leaving only the colored ones. With respect to the quality of the summaries, the SURF summary was the only descriptor that included the contents of all shots, being the closest to the OVP summary. Furthermore, the SIFT summary included two keyframes of a same shot (1st and 2nd frames), and the ORB summary was the one that generated more redundant frames (2nd and 4th frames) than the other descriptors. Comparing to other approaches, SURF performed slightly better than both STIMO and VSUMM, which produced the best summaries among the other approaches. This happens because STIMO included more than one frame of a shot, even though it included at least one frame of every shot, and VSUMM missed the last shot.

For the second video, all of the summaries of each descriptor contain 11 keyframes, one less than the number of video shots. In the redundancy elimination process, it can be noticed that three frames were discarded in the ORB summary (1st, 6th and 7th frames), whereas SIFT summary had two discarded frames (1st and 8th frames) and only one for the SURF summary (7th frame). However, all of the eliminated frames (except for the 7th one of the ORB summary) have a little more information than the remaining consecutive frames of the respective final summaries. Concerning the summary content, SURF selected most of the different shots not only among the descriptors but also the other approaches as well. On the other hand, comparing the summaries of the proposed method to the OVP summary, none of them was able to select a frame from the first shot, as occurred both in DT and VSUMM summaries.

Despite this analysis, it is hard to evaluate how the misdetection of a shot (i.e., when a frame of a shot is not included in the final summary) affects the comprehension of the central message transmitted by a video. For that, a more subjective evaluation must be made, once it requires a deeper content analysis and a general consensus about the degree of relevance of each shot. In other words, even though the summaries produced by each descriptor have more different shots than the ones of other approaches (including the OVP), all of them may have the same relevance in particular situations.

## 5 Conclusions

This paper described a method for video summarization from any genre using a spectral clustering algorithm. Different image descriptors were used to extract features from the video frames, as well as the normalized eigenvectors of the respective affinity matrices. The $K$-means algorithm was used to cluster video frames according to the number of shots detected by a previous procedure that uses a visual rhythm by histogram image to identify shot boundaries. Redundant frames are then discarded to produce the final summaries, which were compared against summaries produced by different video summarization approaches (DT, STIMO, VSUMM and the ground-truth provided by the OVP database).

Despite the slowest processing time, the summaries produced by SURF were the best among the tested descriptors, once they detected most of the different shots and generated less redundant frames than SIFT and ORB. Comparing SURF to other approaches, the results were very close in most cases, although SURF produced more complete summaries. Furthermore, both the shot boundary detection and the redundancy elimination procedures performed well in the analyzed videos, yet they still need some adjustments to improve their accuracy.

**Acknowledgements.** The authors are grateful to FAPESP and CNPq for the financial support.

## References

1. OpenCV: Open Source Computer Vision (2013), `http://opencv.org`

2. The Open Video Project (2013), http://www.open-video.org
3. VSUMM (Video SUMMarization) (2013), https://sites.google.com/site/vsummsite
4. de Avila, S.E.F., Lopes, A.P.B., da Luz Jr., A., de Albuquerque Araújo, A.: VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. Pattern Recognition Letters 32(1), 56–68 (2011)
5. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
6. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
7. Chasanis, V., Likas, A., Galatsanos, N.: Video Rushes Summarization Using Spectral Clustering and Sequence Alignment. In: 2nd ACM TRECVid Video Summarization Workshop, Vancouver, BC, Canada, pp. 75–79 (2008)
8. Damnjanovic, U., Izquierdo, E., Grzegorzek, M.: Shot Boundary Detection Using Spectral Clustering. In: 15th European Signal Processing Conference, Poznan, Poland, pp. 1779–1783 (September 2007)
9. Elhamifar, E., Sapiro, G., Vidal, R.: See All by Looking at a Few: Sparse Modeling for Finding Representative Objects. In: IEEE Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, pp. 1600–1607 (2012)
10. Furini, M., Geraci, F., Montangero, M., Pellegrini, M.: STIMO: STIll and MOving Video Storyboard For The Web Scenario. In: Multimedia Tools and Applications, vol. 46, pp. 47–69. Kluwer Academic Publishers, Hingham (2010)
11. Guimarães, S.J.F., Couprie, M., Araújo, A.D.A., Leite, N.J.: Video Segmentation Based on 2D Image Analysis. Pattern Recognition Letters 24(7), 947–957 (2003)
12. Lowe, D.: Object Recognition from Local Scale-Invariant Features. In: Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
13. Luxburg, U.: A Tutorial on Spectral Clustering. Statistics and Computing 17(4), 395–416 (2007)
14. Money, A.G., Agius, H.: Video Summarisation: A Conceptual Framework and Survey of the State of the Art. Journal of Visual Communication and Image Representation 19(2), 121–143 (2008)
15. Mundur, P., Rao, Y., Yesha, Y.: Keyframe-Based Video Summarization Using Delaunay Clustering. International Journal on Digital Libraries 6, 219–232 (2006)
16. Peng, J., Xiaolin, Q.: Keyframe-Based Video Summary Using Visual Attention Clues. IEEE MultiMedia 17(2), 64–73 (2010)
17. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An Efficient Alternative to SIFT or SURF. In: IEEE International Conference on Computer Vision, Barcelona, Spain (2011)
18. Sanguinetti, G., Laidler, J., Lawrence, N.D.: Automatic Determination of the Number of Clusters Using Spectral Algorithms. In: IEEE Machine Learning for Signal Processing, pp. 28–30 (2005)
19. Shekar, B., Raghurama Holla, K., Sharmila Kumari, M.: Video Shot Detection Using Cumulative Colour Histogram. In: Mohan, S., Kumar, S.S. (eds.) 4th International Conference on Signal and Image Processing. LNEE, vol. 222, pp. 353–363. Springer, Heidelberg (2012)
20. Truong, B.T., Venkatesh, S.: Video Abstraction: A Systematic Review and Classification. ACM Transactions on Multimedia Computing, Communications and Applications 3(1) (February 2007)
21. Zhou, H., Sadka, A.H., Swash, M.R., Azizi, J., Sadiq, U.A.: Feature Extraction and Clustering for Dynamic Video Summarisation. Neurocomputing 73, 1718–1729 (2010)

# Motion Estimation from RGB-D Images Using Graph Homomorphism

David da Silva Pires[1], Roberto M. Cesar-Jr[1], and Luiz Velho[2]

[1] University of São Paulo, São Paulo, Brazil
{davidsp,cesar}@vision.ime.usp.br
http://www.vision.ime.usp.br/creativision

[2] National Institute for Pure and Applied Mathematics, Rio de Janeiro, Brazil
http://www.visgraf.impa.br

**Abstract.** We present an approach for motion estimation from videos captured by depth-sensing cameras. Our method uses the technique of graph matching to find groups of pixels that move to the same direction in subsequent frames. In order to choose the best matching for each patch, we minimize a cost function that accounts for distances on RGB and XYZ spaces. Our application runs at real-time rates for low resolution images and has shown to be a convenient framework to deal with input data generated by the new depth-sensing devices. The results show clearly the advantage obtained in the use of RGB-D images over RGB images.

**Keywords:** motion estimation, graph matching, RGB-D images.

## 1 Introduction

With the advent of devices like Kinect™ (from Microsoft®) and Xtion™ (from ASUS®) that capture texture and depth images from a scene, there are many new challenges and problems to be faced. One of the main applications for data captured by such equipments is generally concerned with natural interaction. These applications typically use anthropometric algorithms to estimate pose, skeleton and the number of users in front of the device. Some systems have specialized algorithms to recognize its users, even if there is identical twins among them [1]. Gesture recognition using Kinect™ has been used as a control to other devices, aiming an easier or more natural interaction and allowing the use of computers with great accessibility [2].

### 1.1 Objective and Motivation

This work aims to show the benefits of using the additional information given by the depth image registered with a texture image, presenting an algorithm, based on graph matching, that detects the direction of movement at real-time rates. The developed procedure shows, with labels identified by colors or, optionally, with the use of arrows, to which direction each group of pixels (rectangular areas, called patches, arranged in a regular grid on the image) is moving.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 487–494, 2013.

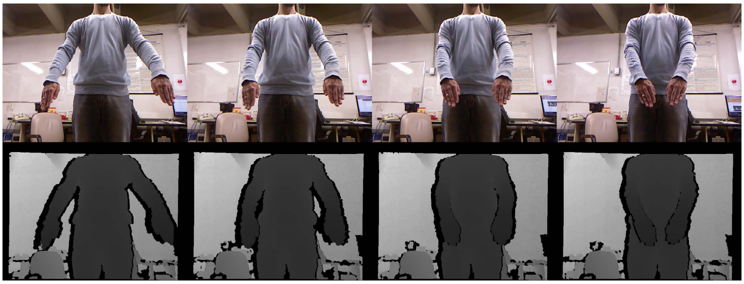


**Fig. 1.** Input data is a video sequence of texture and depth map images per frame, providing the RGB and $(x, y, z)$ values of each pixel

Depth data, when added to the traditional RGB values and texture coordinates, help the delimitation of objects of interest and makes results substantially better when considered as a descriptive feature for each pixel. This characteristic is an advantage with respect to methods that depend on the presence of a well defined pattern on texture, like checkerboard sequences [3].

The technique developed on this paper may be useful on general applications, such as 3D scene reconstruction and augmented reality. Our approach is also an intermediary step to identify the rigid components of an articulated object [4].

Given a video sequence, such as the example shown at Figure 1, the application builds a graph for each frame and compare them subsequently, finding a matching based on distances at RGB and XYZ spaces.

## 2 Related Work

The evolving technology regarding depth-sensing devices was initially created to provide natural interaction to video-games. However, recent Computer Vision and Graphics research has shown a lot of other interesting uses.

The Kinect Identity technology [1] explores a set of three independent identification techniques: face recognition, clothing color tracking and height estimation. These techniques were selected from a major set and demonstrated to be the best ones that, at the same time, were robust, non-CPU and memory intensive and as independent as possible from each other. Such choice indicates the importance of the development of tracking algorithms that uses both kinds of data: texture and depth map.

The lack of a better treatment of the depth data in conjunction with the RGB data and, consequently, the use of them on motion estimation algorithms is felt even by developers of software specialized in gesture recognition, like FAAST, the Flexible Action and Articulated Skeleton Toolkit [5]. They have demonstrated interest in developing real-time head tracking and estimation of the twist of the user's arm. None of these are provided by the middle-ware OpenNI™ and the solutions to these problems certainly involves Computer Vision techniques to be applied to both input data.

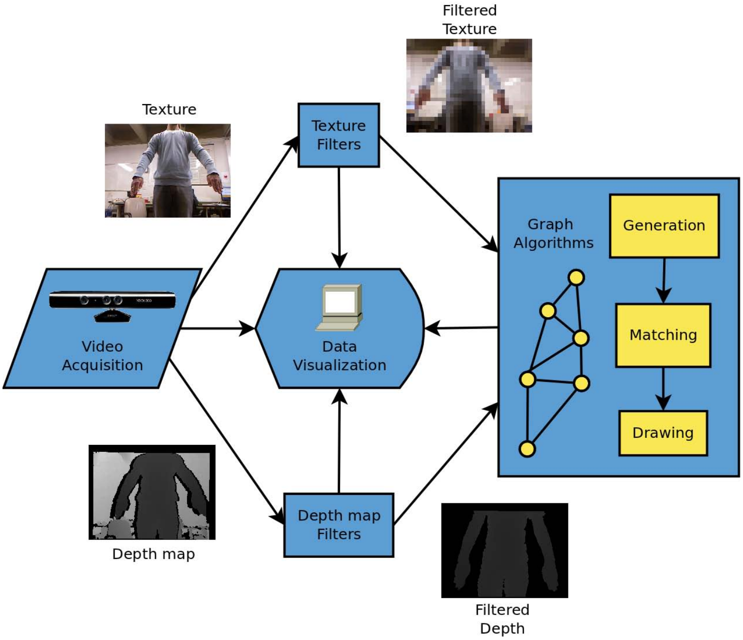


**Fig. 2.** Data flow representing the implemented method

In the present work, we use graph matching to find a correspondence between two point sets. This approach has been used to solve many Computer Vision problems such as interactive natural image segmentation [6], computer-assisted colorization [7] and point matching for non-rigid registration [8], among others [9].

## 3 Methodology

Our method is a kind of discrete optimization for determining optical flow. The data processing occurs according to a specific pipeline that is composed of the following steps: data acquisition; texture filters; depth map filters; graph algorithms; data visualization. The data flow is schematized in Figure 2.

### 3.1 Graph Based Approach

In order to find a matching, we have to consider relevant features that describe the points and to have a way to compare such features. Thus, each frame generates a graph whose vertices are derived from patches properties. We used six values for each pixel on the input data: RGB data, extracted from color channels, and $(x, y, z)$ data, with $x$ and $y$ being texture coordinates and $z$ being the distance to the capturing device. The frame representation is given by an attributed relational graph (ARG), allowing storage and comparison of structural, temporal and quantitative information. The recognition of the direction of the movement is done through an inexact graph matching. This approach allows differences between model and input graph [10]. In the present paper, each pair of subsequent frames generates the model and input graphs.

An ARG is a graph whose vertices represent objects while edges denote relations among them. Objects can be characterized by a finite number of attributes (numerical or symbolic), such as area, perimeter, color and shape. The relations often correspond to distances and relative orientation between objects, although more rich spatial relations may be adopted. With the contents of each frame being represented by an ARG, motion estimation resumes to a graph matching, consisting of a determination of a mapping of the vertex set of an ARG to the vertex set of another one.

Each graph is treated as a complete graph, in the sense that every vertex is connected to all the other vertices. We compute a cost function (see Section 3.3) involving the distance between the RGB and depth values of each pair of vertices: $(v_m, v_i)$, where $v_m$ and $v_i$ are vertices from the model and the input graphs, respectively. The pair that minimizes this cost function is added to the matching set.

While graph vertices store point sets, including position information about these points, structural relations are stored at graph edges.

### 3.2 Graph Generation

The model and the input graph are built from consecutive pairs of texture and depth map input frames. Thus, at the beginning of the acquisition procedure, we can build just one graph. As the subsequent frames are captured, the input graph relative to the immediate past frame is assigned to the model graph and a new input graph is built from the new data acquired.

In order to build the graph, we consider the representation of the input images (texture and depth map) composed by patches. Given an image and the patches' parameters, we can compute how much patches compose the new representation, being sufficient to make a division between the number of rows and columns of each one. Thus, patches' dimensions are directly related with the size of the graphs that are created, highly influencing the performance of the application.

Each patch is a candidate to have a vertex representing it on the graph, being elected based on its $z$ (depth) value. A new vertex is created and inserted on the graph only if its $z$ value do not belong to shadow areas on depth map or if it is not too close nor too far of the capturing device. This perspective treats the depth map as a valid mask to texture pixels and allows easy background elimination based on a predefined threshold.

### 3.3 Graph Matching

The matching is done between two graphs: model and input. The model graph represents the last pair of frames (texture and depth images), captured before the current one, which is represented by the input graph.

A matching is an ordered pair of vertex descriptors, the first one referring to a model graph vertex and the second one relative to an input graph vertex.

For each vertex belonging to the model graph, we find the vertex on the input graph that minimizes the cost function. Eventually, the matched vertices have the same $(x, y)$ texture coordinates, indicating that no movement has occurred at that location between the two pairs of frames.

The cost function $c$ is given by a convex combination of two distances, $d_{RGB}$ and $d_{XYZ}$:

$$c = \alpha \cdot d_{RGB} + (1 - \alpha) \cdot d_{XYZ}. \tag{1}$$

The $d_{RGB}$ value measures the distance of the color of the patches being compared on RGB space, while the $d_{XYZ}$ value measures the distance between the $(x, y)$ texture coordinates and between the $z$ depth values. As it can be seen, the parameter $\alpha$ controls how much each distance is considered at the final value of the cost function.

When calculating the cost function, we need to decide about which distance function to use. Two different distances have been implemented, the city block (Manhattan):

$$d_{RGB} = \left|v_R^M - v_R^I\right| + \left|v_G^M - v_G^I\right| + \left|v_B^M - v_B^I\right|, \tag{2}$$

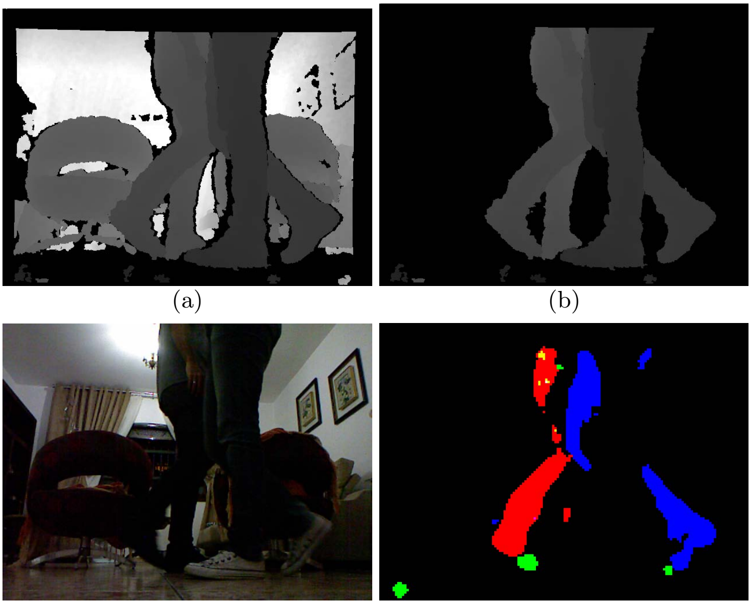
and the Euclidean distance:

$$\begin{aligned} d_{RGB} &= \left\|v_{RGB}^M - v_{RGB}^I\right\|_2 \\ &= \left\|\left(v_R^M - v_R^I, v_G^M - v_G^I, v_B^M - v_B^I\right)\right\|_2, \end{aligned} \tag{3}$$

where the raised indexes $M$ and $I$ indicate if the vertex belongs to the model or input graph, respectively, while the sub-indexes $R$, $G$ and $B$ indicate the channel color being considered. The same formulas are applied to the $(x, y, z)$ depth map values. Best results were achieved with the use of Euclidean distance.

## 4 Results and Discussion

The present section illustrates the results produced by the system with some real cases. All the examples were captured at 30 fps, with VGA resolution $(640 \times 480)$ and run in an Intel Core 2 duo computer. Figure 3 shows the captured depth of two subjects walking at opposite directions, with occlusion occurring between them and also between their respective legs. The image at Figure 3(b) shows the same depth at Figure 3(a) after background elimination filtering. Figure 3(c) shows the texture for the same scene. Note that the depth is already registered with the texture. Finally, Figure 3(d) shows the detected motion represented as color labels, with the colors red, green and blue representing movement to right, up and left directions, respectively. As we can see, the method accounts for the effects of occlusion. This experiment shows the identification of motion present on the scene. The leftmost subject is more distant to the capturing device, as indicated by the gray levels in the depth map; it is walking from left to right.



(a) (b)

(c) (d)

**Fig. 3.** Result obtained at a scene where two subjects are walking at opposite directions. This also exemplifies a case of occlusion between moving objects. **(a)** Depth image. **(b)** The same depth at image (a) after filtering for background elimination. **(c)** Texture image for the same scene shown at (a). **(d)** Detected motion represented as color labels.

The other subject, closer to the capturing device, executes a movement from right to left. Note the correct classification of both movements, even on the region where they intercept each other. The green pixels that arise on Figure 3(d) were identified as moving up, a reasonable result, except for the green blob that appears at left-bottom corner: it appears due to error on depth capturing. This same experiment is also an example of how our method gets successful results even in the presence of occlusion of moving objects. Note how a leg is partially occluded by another and still has its motion correctly identified.

Figure 4 shows two RGB-D frames on which a couple of dancers executes a movement to right with a subtle motion of the arms to up. The images at the bottom of the figure show the output obtained for patches of dimension 4 × 6 when the parameter $\alpha$ takes the values 0.00, 0.25, 0.50, 0.75 and 1.00. The best result was obtained for $\alpha = 0.50$. Black pixels indicate absence of motion.

For the input shown at left of Figure 5, we obtained the output shown at right of the same figure. This experiment shows the nice visual appealing of the matchings provided by the arrows. Instead of color labels, arrows with circled tips are used to indicate the motion.

Various values of the parameter $\alpha$ together with different patches sizes have been evaluated. As expected, our motion estimation achieved better results with

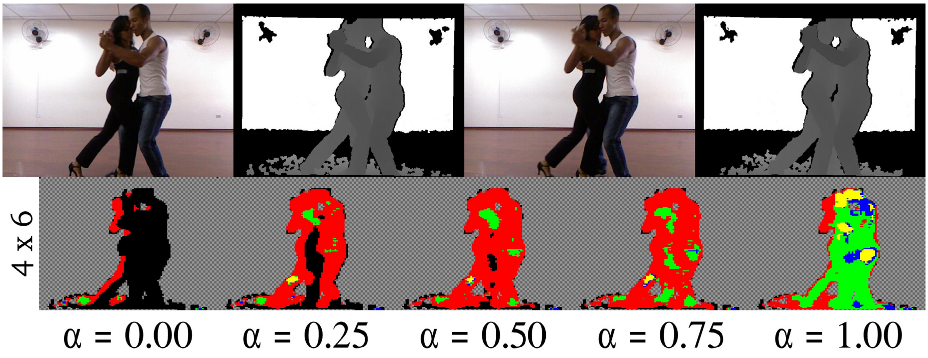


**Fig. 4.** Varying the values of parameter $\alpha$

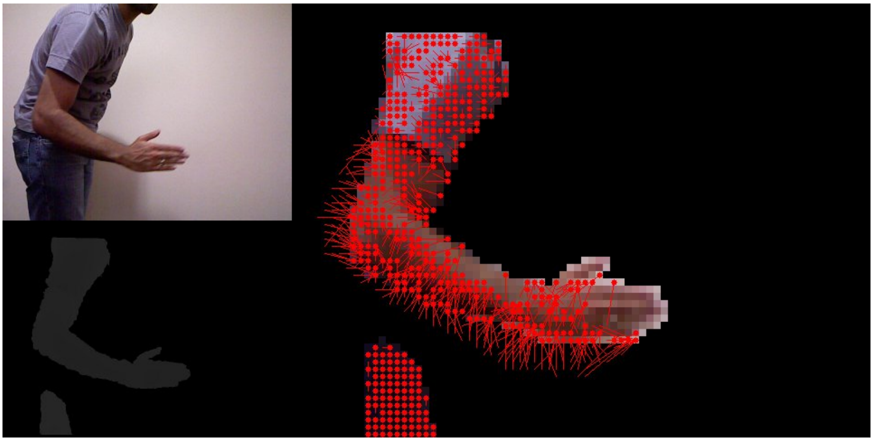


**Fig. 5.** Video sequence with a subject moving his arm. Matchings are shown as arrows, indicating that the arm is moving up and to the right direction.

intermediate values of $\alpha$. For values close to the extremes of the valid range (0 and 1), the results were poor, indicating matchings that were not consistent with the observed movement.

### 4.1 Conclusions and Future Works

Our algorithm takes as input a sequence of pairs of registered RGB texture and depth map. Since the acquired input depth map is already registered to the texture image, there is no need for using knowledge about intrinsic or extrinsic calibration parameters between the infra-red light receptor and the camera.

The developed system has shown to be a convenient framework to deal with input data generated by the new depth-sensing devices. The application is capable of doing image processing and execute computer vision algorithms, thus allowing easy evaluation of results at real-time rates.

There are many features that may be considered to improve the results and the performance:

– Enforce coherence in the motion of nearby patches, which often present similar motions, turning the technique less sensitive to image noise and ambiguous results.
– The direction of the movement of the patches is calculated on 3D space but the visualization of the results is done only in the plane that is determined by texture coordinates. The development of a 3D visualization have already been started.
– It is possible to apply this method to detect the rigid parts of an articulated object.
– Compare our results with other 3D motion segmentation algorithms [3].

Our ongoing work include all these possibilities and new advances will be reported in due time.

## References

1. Leyvand, T., Meekhof, C., Wei, Y.C., Sun, J., Guo, B.: Kinect Identity: Technology and experience. Computer 44, 94–96 (2011)
2. Gallo, L., Placitelli, A., Ciampi, M.: Controller-free exploration of medical image data: Experiencing the Kinect. In: CBMS, pp. 1–6 (June 2011)
3. Tron, R., Vidal, R.: A benchmark for the comparison of 3-D motion segmentation algorithms. In: IEEE CVPR, pp. 1–8 (2007)
4. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Learning layered motion segmentations of video. In: ICCV (2005)
5. Suma, E.A., Lange, B., Rizzo, A.S., Krum, D.M., Bolas, M.: FAAST: The Flexible Action and Articulated Skeleton Toolkit. IEEE Virtual Reality, 245–246 (March 2011)
6. Noma, A., Graciano, A.B.V., Cesar Jr., R.M., Consularo, L.A., Bloch, I.: Interactive image segmentation by matching attributed relational graphs. Pattern Recognition 45(3), 1159–1179 (2012)
7. Noma, A., Velho, L., Cesar Jr, R.M.: A computer-assisted colorization approach based on efficient belief propagation and graph matching. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 345–352. Springer, Heidelberg (2009)
8. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Comput. Vis. Image Underst. 89(2-3), 114–141 (2003)
9. Conte, D., Foggia, P., Sansone, C., Vento, M.: Thirty years of graph matching in pattern recognition. Intl. Journal of Pattern Recognition and Artificial Intelligence 18(3), 265–298 (2004)
10. Cesar Jr, R.M., Bengoetxea, E., Bloch, I., Larranaga, P.: Inexact graph matching for model-based recognition: Evaluation and comparison of optimization algorithms. Pattern Recognition 38(11), 2099–2113 (2005)

# MoCap Data Segmentation and Classification Using Kernel Based Multi-channel Analysis

Sergio García-Vega, Andrés Marino Álvarez-Meza,
and César Germán Castellanos-Domínguez

Universidad Nacional de Colombia, Sede Manizales,
Signal Processing and Recognition Group
km 7 vía al Magdalena, Colombia
{segarciave,amalvarezme,cgcastellanosd}@unal.edu.co
http://portal.manizales.unal.edu.co/gta/signal/

**Abstract.** A methodology for automatic segmentation and classification of multi-channel data related to motion capture (MoCap) videos of cyclic activities are presented. Regarding this, a kernel approach is employed to obtain a time representation, which captures the cyclic behavior of a given multi-channel data. Moreover, we calculate a mapping based on kernel principal component analysis, in order to obtain a low-dimensional space that encodes the main cyclic behaviors. From such, low-dimensional space the main segments of the studied activity are inferred. Then, a distance based classifier is used to classified each MoCap video segment. A well-known MoCap database is tested which contains different activities performed by humans. Attained results shows how our approach is a simple alternative to obtain a suitable classification performance in comparison to complex methods for MoCap analysis.

**Keywords:** Multi-channel data, kernel methods, MoCap, human activity recognition.

## 1 Introduction

Human action recognition from video data are a growing area of study in the computer vision field. For a correct recognizing, it is necessary to develop a system that allows to identify and classify characteristic patterns from the input data [1] [2]. In real life, there are some human activities involving a cyclic behavior along the time, such as: walking, running, swimming, among others. Commonly, it is important to identify the main cyclic behavior that describes each action to find relevant information about the process [3]. For such purpose, it is necessary to develop three main stages: preprocessing, segmentation, and classification. However, the segmentation stage is not always developed in an automatic way, which can lead to unstable results and low classification performances. Moreover, when the data segmentation stage is fixed manually, it could lead in a time demanding process for the user. Then, it is necessary to develop an automatic segmentation stage that allows to obtain a suitable data analysis.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 495–502, 2013.

There are some works in the state-of-the-art related to the analysis of Motion Capture - MoCap data for human activity recognition. In [4], it is used a well-known MoCap database and the dynamics of each action class is modeled by a Bayesian based approach using Hidden Markov Models - HMM. The achieved accuracy classification results are over the 90%, however, the system requires that the multi-channel data is previously segmented, such that each segment contains a whole course of one action. Moreover, a complex classifier is employed to train the data, which requires a high computational load. Other approaches that require a manual MoCap data segmentation can be found in [5].

Here, a methodology for automatic segmentation and classification of multi-channel data is proposed. In this sense, a kernel function is employed to discover the time relationships among multi-channel data. Our aim is to highlight the cyclic behavior of the studied process, which is assumed to be hidden into the input samples. Indeed, an eigen-based decomposition is used to find a low-dimensional space that allows to segment the cyclic segments of the input data. Thus, proposed methodology is able to capture cyclic behaviors hidden into multi-channel data, avoiding the need of a manual segmentation that could lead in biased and unstable results. A well-known MoCap database is tested, which contains different activities executed by humans. Furthermore, two classification alternatives are studied: by considering each MoCap frame as an unique sample, and by considering a set of frames.

The remainder of this work is organized as follow. Section 2 introduces the proposed methodology for automatic segmentation and recognition of multi-channel data using kernel based methods. In Sections 3 and 4, the experimental results are described and discussed, respectively. Finally, in Section 5, the work conclusions are presented.

## 2 Kernel Based Multi-channel Data Representation

Let $\boldsymbol{X} \in \Re^{N \times P}$ be a multi-channel input matrix, with $P$ channels and $N$ samples, where $\boldsymbol{x}_i \in \Re^{1 \times P}$ is a row vector containing the information of all the provided channels at different time instants, with $i \in \{1, \dots, N\}$. Our aim is to identify the main relationships that the channels share along the time to highlight hidden cyclic patterns into the studied process. For such purpose, a kernel function is employed to discover such relationships taking into account a non-linear mapping $\varphi : \Re^{N \times p} \to \mathcal{H}$, where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space - RKHS [6]. Thus, the kernel based representation allows to deal with nonlinear structures that can not be directly estimated by traditional operators, such as, the linear correlation function. Regarding this, the inner product between two samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is computed in RKHS as $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$, being $\kappa(\cdot, \cdot)$ a Mercer's kernel [6]. Taking advantage of the so-called kernel trick, the kernel function can be computed directly from $\boldsymbol{X}$. Here, the well-known Gaussian kernel is considered, which can be defined as

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), \tag{1}$$

being $\sigma \in \Re^+$ the kernel band-width. Then, from equation (1) the similarity matrix $\boldsymbol{S} \in \Re^{N \times N}$ can be estimated as $S_{i,j} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It is important to note that other kind of kernels could be used, e.g. linear, polynomial, Laplacian, tangential, among others. However, due to the smooth nature of the input data and considering the universal approximating capability, the Gaussian kernel is used [7]. Each application task could be adapted or not to each kernel function according to the prior knowledge about the input data (see [6,8]). In this sense, $\boldsymbol{S}$ encodes the temporal dynamics of the multi-channel input data. Analyzing the pair similarities information into $\boldsymbol{S}$, it is possible to cluster (segment) samples that are related to a cyclic behavior of the studied process. Note that, the above mentioned kernel representation assumes that the multi-channel data shares an unique cyclic behavior. In case that the input data is composed by different processes, or when the multi-channel data is non-stationary, an unique kernel function could be not enough to deal with such changes along the time, not mentioning the need to consider the time structure of such kind of processes.

### 2.1 Automatic Multi-channel Data Segmentation

From the above mentioned kernel based multi-channel representation, and in order to find out the cyclic behavior into $\boldsymbol{X}$, we propose to use an eigen-based decomposition of $\boldsymbol{S}$ to calculate a low-dimensional space $\boldsymbol{Y} \in \Re^{N \times m}$, with $m < P$, which reveals the main components of $\boldsymbol{X}$. Therefore, the well-known Kernel Principal Components Analysis - KPCA algorithm is performed over $\boldsymbol{S}$. KPCA is a nonlinear generalization of PCA in the sense that it performs PCA in $\mathcal{H}$, which can be viewed as a feature space of arbitrarily large dimensionality [6]. Before applying KPCA, a Laplacian based normalization is employed to avoid the effect of outliers, thus, the matrix $\boldsymbol{L_M} \in \Re^{N \times N}$ is computed as $\boldsymbol{L_M} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$, where $\boldsymbol{D} \in \Re^{N \times N}$ is a diagonal matrix with elements $d_{ii} = \sum_{i=1}^{n} S_{ij}$. Afterwards, the low-dimension KPCA mapping is obtained as $\boldsymbol{Y} = \ \boldsymbol{L_M}\boldsymbol{V}$, where $\boldsymbol{V} \in \Re^{P \times m}$ is a matrix containing the first $m$ eigenvectors of $\boldsymbol{L_M}$, after discarding the first one as trivial solution.

As a result, we obtain a low-dimensional representation that contains the main cyclic components of $\boldsymbol{X}$. Hence, we find the local maxima or *peaks* vector $\boldsymbol{\rho} \in \Re^B$, where $B$ indicates the number of found *peaks* into the first coordinate (column vector) $\boldsymbol{y}$ of $\boldsymbol{Y}$. Note that, each column vector of $\boldsymbol{Y}$ could be related to a different cyclic component of $\boldsymbol{X}$. However, for complex dynamics and/or non-stationary environments, such components can mix more than one cyclic behavior. As a first approach, here, we assume that the given multi-channel data encodes an unique cyclic dynamic. Besides, the signal to noise ratio is high enough to ensure stable performances. Then, each element of $\boldsymbol{\rho}$ is estimated as follows. We compare each element $y_i$ against its two nearest neighbors $y_{i-1}$ and $y_{i+1}$. If $y_i$ value is higher than the value of its neighbors, so, $y_i$ is labeled as a local *peak* and $\rho_b = i$, with $b \in \{1, \ldots, B\}$. After that, we compute the differences between adjacent elements of $\boldsymbol{\rho}$ and finally, take into account the amount of *peaks* found $B$, we obtain $B - 1$ segments of $\boldsymbol{X}$. Fig. 1 illustrates

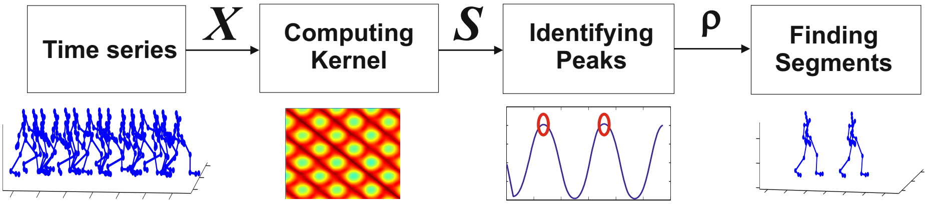


**Fig. 1.** Proposed methodology for multi-channel data segmentation

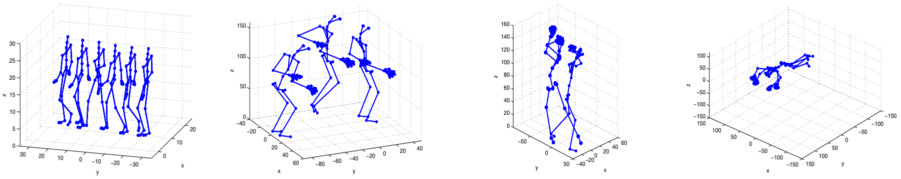
the proposed approach for automatic multi-channel segmentation using a kernel based representation (a motion capture video analysis example is described).

## 3 Experimental Set-up and Results

We test our automatic segmentation and classification methodology for multi-channel data analysis, using a well-known Motion Capture Database - MoCap database, with the purpose to find the main cyclic patterns of human motion activities. In this sense, the CMU MoCap is used[1]. Such data were recorded in a MoCap lab at Carnegie Mellon University, which contains 12 Vicon infrared MX-40 cameras, each of which is capable of recording 120 $Hz$ with images of 4 mega pixel resolution. The cameras are placed around a rectangular area, of approximately $3m \times 8m$, in the center of the room. Subjects wear a black jump suit and have 57 markers taped on, and the Vicon cameras see the markers in infra-red. The images that the various cameras pick up are triangulated to get 3D data representation. The subjects are asked to perform several human motions activities, which are captured by the MoCap system. Then, a video in BVH format for each motion activity by a given subject is recorded. Thus, 146 videos of 31 different subjects are considered for 11 different activities: *jump*, *walk*, *run*, *marching*, *salsa dance*, *golf*, *boxing*, *swimming*, *yoga*, *monkey (human subject)* and *chicken (human subject)*. For each video, an input multi-channel matrix $\boldsymbol{X} \in \Re^{N \times P}$ is obtained, where $P = 57 \times 3 = 171$ corresponds to the 57 joints in 3D coordinates, and $N$ represents the number of frames in the BVH file. As seen from Fig. 2, it is possible to notice some examples from the database used on this work.

In order to avoid the bias effect due to the subject translation along the 3-D space when performing a human activity, e.g. walking and running, a preprocessing stage is carried out, where each input frame is normalized with respect to the Hips joint 3-D coordinates. Thus is, this joint will be always centered at the $(0, 0, 0)$ position for every time instant. After that, we compute the kernel matrix as shown in equation 1, where $\sigma$ is computed according to the empirical estimation of the Gaussian kernel band-width by the *Sylverman's rule* [9]. From such kernel based representation, we estimate the different segments for

[1] http://mocap.cs.cmu.edu/



(a) walk (b) jump (c) golf (d) swimming

**Fig. 2.** Some MoCap human activities representative frames

each video as described in subsection 2.1, fixing $m = 1$. Table 1 describes the amount of segments found automatically for each activity with our approach. Such segments contain the main cycles of the dynamics for the different considered activities. As result we found 910 different segments that represent 112045 frames. The main stages for the proposed automatic segmentation approach are presented in Fig. 3 for two MoCap videos examples.

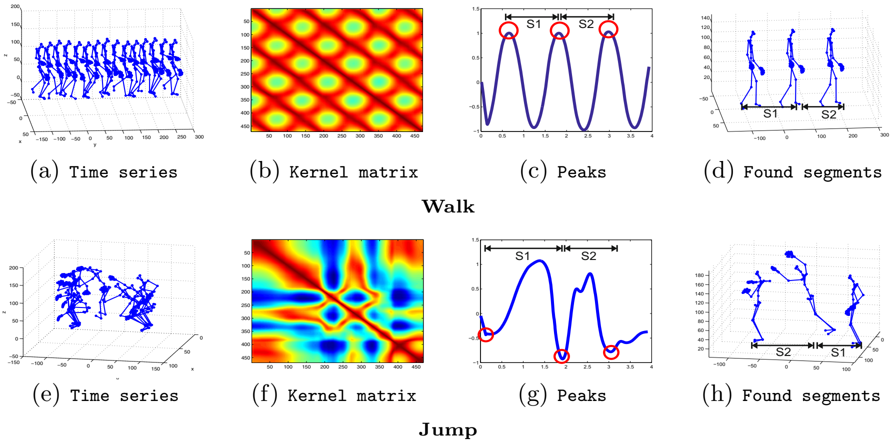
Furthermore, given the computed segments, the generalization abilities for the provided experimental conditions are tested by using a 10-fold cross-validation scheme. Regarding this, a $k$-nearest neighbors (KNN) classifier is used to recognize automatically different activities. The number of neighbors for this classifier is optimized with respect to the leave-one-out error of the training set. In this case, two kind of experiments are provided. First, each frame is employed as an unique sample. Second, for a given video segment, its class membership is estimated as the mode of the labels of the frames within the segment. In Tables 2 and 3 the mean confusion matrices for the above mentioned classification conditions are presented. Finally, at the bottom of the Table 3, the performance of the proposed methodology is compared against the results obtained in [4].

## 4 Discussion

At the top of the Fig. 3, it is possible to see the main segmentation results by using the proposed approach to analyze a walk MoCap video. Particularly, Fig. 3 *(b)* shows the computed kernel matrix, which properly identifies the cyclic similarities (green circles) into the video. Now, Fig. 3 *(c)* describes how our method

**Table 1.** Number of identified segments per each human activity

| Activity | Found Segments | Activity | Found Segments |
|---|---|---|---|
| jump | 68 | golf | 23 |
| walk | 127 | boxing | 36 |
| run | 78 | swimming | 41 |
| marching | 81 | yoga | 86 |
| salsa dance | 114 | monkey(HS) | 135 |
| chicken(HS) | 121 | | |
| **Total: 910 Segments** | | | |



(a) Time series (b) Kernel matrix (c) Peaks (d) Found segments

**Walk**

(e) Time series (f) Kernel matrix (g) Peaks (h) Found segments

**Jump**

**Fig. 3.** Some automatic segmentation results - Main stages

**Table 2.** Mean confusion matrix - frames classification results

| | Jump | Walk | Run | Marching | Salsa dance | Chicken (HS) | Golf | Boxing | Swimming | Yoga | Monkey (HS) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Jump | 92,94 | 0,16 | 1,75 | 0,52 | 0,96 | 0,00 | 0,60 | 0,00 | 0,33 | 1,97 | 0,46 |
| Walk | 1,06 | 98,22 | 6,00 | 0,84 | 0,52 | 0,00 | 0,70 | 0,00 | 0,08 | 0,10 | 0,00 |
| Run | 0,70 | 1,06 | 79,84 | 3,25 | 1,24 | 0,00 | 0,58 | 0,00 | 1,93 | 0,63 | 0,00 |
| Marching | 0,37 | 0,42 | 5,16 | 94,34 | 0,95 | 0,00 | 0,21 | 0,00 | 1,47 | 0,84 | 0,00 |
| Salsa dance | 0,69 | 0,05 | 3,38 | 0,89 | 93,13 | 0,00 | 1,50 | 0,00 | 0,39 | 2,86 | 0,00 |
| Chicken (HS) | 0,00 | 0,00 | 0,00 | 0,00 | 0,08 | 100,00 | 0,00 | 0,00 | 0,00 | 0,14 | 0,00 |
| Golf | 0,09 | 0,00 | 0,62 | 0,00 | 0,97 | 0,00 | 92,28 | 1,47 | 0,00 | 0,80 | 0,00 |
| Boxing | 0,22 | 0,00 | 0,00 | 0,00 | 0,21 | 0,00 | 2,75 | 97,73 | 0,00 | 0,98 | 0,00 |
| Swimming | 0,00 | 0,09 | 2,40 | 0,00 | 0,12 | 0,00 | 0,39 | 0,10 | 95,80 | 0,00 | 0,00 |
| Yoga | 3,83 | 0,00 | 0,34 | 0,15 | 1,38 | 0,00 | 0,77 | 0,44 | 0,00 | 90,02 | 0,62 |
| Monkey (HS) | 0,10 | 0,00 | 0,52 | 0,00 | 0,45 | 0,00 | 0,23 | 0,27 | 0,00 | 1,64 | 98,92 |

**Table 3.** Mean confusion matrix - video segments classification results

| | Jump | Walk | Run | Marching | Salsa dance | Chicken (HS) | Golf | Boxing | Swimming | Yoga | Monkey (HS) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Jump | 98,57 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,83 |
| Walk | 0,00 | 100,00 | 2,50 | 0,00 | 0,91 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Run | 0,00 | 0,00 | 93,75 | 1,25 | 0,00 | 0,00 | 0,00 | 0,00 | 2,36 | 0,00 | 0,00 |
| Marching | 0,00 | 0,00 | 0,00 | 98,75 | 0,00 | 0,00 | 0,00 | 0,00 | 2,50 | 0,71 | 0,00 |
| Salsa dance | 0,00 | 0,00 | 2,50 | 0,00 | 99,09 | 0,00 | 0,00 | 0,00 | 0,00 | 1,48 | 0,00 |
| Chicken (HS) | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 100,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Golf | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 97,50 | 2,00 | 0,00 | 0,00 | 0,00 |
| Boxing | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 2,50 | 98,00 | 0,00 | 0,00 | 0,00 |
| Swimming | 0,00 | 0,00 | 1,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 95,14 | 0,00 | 0,00 |
| Yoga | 1,43 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 96,32 | 0,00 |
| Monkey (HS) | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,48 | 99,17 |

| | Simple Activities | | Complex Activities | |
|---|---|---|---|---|
| | Benchmark [4] | Kernel Multi-Channel | Benchmark [4] | Kernel Multi-Channel |
| Mean accuracy | 91.96% | 97.77% | 92.14% | 97.88% |

models, in one-dimensional coordinate, the main relationships among frames. In this case, due to walking is a slow motion with smooth changes between adjacent frames, the KPCA mapping can be related to a *sin* function. Now, from Fig. 3

*(d)* 3 peaks are calculated (red circles), which properly identify the 2 gait cycles performed by the subject. Analogously, at the bottom of the Fig. 3, we have the main segmentation results for a jump MoCap video. Note that, even when jumping is a cyclic activity with a stronger dynamic of change than walking, our approach is able to infer such behaviors. As seen in Fig. 3 *(f)* the computed kernel matrix highlights 2 set of frames that share a strong similarity into them. Such segments can be identified in the first KPCA coordinate as presented in Fig. 3 *(g)*. Again, note that how our approach is able to track the activity cyclic behavior, even when *S1* is smoother and longer than *S2*.

Regarding to the classification results, as can be seen in Table 2, the mean confusion matrix for the frame based classification scheme demonstrates how our approach obtains a suitable recognition accuracy. Overall, performances over the 90% are attained for all the provided classes. The worst result is obtained for *run*, where the system is confused with *walk* and *march* classes. Above drawback is expected considering that *run* is the class with lowest number of segmented sequences (see Table 1). Moreover, a frame based classification could not be the best alternative to differentiate between activities that share many MoCap poses, e.g., run and walk. Thus is, such video segments are conformed by some frames where the spatial position of the human body joints are similar for both activities. It is important to note that our method, in most of the cases, obtains a better frame based classification performance in comparison to a closed work presented in [4]. Moreover, our approach is a simple solution that includes both, data segmentation and classification.

Now, taking into account the segment based classification scheme results presented in Table 3, it is possible to see how such alternative is more stable than the frame based classification. Attained results describe an average accuracy over the 95%. Particularly, the worst frame based classification performance (*run*) is improved from 79, 84% to 93, 75%. Above system behavior can be explained by the fact that a segment classification decision considers the mode of the frame labels as the segment membership. So, the mode function can be viewed as a filter that is robust against wrong decisions due to pose mistakes (human body joint similarities). Finally, at the bottom of the Table 3, the performance of the proposed methodology is compared against the results obtained [4]. The classification success of our method lies in the automatic segmentation approach, which suitable identifies the main dynamic cycles of the process.

## 5 Conclusions

A methodology for automatic segmentation and classification of multi-channel data was presented. In this sense, a kernel based representation is employed to find out the time relationships among channels. Then, a KPCA mapping is calculated to highlight the main dynamics of the studied process in a low-dimensional space. From such low-dimensional space, a local minimum based method is used to cluster different time segments that share a common behavior. Therefore, our approach is able to capture cyclic behaviors hidden into multi-channel data.

A well-known MoCap database was tested, which contains different activities executed by humans. For concrete testing, proposed approach is used to segment automatically the video data. Such segments are employed to train a *k-nearest* neighbors (KNN) classifier for recognizing automatically different activities. Besides, two kind of classify experiments are carried out: by considering each frame as an unique sample, and by considering a set of frames (video segment). The attained results showed that our approach is a simple but efficient alternative to obtain a suitable classification performance in comparison to other complex state of the art methods related to MoCap data classification. Besides, state art methods employs, in most of the cases, a manually video segmentation, which can lead to subjectively results and inefficient real-world implementations. As future work, we are interested in test our methodology in other kind of human activities that involve different cyclic patterns and non-stationary environments by coupling the proposed method with an online based adaptive filter scheme.

**Acknowledgements**. Research carried out under grants provided by Jóvenes Investigadores e Innovadores - 2012, and a Ph.D. scholarship funded by Colciencias.

## References

1. Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., Shotton, J.: Real-time human pose recognition in parts from single depth images. In: CVPR, pp. 1297–1304 (2011)
2. Gan, Y.W.C., Wang, X.: Human motion segmentation by rpca with augmented lagrange multiplier. In: ICALIP, pp. 379–383 (2012)
3. Murugappan, M., Basah, S.N.B., Yaacob, S.B., Ismail, K.N.S.B.K.: Human postures modeling using motion analysis: A review. In: International Conference on Biomedical Engineering (ICoBE), pp. 280–285 (2012)
4. Lv, F., Nevatia, R.: Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 359–372. Springer, Heidelberg (2006)
5. Lan, Z.-Z., De la Torre, F., Hoai, M.: Joint segmentation and classification of human actions in video. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3265–3272 (2011)
6. Scholkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge (2002)
7. Liu, W., Príncipe, J.C., Haykin, S.: Kernel Adaptive Filtering: A Comprehensive Introduction. John Wiley & Sons, Inc. (2010)
8. Genton, M.G., Cristianini, N., Shawe-taylor, J., Williamson, R.: Classes of kernels for machine learning: a statistics perspective. Journal of Machine Learning Research 2, 299–312 (2001)
9. Sheather, S.J.: Density estimation. Statistical Sci. 19, 588–597 (2004)

# Structural Cues in 2D Tracking: Edge Lengths vs. Barycentric Coordinates

Nicole M. Artner and Walter G. Kropatsch

Vienna University of Technology
Pattern Recognition and Image Processing Group, Vienna, Austria
{artner,krw}@prip.tuwien.ac.at

**Abstract.** Graph models offer high representational power and useful structural cues. Unfortunately, tracking objects by matching graphs over time is in general NP-hard. Simple appearance-based trackers are able to find temporal correspondences fast and efficient, but often fail to overcome challenging situations like occlusions, distractors and noise. This paper proposes an approach, where an attributed graph is used to represent the structure of the target object and multiple, simple trackers in combination with structural cues replace the costly graph matching. Thus, the strengths of both methodologies are combined to overcome their weaknesses. Experiments based on synthetic videos are used to evaluate two possible structural cues. Results show the superiority of the cue based on barycentric coordinates and the potential of the proposed tracking approach in challenging situations.

## 1 Introduction

Even though there exists a vast amount of approaches for video tracking [1,2], this field of research still has some open problems and challenges. The aim of this paper is to show, which challenges can be overcome by choosing a graph-based representation for the target object and by employing structural cues deduced from this representation in tracking. We will study the following challenges from [1]: (1) Distractors: neighboring objects with similar appearance as the target object; (2) Occlusions: varying degrees of partial occlusions; (3) Varying object pose: translation and rotation in 2D and global scaling; (4) Noise: Gaussian white noise and Salt & Pepper.

The concept of the proposed approach is to represent the target object by a graph, where its vertices represent salient features describing the target object and its edges encode their spatial relationships. Instead of graph matching, appearance-based trackers are employed to find the temporal correspondences of the vertices with the help of structural cues deduced from the graph representation. Hence, the proposed approach is able to benefit from the strengths of graph-based representations to overcome challenges during tracking (see Tab. 1). In this paper, two structural cues are compared: *edge cue* and *triangle cue*. The **contributions** of this paper are:

1. A novel structural cue based on barycentric coordinates (triangles);

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 503–511, 2013.

**Table 1.** Strength ⊕ and weaknesses ⊖ of simple trackers and graph-based trackers

| **Simple tracker** | **Graph-based tracker** |
|---|---|
| ⊕ fast correspondence finding | ⊖ costly graph matching |
| ⊖ sensitive against partial occlusions | ⊕ robust against partial occlusions |
| ⊖ sensitive against noise | ⊕ robust against noise |
| ⊖ sensitive against distractors | ⊕ robust against distractors |

2. Comparison of performance of two structural cues (edges and triangles);
3. Evaluation of structural cues under challenges;
4. Analysis of the influence of different parameters on the proposed method.

The edge cue is related to pictorial structures introduced in 1973 by Fischler et al., where the target object is described by a set of parts in a deformable configuration. Felzenszwalb et al. [3] continued and improved the ideas of Fischler et al. to do part-based object recognition for faces and articulated objects. Ramanan et al. apply in [4] the ideas from [3] in tracking people. In comparison to the related work, the edge cue in this paper can be calculated from arbitrary graphs and instead of using structure to verify statistical hypothesis, the proposed structural cues emerge from the underlying structure.

The triangle cue in this paper is determined from barycentric coordinates, which were introduced by August Ferdinand Möbius in 1827. Barycentric coordinates are particularly important in computer graphics, but are also used in computer vision. Salzmann et al. [5] represent surfaces as triangulated meshes and try to recover their 3D shape from 2D correspondences. Barycentric coordinates are used to describe the surface coordinates of each pixel through the triangle inside which they lie. In [6], Dornaika et al. track faces in a particle filter based framework using a statistical facial appearance model. After a general 3D face model is adapted to the face in the input video, barycentric coordinates are used to describe the position of each pixel within its associated triangle. In comparison to [5] and [6], we calculate the barycentric coordinates of vertices outside of triangles (see Fig. 1).

**Overview of Paper.** Sec. 2 describes the proposed graph model. Sec. 3 shortly presents the appearance-based tracker. Sec. 4 introduces the edge cue and Sec. 5 the novel triangle cue. In Sec. 6 the combined iterative tracking is described. Sec. 7 covers the evaluation of the proposed structural cues. Conclusions are given in Sec. 8.

## 2 Structural Model: Attributed Graph

An attributed graph **G** consists of a set of vertices **V**, which are connected via a set of edges **E**. The edges **E** are inserted following the rules of the*Delaunay triangulation*. Hence, there is also a set of triangles **F**, where $c : \mathbf{F} \mapsto V^3; c(f) = \{v_1, v_2, v_3\}$ and $\{e_1(v_1, v_2), e_2(v_2, v_3), e_3(v_3, v_1)\} \in \mathbf{E}$ are the corresponding edges. The model stores attributes with vertices, edges and triangles.

**Attributes of Vertices.** Each vertex $v \in \mathbf{V}$ stores a set of attributes $\{\mathbf{p}, \mathbf{B}, \mathbf{a}\}$.

- $\mathbf{p} : \mathbf{V} \times \mathcal{T} \mapsto \mathbf{R}^2; \mathbf{p}(v,t) = (x,y)^T$ is the 2D position of vertex $v$ at time $t \in \mathcal{T}$. These coordinates are updated in every iteration of the tracking algorithm.
- $\mathbf{B} : \mathbf{V} \times \mathbf{F}' \mapsto \mathbf{R}^3; \mathbf{B}(v, \mathbf{F}')$ is a set of barycentric coordinates of vertex $v$ for each triangle $f \in \mathbf{F}'$, where $\mathbf{F}' = \{f \in \mathbf{F} | v \notin c(f)\}$. The barycentric coordinates are determined during initialization and are constant over time (see Sec. 5).
- $\mathbf{a} : \mathbf{V} \mapsto \mathbf{R}^n; \mathbf{a}(v)$ delivers features for vertex $v$ from an image window $I_{n \times n}$ centered at position of $\mathbf{p}(v,0)$. It is calculated during initialization and is constant over time. Any arbitrary feature can be employed in the model.

**Attribute of Edges.** For each edge $e = (v,w) \in \mathbf{E}$ the length $l : \mathbf{E} \times \mathcal{T} \mapsto \mathbf{R}; l(e,t) = ||\mathbf{p}(v,t) - \mathbf{p}(w,t)||_2$ is the Euclidean distance of the vertices $v$ and $w$ at time $t$. These lengths are updated at each frame to deal with global scaling.

**Attribute of Triangles.** Each triangle $f \in \mathbf{F}$ stores the ratios of its edge lengths $\mathbf{r} : \mathbf{F} \times \mathcal{T} \mapsto \mathbf{R}^3$, where $\mathbf{r}(f,t) = \{\frac{l(e_1,t)}{l(e_2,t)}, \frac{l(e_1,t)}{l(e_3,t)}, \frac{l(e_2,t)}{l(e_3,t)}\} = \{r_{12}^t, r_{13}^t, r_{23}^t\}$ are their ratios at time $t$. These ratios are updated at each frame.

## 3 Appearance-Based Tracker

Mean Shift [7] is employed as appearance-based tracker to associate the vertices of the graph over time. In each frame, it finds the locally optimal position $\mathbf{p}$ for each vertex $v$. This is achieved in an iterative process, where starting from the position from the last frame, Mean Shift searches in a local neighborhood for a position which maximizes the similarity $\mathcal{A} : R^n \times R^n \mapsto [0,1]$ to the appearance $\mathbf{a}(v)$ of the model. Similarity in appearance is determined as follows:

$$\mathcal{A}(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v,t_i))) = 1 - \delta(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v,t_i))), \tag{1}$$

where $\mathbf{I} : \mathbf{R}^2 \mapsto \mathbf{R}^n$ extracts a feature vector around position $\mathbf{p}(v,t_i)$. $\delta$ can be any distance metric suitable for the employed features. In this paper, it is the Bhattacharyya distance as described in [7]. The offset generated at time $t_i$ points to the position maximizing $\mathcal{A}$: $\mathbf{p}(v,t_i) = \mathbf{p}(v,t_{i-1}) + \boldsymbol{m}(v,t_i)$.

## 4 Structural Cue Based on Edges

Under the assumption that the target object is rigid and its motion is limited to the image plane, the length of edges does not change over time. Fig. 1 visualizes the idea behind the edge cue. This cue has already been presented in a similar form in [8], but has been improved and simplified for this paper.

The edge cue is determined several times during the iterative process (see Sec. 6) in each frame of a video. $t_i$ indicates a point in time within the current frame starting at time $t_0$. For each vertex $v \in \mathbf{V}$ an edge cue can be determined

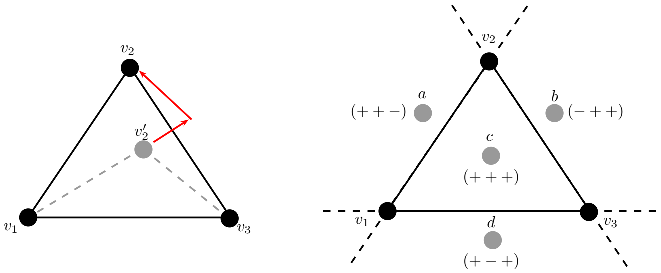


**Fig. 1.** Structural cues. Left: edge cue; Right: triangle cue.

based on the local, spatial deformation of graph $\mathbf{G}$. The local (deformation) energy $\mathcal{E}$ in vertex $v$ at time $t_i$ can be quantified as follows:

$$\mathcal{E} : \mathbf{V} \times \mathcal{T} \mapsto \mathbf{R}; \mathcal{E}(v, t_i) = \min\left(1, \sum_{e=(v,w)\in\mathbf{E}} \left|1 - \tfrac{||\boldsymbol{p}(v,t_i)-\boldsymbol{p}(w,t_i)||_2}{l(e,t_0)}\right|\right). \tag{2}$$

$\mathcal{E}$ is a weight used to calculate the edge cue $\boldsymbol{d} : V \times \mathcal{T} \mapsto \mathbf{R}^2$:

$$\boldsymbol{d}(v, t_i) = \sum_{e=(v,w)\in\mathbf{E}} \mathcal{E}(w, t_i) \cdot |\ ||\boldsymbol{p}(v, t_i) - \boldsymbol{p}(w, t_i)||_2 - l(e, t_0)| \cdot \tfrac{\boldsymbol{p}(v,t_i)-\boldsymbol{p}(w,t_i)}{||\boldsymbol{p}(v,t_i)-\boldsymbol{p}(w,t_i)||_2}, \tag{3}$$

which is an offset vector pointing towards the structurally optimal position.

## 5 Structural Cue Based on Triangles

Triangles are 2D entities, which are able to describe the geometry of planar objects and approximate curved objects (triangle mesh). By knowing the correspondence of three points at two time instances, it is possible to estimate their affine transformation in and out of the image plane.

Barycentric coordinates are an elegant way to transfer the motion information of a triangle to the neighboring vertices in a graph. The position of a vertex $v$ can be calculated with the help of the barycentric coordinates $\{\beta_1, \beta_2, \beta_3\}$ of the three corners $c(f)$ of any triangle $f \in \mathbf{F}$. Figure 1 illustrates this concept.

During the intra-frame, iterative process, the *triangle cue* for a vertex is determined from the barycentric coordinates of a triangle $f^* \in \mathbf{F}'$. Let $f^*$ be the triangle with the highest confidence $\mathcal{F} : \mathbf{F}' \times \mathcal{T} \mapsto \mathbf{R}$. $\mathcal{F}(f, t_i)$ is based on two properties of triangles: change of shape $\mathcal{R} : \mathbf{F} \times \mathcal{T} \mapsto \mathbf{R}$ and similarity in appearance $\mathcal{A}$ in their corners $\mathbf{c}(f)$ (see (1)). Change in ratios $\mathcal{R}$ is determined as $\mathcal{R}(f, t_i) = \min(|1 - \frac{r_{12}(t_i)}{r_{12}(t_0)}| + |1 - \frac{r_{13}(t_i)}{r_{13}(t_0)}| + |1 - \frac{r_{23}(t_i)}{r_{23}(t_0)}|, 1)$. From this, the confidence is calculated as follows:

$$\mathcal{F}(f, t_i) = \frac{1 - \mathcal{R}(f, t_i) + \min\limits_{v\in c(f)} (\mathcal{A}(\mathbf{a}(v_j), \mathbf{I}(\mathbf{p}(v_j, t_i)))}{2} \tag{4}$$

**Algorithm 1.** Combined, iterative tracking within one frame.

```
ITERATIVETRACKING
    $\varepsilon_i$, $\varepsilon_{\mathcal{A}}$, $\varepsilon_{\mathcal{E}}$ thresholds for iterations, similarity, energy
    $i \leftarrow 1$                                                    ▷ iteration counter
    while $i < \varepsilon_i \wedge (\underset{v \in \mathbf{V}}{\operatorname{argmin}}(\mathcal{A}(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v, t_i)))) < \varepsilon_{\mathcal{A}} \vee \underset{v \in \mathbf{V}}{\operatorname{argmax}}(\mathcal{E}(v, t_i)) > \varepsilon_{\mathcal{E}})$ do
        sort $\mathbf{V}$                                               ▷ for more details see Sec. 7
        for each vertex $v \in \mathbf{V}$ do
            determine appearance cue $\mathbf{m}(v, t_i)$ using Mean Shift
            if $i > 1$ then                                             ▷ first iteration is Mean Shift only
                determine structural cue $\mathbf{s}(v, t_i)$
                combine cues $\mathbf{p}(v, t_i) = \mathbf{p}(v, t_{i-1}) + (\omega \cdot \boldsymbol{m}(v, t_i) + (1 - \omega) \cdot \boldsymbol{s}(v, t_i))$
            else
                Mean Shift only $\mathbf{p}(v, t_i) = \mathbf{p}(v, t_{i-1}) + \boldsymbol{m}(v, t_i)$
            end if
        end for
        update: $\mathcal{A}(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v, t_i)))$, $\mathcal{E}$ and $\mathcal{V}$ of $v \in \mathbf{V}$, $\mathcal{F}$ of $f \in \mathbf{F}$
        $i \leftarrow i + 1$
    end while
  end
  update: $l$ of $e \in \mathbf{E}$ and $\mathbf{r}$ of $f \in \mathbf{F}$
```

The most stable $f^*(t_i)$ is selected by $f^*(t_i) = \underset{f \in \mathbf{F}'}{\operatorname{argmax}}(\mathcal{F}(f, t_i))$. Finally, the triangle cue $\mathbf{b} : \mathbf{F}' \times \mathcal{T} \times \mathbf{B} \mapsto \mathbf{R}^2$ is calculated from $\mathbf{B}(v, f^*)$:

$$\mathbf{b}(c(f^*), t_i, \mathbf{B}(v, f^*)) = (x, y, 1)^T = (\beta_1, \beta_2, \beta_3) \cdot \begin{pmatrix} \mathbf{p}(v_1, t_i)^T, 1 \\ \mathbf{p}(v_2, t_i)^T, 1 \\ \mathbf{p}(v_3, t_i)^T, 1 \end{pmatrix} \tag{5}$$

## 6 Combined, Iterative Tracking

The following combined, iterative tracking integrates structural cues into the mode seeking process of Mean Shift (see Sec. 3). By combining the appearance cue $\mathbf{m}(v, t_i)$ with the structural cue $\mathbf{s}(v, t_i)$ (either $\mathbf{d}(v, t_i)$ or $\mathbf{b}(v, t_i) - \mathbf{p}(v, t_{i-1})$) the proposed approach finds a position, which not only maximizes the similarity in appearance, but also the similarity in structure (shape).

During the intra-frame iterations, $\mathbf{s}$ and $\mathbf{m}$ are re-calculated and combined for each vertex $v$ until a position $\mathbf{p}(v, t_i)$ is found where $\mathcal{E}(v, t_i) < \varepsilon_{\mathcal{E}}$ (see (2)) and $\mathcal{A}(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v, t_i))) > \varepsilon_{\mathcal{A}}$ (see (1)). $\mathbf{p}(v, t_i) = \mathbf{p}(v, t_{i-1}) + (\omega \cdot \boldsymbol{m}(v, t_i) + (1 - \omega) \cdot \boldsymbol{s}(v, t_i))$, where $\omega$ is a weight defining the influence of appearance $\mathbf{m}$ and structure $\mathbf{s}$ on the new position.

There are three ways to come up with $\omega$: (i) similarity in appearance $\mathcal{A}$ (see 1), (ii) energy in a vertex $(1 - \mathcal{E})$ (see 2) or (iii) confidence in a vertex $\mathcal{V}$ (see 6). The confidence $\mathcal{V}$ of vertex is determined by combining similarity and energy:

$$\mathcal{V} : \mathbf{V} \times \mathcal{T} \mapsto \mathbf{R}; \mathcal{V}(v, t_i) = \frac{\mathcal{A}(\mathbf{a}(v), \mathbf{I}(\mathbf{p}(v, t_i))) + (1 - \mathcal{E}(v, t_i))}{2}, \tag{6}$$

In Alg. 1, there are two categories of updates: intra-frame and inter-frame. The **intra-frame updates** on $\mathcal{A}$, $\mathcal{E}$, $\mathcal{V}$ and $\mathcal{F}$ are necessary for the combined, iterative process, and the **inter-frame updates** on the lengths $l$ and ratios $\mathbf{r}$ are necessary to adjust the model to global scaling.

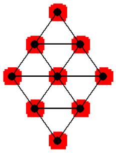
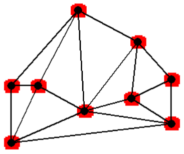
| Regular-sized triangulation | | Irregular-size triangulation | |
|---|---|---|---|
| $\mathbf{p}(v_1, 0) = (10, 50)^T$<br>$\mathbf{p}(v_2, 0) = (40, 50)^T$<br>$\mathbf{p}(v_3, 0) = (70, 50)^T$<br>$\mathbf{p}(v_4, 0) = (40, 10)^T$<br>$\mathbf{p}(v_5, 0) = (40, 90)^T$<br>$\mathbf{p}(v_6, 0) = (25, 30)^T$<br>$\mathbf{p}(v_7, 0) = (55, 30)^T$<br>$\mathbf{p}(v_8, 0) = (25, 70)^T$<br>$\mathbf{p}(v_9, 0) = (55, 70)^T$ |  | $\mathbf{p}(v_1, 0) = (60, 20)^T$<br>$\mathbf{p}(v_2, 0) = (30, 80)^T$<br>$\mathbf{p}(v_3, 0) = (100, 90)^T$<br>$\mathbf{p}(v_4, 0) = (65, 100)^T$<br>$\mathbf{p}(v_5, 0) = (10, 125)^T$<br>$\mathbf{p}(v_6, 0) = (105, 45)^T$<br>$\mathbf{p}(v_7, 0) = (130, 75)^T$<br>$\mathbf{p}(v_8, 0) = (130, 110)^T$<br>$\mathbf{p}(v_9, 0) = (10, 80)^T$ |  |

**Fig. 2.** Graphs of synthetic sequences with their vertices at time $t = 0$. Please note that the proposed approaches is not limited to graphs with 9 vertices.

**Table 2.** Videos used in evaluation are made up of every possible combination in this table. $T$ = Translation; $R$ = Rotation; $S$ = Scaling; $D$ = Distractors; $O$ = Occlusion

| **Layout of G** | **2D Transformations in each frame** | **Challenges** |
|---|---|---|
| regular-sized | $T = (5, 4)^T$ | $D$ |
| 9 vertices | | $D$; $O$: 1 vertex |
| | $T = (7, 5)^T$; $R = \begin{pmatrix} \cos(10^\circ) & \sin(10^\circ) \\ -\sin(10^\circ) & \cos(10^\circ) \end{pmatrix}$ | $D$; $O$: 3 vertices |
| irregular-sized | | $D$; $O$: 6 vertices |
| 9 vertices | $T = (2, 1)^T$; $R = \begin{pmatrix} \cos(5^\circ) & \sin(5^\circ) \\ -\sin(5^\circ) & \cos(5^\circ) \end{pmatrix}$; | $D$; Gaussian white noise |
| | $S = \begin{pmatrix} 1.02 & 0 \\ 0 & 1.02 \end{pmatrix}$ | $D$; Salt & Pepper 10 % |

## 7 Evaluation of Structural Cues

Tab. 2 shows information about the 36 synthetic videos (size = 400 × 600) which are used for this evaluation. Fig. 2 visualizes the two graphs used in the synthetic videos. All vertices have the same appearance $\mathbf{a}(v)$, which makes it difficult for trackers to distinguish between them (challenge: distractors). As a feature, we extracted weighted color histograms around the position of each vertex in a 11 × 11 neighborhood. Three different choices for $\omega$ are evaluated: $0 = \mathcal{A}$; $1 = (1 - \mathcal{E})$; $2 = \mathcal{V}$. Additionally, three different orderings of $v \in \mathbf{V}$ are studied: 0 = fixed ordering; 1 = ascending by $\mathcal{V}$; 2 = descending by $\mathcal{V}$. This results in nine different sets of parameters $\{00, 01, 02, 10, 11, 12, 20, 21, 22\}$ and 324 $(36 \cdot 9)$ test cases for each cue.

The results can be seen in Fig. 3 and 4, where the curves visualize the mean error (Euclidean distance from ground truth position averaged over all vertices in graph) in a vertex at each frame. For both structural cues, the choice of $\omega$ and the ordering of $\mathbf{V}$ have a noticeable influence on the results. For all test cases, the best result of the triangle cue is superior against the best result of the edge cue. The best parameter set for the triangle cue is $\{20\}$ and the worst is $\{00\}$. For the edge cue the best is $\{00\}$ and the worst $\{10\}$. The best parameters for the triangle cue are able to achieve a total error (summed over all test cases) of only $\approx 345$, whereas the best edge cue results in $\approx 1994$.

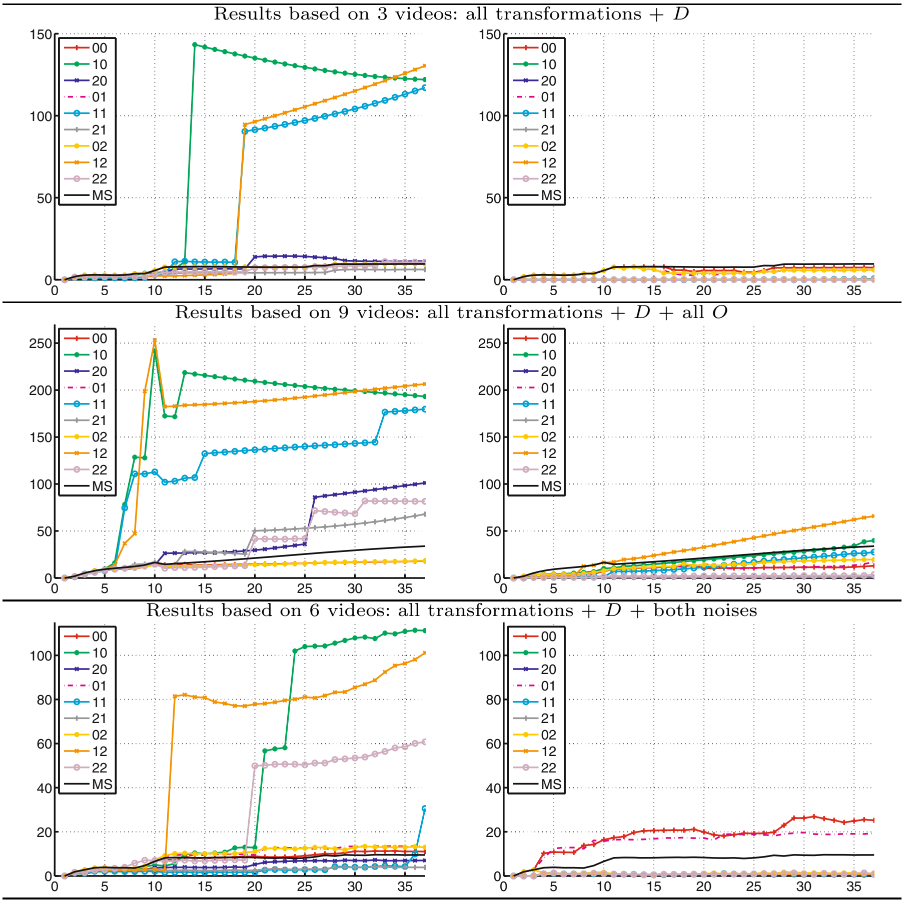
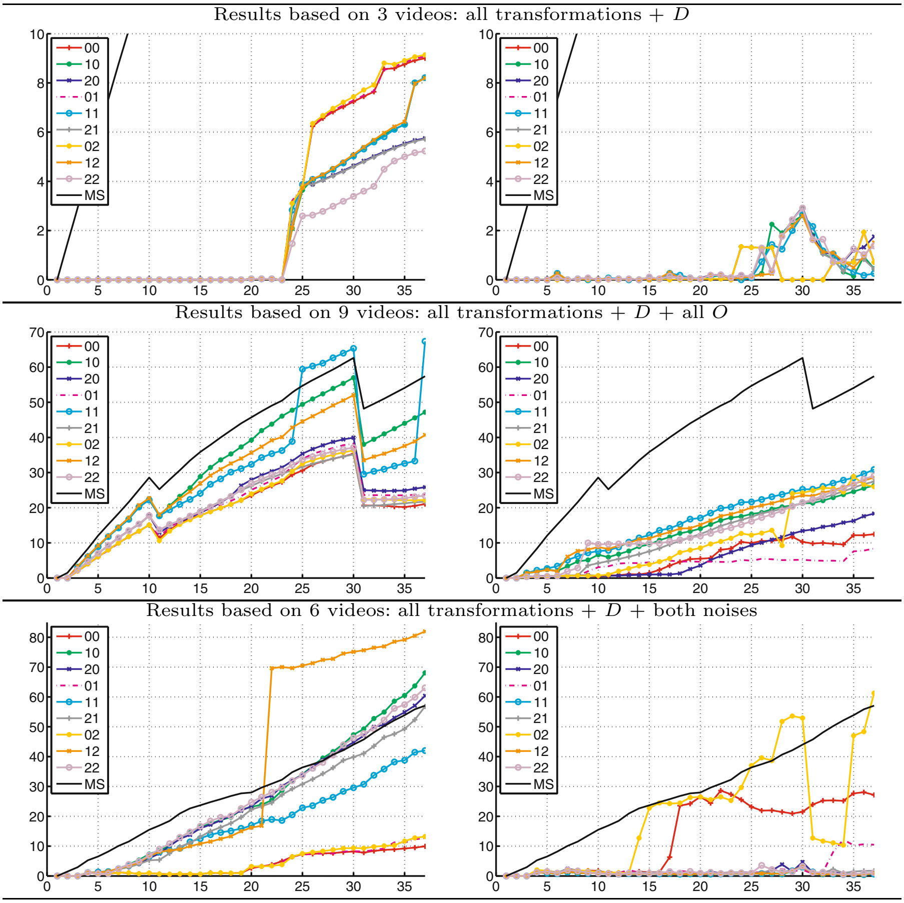


**Fig. 3.** Results with regular-sized triangulation. Left: edge cue; Right: triangle cue. Vertical axis: error; Horizontal axis: frame. MS = Mean Shift (baseline approach).

There are several drawbacks to the edge cue: The quality of this structural cue highly depends on the layout of the edges in the graph. Furthermore, as edges are a one dimensional entity, they are only capable of providing distance information. As this cue is local, there is no direct influence from vertices further away in the graph. Information propagates throughout the whole graph, but in challenging cases this can be problematic.

## 8 Conclusions and Future Work

In this paper, we studied the potential of structural cues in 2D tracking of multiple targets. An attributed graph acted as a model describing the structure of the target object. Iterative tracking combined hypotheses of the appearance-based trackers with the structural cues deduced from the model to establish


Results based on 3 videos: all transformations + $D$

Results based on 9 videos: all transformations + $D$ + all $O$

Results based on 6 videos: all transformations + $D$ + both noises

**Fig. 4.** Results with irregular-sized triangulation. Left: edge cue; Right: triangle cue. Vertical axis: error; Horizontal axis: frame. MS = Mean Shift (baseline approach).

temporal correspondences. This paper evaluated two different structural cues: edge cue and triangle cue. The results of the evaluation showed the superiority of the triangle cue. In the future, we plan to apply the triangle cue in tracking articulated objects and extend this approach to 3D.

## References

1. Maggio, E., Cavallaro, A.: Video Tracking: Theory and Practice. Wiley (2011)
2. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. 38(4) (2006)
3. Felzenszwalb, P.F.: Pictorial structures for object recognition. IJCV 61, 55–79 (2005)
4. Ramanan, D., Forsyth, D.: Finding and tracking people from the bottom up. In: CVPR, vol. 2, pp. 467–474. IEEE (2003)

5. Salzmann, M., Hartley, R., Fua, P.: Convex optimization for deformable surface 3-d tracking. In: IEEE 11th International Conference on Computer Vision, pp. 1–8 (2007)
6. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. Circuits and Systems for Video Technology 16(9), 1107–1124 (2006)
7. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI 25(5), 564–575 (2003)
8. Artner, N.M., Ion, A., Kropatsch, W.G.: Multi-scale 2d tracking of articulated objects using hierarchical spring systems. PR 44(4), 800–810 (2011)

# Hand-Raising Gesture Detection with Lienhart-Maydt Method in Videoconference and Distance Learning

Tiago S. Nazaré and Moacir Ponti

Instituto de Ciências Matemáticas e de Computação — Universidade de São Paulo
13566-590 São Carlos, SP, Brazil
tiagosn@grad.icmc.usp.br, moacir@icmc.usp.br
http://www.icmc.usp.br/~moacir

**Abstract.** In video-conference and distance learning videos, the moment that someone makes a hand-raising gesture is relevant to be included in the video annotation. However, gesture recognition can be challenging in such scenarios. We propose a system to detect faces, the hand-raising gesture and annotate the video. The Lienhart-Maydt object detection method is used, in which each frame is classified. Then, the gesture is detected by analyzing intervals of frames. Our approach was tested in videos with several characteristics. The results show that our method can deal with illumination and background variations, is able to detect multiple gestures and it is robust to confusing gestures. Besides it allow the use of moving cameras.

**Keywords:** Video processing, gesture detection, video annotation.

## 1 Introduction

Gesture recognition is a challenging task that is often addressed with complex sensors and methods, such as the use of depth sensors and multiple classifier systems, under controlled acquisition conditions [8]. Simpler methods, on the other hand, are not useful to specific applications since there is a lower concern about false positives and false negatives minimization. When the problem of gesture is more specific, it is possible to find more viable solutions. It is the case of hand-rising detection, applied to video-conference and distance learning videos to facilitate the annotation task.

In order to help the video annotation, we propose a system to detect faces and the hand raising gesture, i.e. an open hand raised, with the palm of the hand facing forward. A study is presented to investigate the robustness of already existing methods under several conditions and resolutions. A combination of methods is used, in special the improved Viola-Jones or Lienhart-Maydt method [6].

Kölsch e Turk [4],[5] showed that the object visual detector proposed by Viola and Jones [10], originally proposed to detect faces, could be used to detect hand poses. Later, Lienhart e Maydt [6] developed a more efficient method, based on

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 512–519, 2013.

the Viola-Jones original method, not used by Kölsch e Turk [4],[5] to evaluate their experiments.

We believe this investigation can shed a light about the use of this kind of method under conditions such as camera alternation, filming with a moving camera, and illumination changing such as when someone turn on/off the lights. The contributions of our study can be summarized in three parts: i) a **method for a hand-raising gesture detection**, ii) the **study of the limitations of the Lienhart-Maydt method** for this application, iii) and a **new dataset of images** to detect an open hand gesture.

## 2 Related Work

Among related work that specifically address this problem, Yao and Cooperstock [11] assumed that the heads of people in an audience are captured by the camera in a single horizontal line. It looks for movement and human skin in regions above the heads. When such events occur, a straight line is fitted using average points, if the slope of the line is between 45 and 135 degrees, a hand raising gesture is considered to be detected.

Duan and Liu [2] address the problem using indoor human silhouette analysis. It is able to work with moving people and groups. However the camera should remain still. The general pipeline of the method is foreground detection, followed by blob detection, candidate regions extraction (connected components located above the silhouettes), feature extraction using an R-transform and, finally a classification that looks for an arm or raising hand.

This study aims to improve the issues on those related work. We studied different acquisition conditions, including camera movement, changing in illumination, background and partial occlusion.

## 3 Viola-Jones and the Improved Lienhart-Maydt Method

The Viola-Jones method uses integral images and Haar-like operators to obtain several features. A boosting approach [3] is used to select a reduced number of visual descriptors to handle the problem. Finally, it uses a combination of classifiers in cascade, with increasing complexity. This cascade approach eliminates regions of low similarity, dedicating more effort on the classification of regions that are similar to the object of interest. Lienhart and Maydt [6] introduced two changes in the original algorithm: i) a new set of rotated Haar-like features [7] and ii) an improvement on the cascade classifiers based on a stage post-optimization scheme. The authors indicate an increase of 23.8% in the overall performance.

The Lienhart-Maydt [6] version of the object detector proposed by Viola and Jones [10] is used in this study to detect faces and hands in video frames, so that we can calculate the relative position between faces and hands and then detect a hand raising gesture, as described in next section.

# 4 Gesture Detection and Video Annotation

## 4.1 Method

Using the Lienhart-Maydt method, we look for both faces and hands on each video frame. If it detects both faces and hands, we compute the relative position between them, to check if it is compatible with a hand raising gesture.

We consider a hand-raising gesture when a person raises his/her hand open next to his/her face. For this reason, the height of the hand should be at least in the line of shoulders. Thus, for each face the algorithm search for a hand surrounding the face. In order to reduce the search space, only a region proportional to two and a half (2.5) faces is considered to search for a hand in the horizontal direction, both right and left. In the vertical direction, the region proportional to three (3) faces above and half (0.5) face below. Those choices are explained by the average size of proportions face and arm in human beings, as depicted in Figure 1.

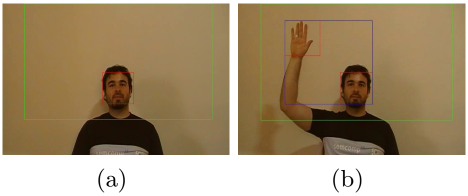


**Fig. 1.** Gesture search space: (a) face detected is showed using a red rectangle and the search space in a green rectangle; (b) hand detected is showed inside a red rectangle and the blue rectangle show the relation between hand and face.

Since the algorithm must detect both right and left hands, it should be trained with both right and left examples. In order to avoid this we trained only with left hands. To detect the right hands, we flip the image and than perform a second search.

A frame labeled as **positive** is those in which at least one pair face-hand satisfies the conditions cited before, otherwise it is considered **negative**. The Algorithm 1 summarizes the steps for the whole procedure.

After all frames are classified, we look for an interval of hand-face detections. A hand-raising gesture is detected inside a time interval when:

1. The duration of the gesture is at least 1 second;
2. The first and last frames are considered positive;
3. Between the first and last frames there are at least 80% of positive frames;
4. There are no sequences of negative frames with duration of more than 1 second of video.

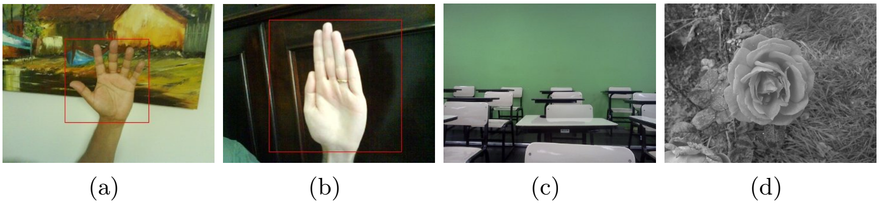
**Algorithm 1.** Hand-raising gesture detection

```
for each video frame do
    detect faces
    detect left hands
    flip the image
    detect right hands
    for each face detected do
        if hand is detected in the search space then
            label the frame as positive
        end if
    end for
end for
```

### 4.2 New Training Hand Dataset and Implementation

We used the OpenCV library version 2.1 [1] was used to implement the detector. This library has cascade classifiers trained to detect faces, available in XML files. It is also possible to create new cascade classifiers using positive and negative examples, and also store it in a XML file.

The face detection was performed using the already available classifier. To detect hands, we used 905 images of open left hands with different backgrounds, illumination variations, and changing finger positions (open and close). We also collected 1000 negative examples, both RGB and grayscale images (most of grayscale images were collected from [9]). In Figure 2 negative and positive examples are shown. The image dataset and the XML files are available in the project webpage [1].



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

**Fig. 2.** Examples of images collected: (a-b) positive examples (c-d) negative examples

## 5 Experiments

A total of 16 videos were produced in order to test the method. The Table 1 summarizes the characteristics of each video. Three of those videos were recorded in a distance learning context (# 5, 6 and 16), and the remaining in videoconferencing context. The resolutions indicated are: (A) 720$x$480, (B) 640$x$480,

[1] `http://www.icmc.usp.br/~moacir/project/VideoProcessing/`

(C) 1280$x$720. In order to check the robustness of the method, some situations were included in the videos, as indicated in Table 1:

1. Artificial illumination variation during the video (switching lights on and off, open a window, etc.);
2. Confusing gestures (scratch the head, spreading the arms, holding the head with an open hand, etc.);
3. External (natural) illumination;
4. Camera movement;
5. Multiple gestures (two or more people raising their hands simultaneously);
6. Partial occlusion of the hand and/or face;
7. Two or more people in different distances to the camera;
8. Variation of hand position (scale) in the same gesture;

**Table 1.** Characteristics of each video

| Video ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Persons | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 13 |
| Tests | 1 | 2 | 1,2 | 3,4 | 5,6 | 5 | 6,2 | 6,2 | 6 | 6,2 | 6,2 | 6 | 1,2,6 | 1,8 | 2 | 4,5,7 |
| Resolution | A | A | A | A | A | A | B | B | B | B | B | B | B | B | B | C |
| FPS | 29 | 29 | 29 | 29 | 29 | 29 | 24 | 25 | 24 | 24 | 24 | 24 | 25 | 15 | 30 | 29 |

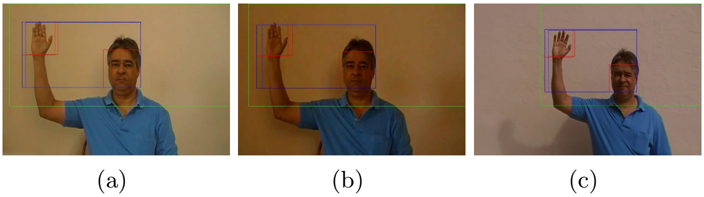

(a) (b) (c)

**Fig. 3.** Neighbor frames with sudden change in ambient lighting and camera movement (a) starting illumination condition (b) lights turned off (c) natural illumination and camera movement

Sudden changes in the illumination during the video are tested in videos #1 and #3. It significantly modifies the pixel values, which often is an issue in movement-based gesture detection [2]. An example of such change (lights turned off and natural illumination), as well as camera movement is shown in Figure 3.

Another issue is the false detection of gestures that are close to a raising hand. During a class or videoconference a person might spread the arms, scratch the face or head, or do something that can be confused with a raising hand gesture. It happens in some videos such as in the examples of Figure 4. Besides, the most difficult problem to overcome when using the Lienhart-Maydt method is the occlusion and partial occlusion. Some examples are shown also in Figure 4.

In both videoconferencing and classroom contexts, there are often multiple gestures and people positioned in different distances in relation to the camera. The videos #3, #5, #6 e #16 include such scenarios as depicted in Figure 5.

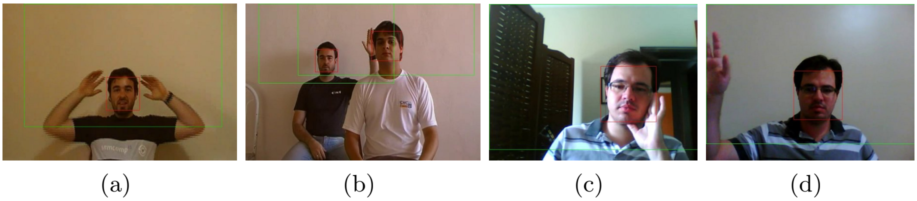


**Fig. 4.** Examples of confusing gestures (a) and (c), and partial occlusion (b) and (d)

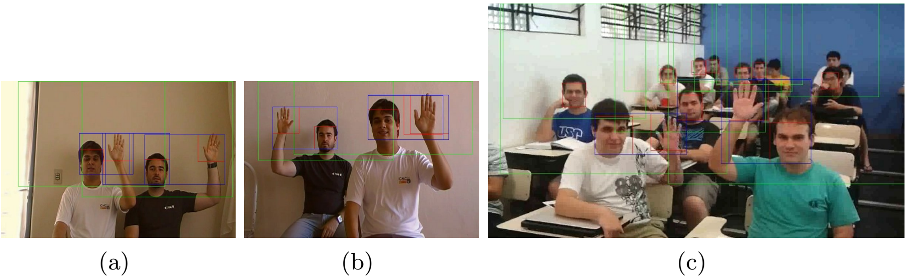


**Fig. 5.** Detected frames with multiple simultaneous gestures

## 6 Results and Discussion

The Table 2 shows, for each video, *TP*: true positive rate, *FP*: false positive rate, *FN*: false negative rate, *TN*: true negative rate, *FPS*: frames per second, precision, accuracy and running time (seconds).

All processing is performed offline, after the video is recorded, and the video annotated for future search.

The proposed method achieved the following robustness results:

- **Illumination changes**: did not affected the performance, including when the change occur during a gesture.
- **Confusing gestures**: did not affected the performance, the method seems to be very good on discharging false or confusing hand gestures.
- **Occlusion**: the method cannot handle partial occlusion in the external borders of the frame, since the detector cannot center a rectangle in the region. This is the reason why the videos #7–13 have lower accuracy and precision values. For partial occlusion inside the frame, of both hands and faces, we observed that the method could not handle more than 15% of occlusion.
- **Multiple gestures and scale variation**: the method handle well multiple gestures, as indicates the results for videos #5, #6 and #16. The scale is also not an issue if well managed. A case of failure is the video #16 in which gestures are not detected when people are too far from the camera, as shown in Figure 5. For this reason, it is important to control the audience position in order to assure a good result.
- **Camera movement**: it did not affected the results, all gestures were detected under moderate movement.

**Table 2.** Results

| ID# | TP | FP | FN | TN | FPS | Precision | Accuracy | Running Time (s.) |
|---|---|---|---|---|---|---|---|---|
| 1 | 838 | 4 | 10 | 811 | 29 | 0.9952 | 0.9915 | 545 |
| 2 | 401 | 7 | 6 | 887 | 29 | 0.9828 | 0.9900 | 291 |
| 3 | 1103 | 28 | 1 | 815 | 29 | 0.9752 | 0.9851 | 739 |
| 4 | 699 | 11 | 7 | 403 | 29 | 0.9845 | 0.9839 | 707 |
| 5 | 973 | 54 | 84 | 418 | 29 | 0.9474 | 0.9097 | 831 |
| 6 | 429 | 29 | 18 | 363 | 29 | 0.9366 | 0.9439 | 348 |
| 7 | 2 | 5 | 91 | 454 | 24 | 0.2857 | 0.8260 | 299 |
| 8 | 68 | 8 | 355 | 712 | 25 | 0.8947 | 0.6824 | 627 |
| 9 | 42 | 2 | 180 | 339 | 25 | 0.9545 | 0.6767 | 268 |
| 10 | 112 | 0 | 182 | 254 | 24 | 1.0000 | 0.6678 | 151 |
| 11 | 167 | 25 | 162 | 369 | 24 | 0.8697 | 0.7413 | 226 |
| 12 | 35 | 1 | 223 | 167 | 24 | 0.9722 | 0.4741 | 179 |
| 13 | 159 | 115 | 76 | 460 | 24 | 0.5802 | 0.7641 | 243 |
| 14 | 234 | 0 | 55 | 69 | 15 | 1.0000 | 0.8463 | 128 |
| 15 | 645 | 52 | 286 | 618 | 30 | 0.9253 | 0.7888 | 892 |
| 16 | 168 | 27 | 94 | 114 | 29 | 0.8615 | 0.6997 | 805 |

## 7 Conclusion

The Lienhart-Maydt method, used as basis for our method, was able to overcome many issues of previous works that tried to detect the same gesture, since it is robust to scale and illumination changes. The frame-by-frame analysis and the smoothness of the gesture detection in intervals is probably the cause of the success in other conditions such as camera movement and multiple gestures. Our method is not dependent on a specific camera or pose, can handle variations in illumination even during the gesture, is able to detect gestures with moving cameras, and work with different backgrounds and groups of people.

The drawbacks of the method include failure of detecting faces and hand with partial occlusion, and the necessity of filming the audience facing front, i.e., towards the camera, with small tolerance for angles (up to 15, as tested by video #16). Also, the detector runs three times, to detect faces, right and left hands, hindering the possibility of online processing. For annotation purposes it is not an issue since it is often performed after the recording.

It is important to note that the method has flexibility to detect different objects, so it can be a good choice to help on semi-automatic annotation systems. To improve the running time, we suggest the use of GPUs to implement the detector. Also, a better occlusion treatment is a matter of future studies.

**Acknowledgment.** The authors are grateful to CNPq for the student scolarship and FAPESP (proc. n. 2011/16411-4) for supporting the project.

## References

1. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
2. Duan, X., Liu, H.: Detection of hand-raising gestures based on body silhouette analysis. In: IEEE Int. Conf. Robotics and Biomimetics, pp. 1756–1761 (2009)
3. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proc. 13th International Conference on Machine Learning (ICML 1996), pp. 148–156 (1996)
4. Kölsch, M., Turk, M.: Analysis of rotational robustness of hand detection with a viola-jones detector. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 3, pp. 107–110 (2004)
5. Kölsch, M., Turk, M.: Robust hand detection. In: 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Seoul, Korea, pp. 614–619 (2004)
6. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. In: IEEE Int. Conf. Image Processing (ICIP), pp. 900–903 (2002)
7. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 349–361 (2001)
8. Qin, S., Zhu, X., Yang, Y., Jiang, Y.: Real-time hand gesture recognition from depth images using convex shape decomposition method. Journal of Signal Processing Systems, 1–12 (2013)
9. Seo, N.: Tutorial: OpenCV Haar training (rapid object detection with a cascade of boosted classifiers based on Haar-like features) (2011), `http://tutorial-haartraining.googlecode.com/svn/trunk/data/negatives/`
10. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001), pp. 511–518 (2001)
11. Yao, J., Cooperstock, J.R.: Arm gesture detection in a classroom environment. In: Proc. 6th IEEE Work. Appl. Computer Vision, pp.153–157 (2002)

# Statistical Analysis of Visual Attentional Patterns for Video Surveillance

Giorgio Roffo[1], Marco Cristani[1,2], Frank Pollick[3], Cristina Segalin[1], and Vittorio Murino[2,1]

[1] Department of Computer Science, University of Verona (IT)
[2] Pattern Analysis and Computer Vision Dept., Istituto Italiano di Tecnologia (IT)
[3] School of Psychology, University of Glasgow (UK)

**Abstract.** We show that *the way* people observe video sequences, other than *what* they observe, is important for the understanding and the prediction of human activities. In this study, we consider 36 surveillance videos, organized in four categories (*confront, nothing, fight, play*): the videos are observed by 19 people, ten of them are experienced operators and the other nine are novices, and the gaze trajectories of both populations are recorded by an eye tracking device. Due to the proved superior ability of experienced operators in predicting violence in surveillance footage, our aim is to distinguish the two classes of people, highlighting in which respect expert operators differ from novices. Extracting spatio-temporal features from the eye tracking data, and training standard machine learning classifiers, we are able to discriminate the two groups of subjects with an average accuracy of 80.26%. The idea is that expert operators are more focused on few regions of the scene, sampling them with high frequency and low predictability. This can be thought as a first step toward the advanced automated analysis of video surveillance footage, where machines imitate as best as possible the attentive mechanisms of humans.

**Keywords:** surveillance, gaze control, eye movement analysis, activity recognition, eye tracking.

## 1 Introduction

The study of eye movements is an innovative way of assessing the skill in monitoring of Closed Circuit Television (CCTV) recording, in which a comparison of the eye movement strategies between experienced operators and novice observers may show important differences that could be used in training an automatic monitoring system. Generally, when we are looking at a video, we consciously or unconsciously focus only on a fraction of the total information that we could potentially process, in other words we perform a perceptual selection process called *attention*. Visually, this is most commonly done by moving our eyes from one place of the visual field to another; this process is often referred to as a change in *overt attention* – our gaze follows our attention shift. The process of selecting

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 520–527, 2013.

visual information is crucial for the subsequent activity understanding, where internal mental representations are built for categorizing the observed events and starting to reason on them, for example to predict future actions.

In this paper, we focus on extracting the spatio-temporal eye patterns which regulate the attentive processes of experienced operators, looking if they substantially differ from those of novice people. Due to the higher ability of experienced operators in predicting violence in surveillance footage [12], we argue that understanding the way visual information is processed can be important for automated video surveillance.

Our approach aims at individuating where the focus of attention is located on the scene and the dynamics of this process. Considering gaze trajectories and modeling them in diverse fashions (e.g., encoding local curvatures, feeding them into heterogeneous classifiers as [6], etc.) did not reveal in our experiments significant differences between experts and novices. Therefore, we follow another strategy, which focuses on two different logical layers, spatial and temporal. Spatial analysis is performed by analyzing the zones of the screen considered most of the time: partitioning the image into cells and counting how many times they have been watched, indicates strongly different patterns among the two classes of observers. For the temporal characterization, we analyze the unpredictability of the movement patterns by adopting entropic measures, capturing in practice the irregularity of the eye trajectories. Spatial and temporal analyses are carried out with standard classifiers (SVM and kNN, respectively), and the fusion of the classification results allows one to consistently separate experts from novices, with an accuracy of 80.26%. In particular, we find that experts are characterized by a spatially more focused analysis (*they know where to look*) with a high level of unpredictability (basically, they switch continuously among different spatial cells), while novices tend to show more regularity in the analysis, considering a larger area of analysis, with a lower speed in accessing the data.

The rest of the paper is organized as follows. In Sec. 2, a review of the related literature is presented, and Sec. 3 details the proposed approach. Experiments are reported in Sec. 4, and, finally, Sec. 5 draws some conclusions and future perspectives.

## 2 Related Work

The selection of good CCTV operators is essential for effective CCTV system functioning. The study of gaze control mechanism is an intriguing way for evaluating the skills of entry level CCTV operators. Indeed, how gaze control operates over complex real-world scenes has recently become of central concern in several core cognitive science disciplines including cognitive psychology, visual neuroscience, and machine vision. For example, an application of psychological principles to Aviation Safety and Welfare (ASW) is suggested in [8], which analyzes the eye movements of expert and novice pilots while performing landings in a flight simulator. They found that expert pilots had significantly shorter dwells, more total fixations and they observe a specific place of interest in the visual scene. Experts were also found to have better defined eye-scanning patterns. In [11], authors conducted a comparison of the eye movement strategies between expert surgeons and novices, while performing a task on a computer-based laparoscopic surgery simulator: the results from eye gaze analysis showed that experts tended to maintain eye gaze on the target, whereas novices were more varied in their behaviours. In general, gaze control differs during complex and well-learned activities such as reading [14], tea and sandwich making [9], and driving [10].

Going back to surveillance, an ongoing research programme is investigating the ability of humans to detect whether or not an individual, captured on CCTV, is carrying weapons [5]. In [2], trained CCTV operators and lay people viewed footage material and were asked to indicate whether or not they thought the surveillance target was carrying a firearm. Our work is in line with this type of research.

## 3 Our Approach

Our approach partitions the screen in a set of 5 × 5 non-overlapped squared cells, of size 288 × 180 pixels each. From this support, we calculate two sets of features: the former models explicitly *where* the attention of the subject has been driven during the monitoring activity, and we call it *spatial feature set.* The latter indicates *how* the attentional analysis has been performed by the subjects, and we call it *temporal feature set.*

The spatial feature set is composed by one feature, which is the **Cell Counting (*Count*)**: a counting matrix, where the $i^{th}$ cell records exactly how many times a participant has been watching the $i^{th}$ cell of the grid. In practice, each videosequence can be summarized by a 25-dim count vector.

In the temporal feature set, the features have been designed upon three temporal basic cues that we will present below. The idea is that eye movement information is recorded, storing for each $i$-th cell a number $f(i)$ of basic cue values, where $f(i)$ indicates the number of times the $i$-th cell has been intercepted by an eye trajectory.

Three are the temporal basic cues:

- **Fixation Duration (*FIXd*)**: a fixation is the state of the eyes during which gaze is held upon a specific region. Humans typically alternate saccadic eye movements and fixations. The term "fixation" can also be referred to as the time between two saccades, during which the eyes are relatively stationary [7,16]. In our experiments, for each video analyzed by a subject, the time spent for each fixation in a particular cell has been recorded, expressed in ms. Therefore, for each cell we have a sequence of fixation duration values.
- **Saccades Velocity (*SACv*)**: the eyes do not remain still when viewing a visual scene; they have to move constantly to build up a mental "map" from interesting parts of the scene. The main reason for this is that only a small central region of the retina, the fovea, is able to perceive with high acuity. The simultaneous movement of both eyes is called a saccade. The duration

of a saccade depends on the angular distance the eyes travel during this movement, the so-called saccade amplitude. A saccade is individuated as a movement exceeding the threshold of $\tau = 30^\circ/sec$ starting after the fixation, lasting at least 20 ms [15,1]. For each cell we record all the saccades$'$ related speed values calculated over it, measured in $degrees/seconds$.

- **Smooth Pursuit Velocity (*PURv*)**: smooth pursuit is the eye movement that results from visually tracking a moving object. Generally, this kind of eye movement has a speed lower than $30^\circ/sec$ [13,16]. The PURv is measured in $degrees/seconds$ and the values are stored as for the previous cues.

In practice, as description of the whole monitoring analysis performed on a video sequence by a subject, we obtain three different cue volumes, each related to the $FIXd$, $SACv$ and $PURv$ feature. In the $i-$th entry of each volume we have all the $f(i)$ feature values collected in the $i-$th cell (i.e., depending on how many times that cell has been visited). At this point, to obtain a unique cue value for each $i-$th entry, we applied the mean operator. As a result, we obtained the $5 \times 5$ maps $\mu_x$, where $x$ stands for $FIXd$, $SACv$ and $PURv$.

At the end, in order to distill a single measure from each map, we calculate its *entropy*: this way, we obtained three entropic values for each analyzed videosequence, dubbed $E_{FIXd}$, $E_{SACv}$ and $E_{PURv}$. The underlying rationale of choosing entropic measures consists in the fact that the entropy gives a measure for assessing how unpredictable is the behavior of the subject: high entropy means that in the whole sequence the subject behaved in a very dynamic fashion, for example steadily focusing on some scene details, then suddenly moving the focus of attention toward distant screen locations. Viceversa, low entropy indicates that the subject kept repeated attentional patterns, patrolling in a mechanical fashion the screen.

Spatial and temporal features become the signature of the attentive behaviour of a single subject: given a pool of subjects belonging to the same class, our approach learns a classifier by employing linear Support Vector Machines (SVM) on the 25-dimensional spatial features, while the 3-dimensional temporal features are processed by kNN classifiers. The choice of the classification machinery supported us with satisfying results, as witnessed in the next section.

## 4 Experiments

In the experiments, we apply our approach to a recent video dataset provided by the University of Glasgow, whose content is detailed in the following.

### 4.1 The Dataset

The dataset has been taken from tens of urban surveillance cameras, highlighting “hot zones”, that is, crossroads near pubs and discotheque areas. In particular, thirty-six 16-second CCTV clips were used. These videos have been grouped in four categories (see Table 1), each composed by 9 videos: in the “Fight” category, behaviours leading up to a violent incident are shown; in the "Confront" category, a sequence of behaviours similar to the fight clip are shown, although no violent/harmful incident occurred; the "Play" category shows people interacting in a playful manner; finally, the "Nothing" category includes a variety of scenes where no violent/harmful behaviour occurs and they were taken from similar locations and with similar camera views. Please note that in the experiments, videos of the Fight category have been truncated, so that fights are not visible: this design was necessary to highlight solely the attentional behavior needed to understand the situation and predict the outcome, and not to analyze the outcome itself. The eye tracking experiment was attended by 19 participants, 10 CCTV operators (3 female, 7 male) aged 21-53 years ($\mu_{age} = 36.3$, $\sigma_{age} = 10.1$); and 9 novices (2 female, 7 male) aged 28-43 years ($\mu_{age} = 33.8$, $\sigma_{age} = 6.0$). All participants were native English speakers, naïve to the goals of the experiment and had not participated in eye tracking experiments in the past. All the participants had normal binocular (Titmus Test) and colour vision (CUCV Test) and corrected binocular visual vision acuity of 6/9 or better. Three of the participants wore eye glasses during the experiment, and two wore contact lenses. The device was an ASL Eye-Trac6 remote eye tracking device, located directly below the display screen and 0.65 meters from the participant's eye. A chin rest was used to minimise head movement and to maintain viewing distance. The video were displayed on a 19 inch LCD monitor with a set resolution of 1440 × 900 pixels which described a 37° × 23° field of view.

**Table 1.** Categories of CCTV clips. A violent incident was defined as an aggressive physical contact with intent to harm, such as a slap, shove, punch, or kick.

| | |
|---|---|
| **Fight clip** | Behaviours leading up to a violent incident. |
| **Confront clip** | Confronts which did not lead to a fight. |
| **Play clip** | People interacting and some playful encounter happens. |
| **Nothing clip** | Scenes where no violent/harmful behaviour<br>occurs, taken from similar locations and with similar camera pans. |

As preliminary analysis of the dataset, basic statistical analysis on standard features has been carried out. In particular, we consider the *mean fixation time* as the percentage of time a subject spends fixating when viewing the clip, the *mean fixation duration* as average duration of all the fixations on a given video and the *mean saccade rate* as the average number of saccades made per second. A main difference among clip categories was observed for the eye movement measures of gazing time and fixation duration. It indicates that there were significant differences in participants$'$ gazing time and fixation duration when viewing different types of clips. In particular:

- Participants exhibited significantly longer gazing time for clips in the matched confront clip category ($\mu = 80.08$, $\sigma = 3.66$), when compared to fight clips

($\mu = 78.31$, $\sigma = 3.99$, $p = 0.008$), to play clips ($\mu = 74.54$, $\sigma = 4.78$, $p < 0.001$) and to nothing clips ($\mu = 76.37$, $\sigma = 4.34$, $p < 0.001$).

- Although not statistically significant, a trend was found that CCTV operators spent lower proportion of time making fixations ($\mu = 76.16$, $\sigma = 4.19$) when compared to novice participants ($\mu = 78.5$, $\sigma = 4.8$). This may suggest that CCTV operators spent more time engaged in saccades and/or smooth pursuit tracking during the clip than novices.
- The mean fixation duration data revealed that CCTV operators exhibited a shorter mean fixation duration ($\mu = 0.34$, $\sigma = 0.02$) in comparison to novice participants ($\mu = 0.36$, $\sigma = 0.04$), even if this difference was not statistically significant.
- A third test was conducted to investigate if there were any significant differences in the mean rate of saccades due to participant experience. This analysis found no main effect of experience.

These results highlight differences between the two groups but do not explain what was observed by the subjects and in what way this happened.

### 4.2 Results

The goal of the classification was to divide novice people from expert operators and this was performed in the following way. First of all, we separate the analysis carried out on the spatial and the temporal features, to assess the contribute of each group of cues. In all the cases, Leave-One-Out cross validation was performed, considering a particular subject as test element, keeping the others as training samples, and exploring all the possible training/test partitions, averaging the classification values at the end. Since each subject watched 9 videos, we build 9 classifiers, i.e., one for each video. Given a test subject, we evaluate its "novice" or "expert" label by majority vote, considering the results of the 9 classifiers. For the *Count* spatial feature, we employ linear SVM as classifier, while for the entropic temporal features $E_{FIXd}$, $E_{SACv}$ and $E_{PURv}$ we adopt the kNN algorithm. The choice of these classifiers gave us the best performances, and their parameters have been chosen by cross-validation.

In the spatial analysis, some *Count* counting matrices have been reported in Fig. 1. Qualitatively, one can see that expert operators are more focused on a central smaller area (which collected the highest number of votes) while novices are more spread over the entire image plane. It is worth noting that these areas were populated by human subjects[1]. The quantitative results are reported in Tab. 2.

In the case of the entropic temporal features, for each subject we considered 9 kNN classifiers, one for each video. The results were quite higher than the spatial counterpart (Tab. 2). For evaluating the effect of including both the spatial and temporal features in the classification process, the majority vote was applied to the all the 18 classifiers, 9 for the spatial features and 9 for the temporal features. We do the same strategy for all the 19 subjects, averaging at the end

[1] The footage cannot be shown for ethical and privacy issues.

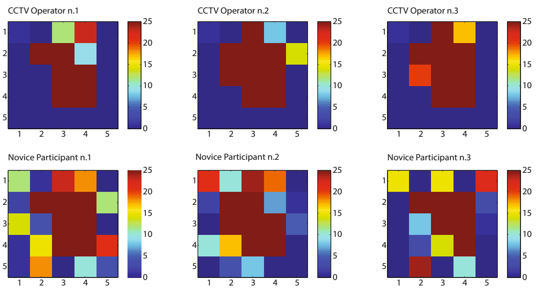


**Fig. 1.** Spatial analysis of the *Count* matrices. The figure shows that the CCTV operators focus on smaller areas than the novices.

**Table 2.** Classification rates while considering Temporal and Spatial cues separately and jointly (third column)

| Activity/Features | Temporal | Spatial | Joint |
|---|---|---|---|
| *Fight* | 78.9% | 68.4% | 84.2% |
| *Play* | 63.1% | 73.7% | 63.2% |
| *Nothing* | 84.2% | 78.9% | 84.2% |
| *Confront* | 73.7% | 68.4% | 89.5% |
| Average | 75.0% | 72.3% | 80.3% |

the accuracy scores obtained for each person. The results are shown in Table 2. We noted that:

- In general (apart from the *Play* class), temporal features were more effective in separating the two classes;
- In general (apart from the *Play* class), the fusion of spatial and temporal features was no worse than the single classifiers, showing a certain complementarity between the two different modeling schemes.

## 5 Conclusions

In this paper we presented an analysis which considers eye tracking data on video surveillance sequences. Our goal was to understand how expert CCTV operators analyze such videos, and if there is a difference with novice participants. Extracting spatio-temporal features, and training SVM and kNN classifiers, we have been able to discriminate the two groups of subjects with an average accuracy of 80.26%: the idea is that expert operators are more focused on few regions of the scene portraying the humans, sampling them with high frequency. This study follows the recent trend of applying a social signal processing perspective to surveillance [3,4], where psychological analyses are exploited to inspire more effective monitoring strategies. In particular, this can be thought as a first step toward the advanced automated analysis of video surveillance footage, where machines imitate as best as possible the attentive mechanisms of humans: in this case, the take-home message is that the dynamics with which people are observed is highly unpredictable but highly focused on them. Even if these results may appear intuitive, they have been obtained by a solid experimental analysis, for the first time.

## References

1. Bahill, A.T., Clark, M.R., Stark, L.: The main sequence, a tool for studying human eye movements. Math. Biosci. (2) (1975)
2. Blechko, A., Darker, I., Gale, A.: Skills in detecting gun carrying from CCTV. In: International Carnahan Conference on Security Technology (2008)
3. Cristani, M., Murino, V., Vinciarelli, A.: Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: CVPRW 2010, pp. 51–58 (2010)
4. Cristani, M., Raghavendra, R., Del Bue, A., Murino, V.: Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing 100, 86–97 (2013)
5. Hales, G., Lewis, C., Silverstone, D.: Gun Crime: The Market in and Use of Illegal Firearms. Findings (Great Britain. Home Office. Research, Development and Statistics Directorate). Home Office (2006)
6. Henderson, J.M., Weeks, P.A., Hollingworth, A.: Multi-feature object trajectory clustering for video analysis. IEEE Transactions on Circuits and Systems for Video Technology 18(11), 1555–1564 (2008)
7. Ji, R., Sun, X., Yao, H.: What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, vol. (3), pp. 1552–1559 (2012)
8. Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., Wickens, C.: Comparison of expert and novice scan behaviors during vfr flight. In: Proceedings of the 11th International Symposium on Aviation Psychology (2001)
9. Land, M.F., Hayhoe, M.: In what ways do eye movements contribute to everyday activities? Vision research 41(25-26), 3559–3565 (2001)
10. Land, M.F., Lee, D.N.: Where we look when we steer. Nature 369, 742–744 (1994)
11. Law, B., Atkins, M.S., Kirkpatrick, A.E., Lomax, A.J.: Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment, pp. 41–48 (2004)
12. Petrini, K., McAleer, P., Neary, C., Gillard, J., Pollick, F.E.: Experience in judging intent to harm modulates parahippocampal activity: an fmri study with experienced cctv operators. In: European Conference on Visual Perception (2012)
13. Pratt, J.: Visual fixation offsets affect both the initiation and the kinematic features of saccades. Experimental Brain Research 118(1), 135–138 (1998)
14. Rayner, K.: Eye movements in reading and information processing: 20 years of research.. Psychological bulletin 124(3), 372–422 (1998)
15. Torralba, A.: Modeling global scene factors in attention. Journal of the Optical Society of America. A, Optics, image science, and vision 20(5), 1407–1418 (2003)
16. Torralba, A., Castelhano, M.S., Oliva, A., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review 113 (2006)

# ReliefF-ML: An Extension of ReliefF Algorithm to Multi-label Learning

Oscar Gabriel Reyes Pupo[1], Carlos Morell[2], and Sebastián Ventura Soto[3]

[1] University of Holguín, Cuba
oreyesp@facinf.uho.edu.cu
[2] Universidad Central "Marta Abreu" de Las Villas, Cuba
cmorellp@uclv.edu.cu
[3] University of Córdoba, Spain
sventura@uco.es

**Abstract.** In the last years, the learning from multi-label data has attracted significant attention from a lot of researchers, motivated from an increasing number of modern applications that contain this type of data. Several methods have been proposed for solving this problem, however how to make feature weighting on multi-label data is still lacking in the literature. In multi-label data, each data point can be attributed to multiple labels simultaneously, thus a major difficulty lies in the determinations of the features useful for all multi-label concepts. In this paper, a new method for feature weighting in multi-label learning area is presented, based on the principles of the well-known ReliefF algorithm. The experimental stage shows the effectiveness of the proposal.

**Keywords:** multi-label learning, feature weighting, ReliefF algorithm.

## 1 Introduction

The multi-label problems have been actively studied in the last years. This is because it has been found that in many applications the multi-label data is a more natural and appropriate form of problem formulation and representation. Particular examples of such applications include text categorization [1], emotions evoked by music [2] and semantic annotation of images [3]. In all of these applications an instance space is typically represented by hundreds or thousands of features, therefore commonly there are features more relevant than others, and this situation affect the effectiveness of the machine learning algorithms.

Several supervised learning methods have been proposed to multi-label classification, however feature weighting and selection methods on multi-label data are less researched problems. How to make feature weighting on multi-label data is still lacking in the literature, furthermore multi-label feature weighting is still a challenging problem.

In this work, a filter-based feature weighting method called ReliefF-ML is proposed. ReliefF-ML is based on the principles of the well-known ReliefF algorithm [4]. Some properties of ReliefF-ML method are that it can be applied to

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 528–535, 2013.

both continuous and discrete problems, it includes interaction among features, and take into account the label dependences.

Due to the fact that lazy learning algorithms use a similarity or distance function based in feature space, these types of algorithms can be easily used to prove the effectiveness of feature weighting methods [5]. In this work, the approach ReliefF-ML was used as a feature weighting, not as a multi-label feature selection method; therefore the comparison with the existent multi-label feature selection methods in the literature was not carried out.

To evaluate the performance of ReliefF-ML, the accuracy of 3 multi-label lazy ranking algorithms using the feature weights provided by ReliefF-ML on 11 multi-label datasets from several fields were compared, showing the effectiveness of the proposal for multi-label problems.

This paper is organized as follows. In section 2, a formal definition of the multi-label learning task and related works on feature weighting methods to multi-label data is presented. In section 3, the ReliefF-ML approach is described. The experimental set up is described in section 4. An analysis of the experiment results appears in section 5. Finally, in section 6 the conclusion of this work are presented.

## 2 Background

### 2.1 Multi-label Learning

The multi-label learning is concerned with learning from examples, where each example is associated with multiple labels. In multi-label learning there can be distinguished two types of tasks: multi-label classification (MLC) and label ranking (LR). In the case of MLC, the goal is to construct a predictive model that will provide a list of relevant labels for a given test instance. On the other hand, the goal in LR is to construct a predictive model that will provide an ordering of the labels according to their relevance for a given test instance. The generalization of these two problems has been called multi-label ranking (MLR). [6]. In general, a multi-label dataset can be defined as follows:
-A feature space $\mathcal{F}$ that consists of tuples of values of primitive data types (discrete or continuos) $\forall x_i \in \mathcal{F}$, $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, where $D$ is the number of descriptive attributes. $x_i$ is the vector of features values for the instance $i$, where $x_{if}$ represents the value of $f$-th attribute for the instance $i$.
-A label space $\mathcal{L}$ with a cardinality equal to $Q$, where $Q$ is the number of labels in the dataset.
-A set of instances (examples) $E = \{(x_i, y_i) | x_i \in \mathcal{F}, y_i \subseteq \mathcal{L}, 1 \leq i \leq N\}$, where $N$ is the number of instances and $y_i$ is the set of relevant labels for the instance $i$. A label $l$ is relevant for an instance $i$ if the instance belongs to the class $l$, a label $l$ is irrelevant for an instance $i$ otherwise.

### 2.2 Related Works

The feature weighting process is a more general method than the feature selection task, in which the features are multiplied by a weight value proportional to the ability of the feature to distinguish pattern classes, whereas the feature selection problem assigns a weight restricted to the binary values 0 or 1 to a feature.

ReliefF [4] is a classical method for feature estimation. ReliefF is able to deal with incomplete and noisy data and can be used for evaluating the feature quality in multi-class problems. Commonly the ReliefF algorithm is used as a feature selection method, however it is a feature weighting method. The feature weighting is an important component of any lazy learning scheme. ReliefF was tested as feature weighting method in [5] and was found to be very useful to improve the performance of lazy algorithms.

In [7] was proposed a feature weighting method that learns a similarity metric to improve the performance of multi-label ranking lazy algorithms. The search process of the best weight vector was performed using a genetic algorithm (GA). This method can be very expensive in complex multi-label datasets.

An approximation of ReliefF algorithm to multi-label data was presented in [8]. The authors decompose the multi-label problem into a set of pairwise multi-label 2-class problems. The algorithm excludes those examples that fall into *Hits* and *Misses* neighbors at the same time. The authors expose that the occurrence of these cases is very small, and therefore the exclusion of these instances will not affect the results significantly. However, this reasoning was done because the two specific datasets used in the experiment present this characteristic. In multi-label datasets a very high number of examples can fall into *Hit* and *Misses* neighbors at the same time, therefore excluding these examples can affect the results significantly.

In [9] other adaptation of ReliefF algorihtm to multi-label data was presented. It uses the standard ReliefF for single-label, where is measured the contribution of each feature according to each label. Afterwards, the average of the score of each feature across all labels is considered, and features with an averaged score greater than a threshold are selected. This approach use the Binary Relevance [10] approach to decompose the multi-label problem into several binary classification problems, therefore it does not consider label correlations.

## 3 The ReliefF-ML Algorithm

The biggest problem for the multi-label feature weighting process is that an instance is assigned to multiple labels simultaneously, therefore nearest *Hits* and *Misses* cannot be used in a strict sense as in classic ReliefF algorithm. Given a multi-label dataset, the prior probability of a label $l$ is computed as follows:

$$P_l = \frac{C_l + s}{N + 2s} \tag{1}$$

, where $C_l$ is the number of instances in the dataset that belong to label $l$ and $s$ is the smoothing parameter controlling the strength of uniform prior ($s = 1$ yields the Laplace smoothing).

Given the instances $i$ and $j$, the distance between the sets of labels of $i$ and $j$ is calculated by the Hamming Distance (see equation 2). The distance $d_{\mathcal{L}}$ represents a measure of how much differ the sets of labels of two instances.

$$d_{\mathcal{L}}(i,j) = \frac{\mid y_i \triangle y_j \mid}{Q} \tag{2}$$

ReliefF-ML uses the HEOM distance(Heterogeneous Euclidean Overlap Metric) [11](equation 3) to retrieve the $k$-nearest neighbors of an instance $i$ according to the feature space.

$$d_{\mathcal{F}}(i,j) = \sqrt{\sum_{\forall f \in \mathcal{F}} \delta(x_{if}, x_{jf})^2} \tag{3}$$

$$\delta(x_{if}, x_{jf}) = \begin{cases} 1 & discrete,\ x_{if} \neq x_{jf} \\ 0 & discrete,\ x_{if} = x_{jf} \\ \frac{|x_{if} - x_{jf}|}{max(f) - min(f)} & continuous \end{cases} \tag{4}$$

For each relevant and irrelevant label of an instance $i$ a group of $k$-nearest neighbors is defined. Therefore, the following groups of *Hits* $(H_i^l)$ and *Misses* $(M_i^l)$ respect to an instance $i$ are defined:
-$H_i^l$: $k$-nearest neighbors that have the relevant label $l$ of $i$ as relevant label
-$M_i^l$: $k$-nearest neighbors that have the irrelevant label $l$ of $i$ as relevant label

Based in the defined groups of *Hits* and *Misses* the following "probability" was defined, it is modelled with the distance between the sets of labels of two learning instances.

$$P_{G_i^l} = \frac{\sum_{\forall j \in G_i^l} d_{\mathcal{L}}(i,j)}{k} \tag{5}$$

, where:
-$P_{H_i^l}$: is the probability that two nearest instances that share the label $l$ as relevant, belong to different set of labels.
-$P_{M_i^l}$: is the probability that two nearest instances belong to different set of labels, where $i$ has the label $l$ as irrelevant and the $k$-nearest neighbors have the label $l$ as relevant.

In ReliefF-ML the dependence among labels is taken into account through the calculus of $P_{H_i^l}$ and $P_{M_i^l}$ for each relevant and irrelevant label respectively of a sampling instance. Each feature weight reflects its ability to distinguish class labels, thus a high weight indicates that there is differentiation in this attribute among instances with very different sets of labels and has similar values for instances with similar sets of labels otherwise. The weight updating of an attribute $f$ uses the equation (6).

$$W_f = W_f - \sum_{l \in y_i} \left( \frac{P_l}{\sum_{q \in y_i} P_q} \frac{1 - P_{H_i^l}}{1 + P_{H_i^l}} \sum_{j \in H_i^l} \frac{\delta(x_{if}, x_{jf})}{mk} \right) + \sum_{l \notin y_i} \left( \frac{P_l}{\sum_{q \notin y_i} P_q} P_{M_i^l} \sum_{j \in M_i^l} \frac{\delta(x_{if}, x_{jf})}{mk} \right) \tag{6}$$

The contributions of each relevant and irrelevant label are weighted by the factors $\frac{P_l}{\sum_{q \in y_i} P_q}$, $\frac{1 - P_{H_i^l}}{1 + P_{H_i^l}}$ and $\frac{P_l}{\sum_{q \notin y_i} P_q}$, $P_{M_i^l}$ respectively.

**Algorithm 1.** Pseudocode of ReliefF-ML algorithm

```
Input: $E$: learning multi-label instances, $m$: sampling parameter, $k$: number of nearest neighbors to retrieve
Output: weight vector $W$
for each $l \in \mathcal{L}$ do Calculate $P_l$ end for;
for each $f \in \mathcal{F}$ do Set $W_f = 0$ end for;
for $n = 1$ to $m$ do
    Pick randomly an instance $i$ from $E$
    for each relevant label $l \in y_i$ do
        Get $k$-nearest Hits $H_i^l$
        Calculate $P_{H_i^l}$
    end for
    for each irrelevant label $l \notin y_i$ do
        Get $k$-nearest Misses $M_i^l$
        Calculate $P_{M_i^l}$
    end for
    for each attribute $f \in \mathcal{F}$ do
        Calculate $W_f$ by expression (6)
    end for
end for
Scale the weights in the range [0..1]
```

ReliefF-ML picks randomly a predefined number of instances ($m$) from the $E$ set to estimate the feature weights. It uses the whole training set to retrieve the $k$ nearest neighbors of a selected instance. To fix the number of instances to be selected to estimate the feature weights the following rules were used:
1.***if*** $(|E| \leq 5000)$ ***then*** $(m{=}0.1 \times |E|)$
2.***if*** $(|E| > 5000 \, and \mid E \mid \leq 10000)$ ***then*** $(m{=}0.05 \times |E|)$
3.***if*** $(|E| > 10000)$ ***then*** $(m{=}0.01 \times |E|)$

## 4 Experimental Section

In [12] a lazy algorithm named ML-$k$NN was proposed, it uses the maximum a posteriori principle (MAP) in order to determine the set of labels of a query instance. DML-$k$NN [13] can be considered a generalization of the ML-$k$NN based approach where the dependencies among labels are considered. MLC-W$k$NN appears in [14], the author constructs a weighted $k$NN version for multi-label learning according to the Bayesian theorem.

To prove the effectiveness of the proposal, each lazy algorithm using the weights reached by ReliefF-ML were tested, and then the results were compared with the original methods. The modified algorithms were named ML-$k$NN-WF, DML-$k$NN-WF and MLC-W$k$NN-WF to differentiate them from the original methods. In the adapted lazy algorithms the function used originally to retrieve the $k$-nearest neighbors was replaced by the Weighted HEOM distance version, which takes into account the feature weights.

ReliefF-ML and the lazy algorithms were implemented on MULAN [15], that is a Java library which contains several methods for multi-label learning. For each possible combination of algorithms and datasets a stratified 10-fold cross validation strategy was used. For each fold in the training phase, ReliefF-ML finds the weight vector by picking randomly the sampling instances from the training set. The lazy learning algorithms use the weight vector in the distance functions to retrieve the $k$ nearest neighbors of an instance. The best value for the parameter $k$ used by ReliefF-ML and the lazy algorithms on each dataset was determined. As for comparison between the originals and adapted methods, the Wilcoxon signed ranks test was used as proposed in [16].

The algorithms were tested with 11 multi-label datasets from different domains. Selection was made in order to understand the behaviour of our approach in datasets with diverse characteristics. All datasets are available for download at the web page http://mlkd.csd.auth.gr/multilabel.html. In order to verify the effectiveness of the proposal, 4 evaluation measures that have been suggested for MLR problems in [10] were used. The Hamming Loss ($H_L$) reports how many times on average, the relevance of an example to a class label is incorrectly predicted. Accuracy ($A_{cc}$) returns the proportion of the predicted correct labels to the total number (predicted and actual) of labels for that instance, over all instances. One Error ($O_E$) measures how many times the top ranked predicted label is not in the set of true labels of the instances. Ranking Loss ($R_L$) evaluates the average proportion of label pairs that are incorrectly ordered for an instance.

## 5 Results and Dicussion

The performance of the ReliefF-ML was evaluated through comparisons of the algorithms ML-$k$NN, DML-$k$NN and MLC-W$k$NN, and their respective extensions ML-$k$NN-WF, DML-$k$NN-WF and MLC-W$k$NN-WF. In all cases the best results are highlighted in bold typeface in the tables. Tables 1 to 4 show the results of $H_L$, $A_{cc}$, $O_E$ and $R_L$ on the 3 selected algorithms.

The results shows that the adapted algorithms perform better than the original algorithms in almost all datasets with the 4 measures used in the experiment. Table 5 shows Wilcoxon's signed rank test; it summarizes the positive ($R^+$) and negative ($R^-$) ranks, ties and if the hypothesis is rejected (R) or not (NR) with a significance $\alpha$ equals to 0.01.

The evidences suggest that ML-$k$NN-WF, DML-$k$NN-WF and MLC-W$k$NN-WF are statistically better than the original algorithms in all the measures used. The results obtained show that the proposed approach is robust, it does well in datasets with different characteristics. Furthermore, the proposed method to multi-label feature weighting improves the performance of multi-label lazy learning algorithms.

**Table 1.** $H_L$ results

| **Dataset** | **ML-*k*NN** | | **DML-*k*NN** | | **MLC-W*k*NN** | |
|---|---|---|---|---|---|---|
| | **-** | **WF** | **-** | **WF** | **-** | **WF** |
| Emotions | 0.1963 | **0.1812** | 0.1965 | **0.1840** | 0.1884 | **0.1800** |
| Yeast | 0.1925 | **0.1915** | 0.1924 | **0.1910** | 0.1935 | **0.1915** |
| Scene | 0.0868 | **0.0865** | 0.0872 | **0.0859** | 0.0846 | **0.0840** |
| Cal500 | 0.1387 | **0.1382** | 0.1377 | **0.1373** | **0.1472** | **0.1472** |
| Genbase | 0.0043 | **0.0036** | 0.0046 | **0.0043** | 0.0012 | **0.0009** |
| Medical | 0.0151 | **0.0136** | 0.0157 | **0.0145** | 0.0146 | **0.0137** |
| Enron | 0.0526 | **0.0525** | 0.0520 | **0.0518** | 0.0558 | **0.0557** |
| TMC2007-500 | 0.0649 | **0.0620** | 0.0646 | **0.0620** | 0.0380 | **0.0366** |
| Mediamill | 0.0281 | **0.0279** | 0.0282 | **0.0280** | 0.0246 | **0.0245** |
| Corel5k | **0.0094** | **0.0094** | **0.0094** | **0.0094** | **0.0096** | **0.0096** |
| Corel16k | **0.0175** | **0.0175** | **0.0175** | **0.0175** | 0.0181 | **0.0180** |

**Table 2.** $A_{cc}$ results

| **ML-*k*NN** | | **DML-*k*NN** | | **MLC-W*k*NN** | |
|---|---|---|---|---|---|
| **-** | **WF** | **-** | **WF** | **-** | **WF** |
| 0.5344 | **0.5645** | 0.5352 | **0.5645** | 0.5518 | **0.5789** |
| **0.5201** | 0.5188 | **0.5196** | **0.5196** | 0.5268 | **0.5359** |
| 0.6665 | **0.6784** | 0.6665 | **0.6800** | **0.6879** | 0.6878 |
| 0.1954 | **0.1998** | 0.1914 | **0.1959** | 0.2216 | **0.2217** |
| 0.9499 | **0.9618** | 0.9453 | **0.9501** | 0.9894 | **0.9895** |
| 0.5828 | **0.6412** | 0.5288 | **0.5858** | 0.5815 | **0.6198** |
| 0.3032 | **0.3046** | 0.2978 | **0.3025** | 0.3162 | **0.3168** |
| 0.5296 | **0.5567** | 0.5285 | **0.5559** | 0.7264 | **0.7351** |
| 0.4727 | **0.4728** | **0.4700** | 0.4691 | 0.5517 | **0.5521** |
| 0.0148 | **0.0170** | 0.0026 | **0.0039** | 0.0344 | **0.0378** |
| 0.0076 | **0.0083** | 0.0043 | **0.0053** | 0.0339 | **0.0360** |

**Table 3.** $O_E$ results

| **Dataset** | **ML-*k*NN** | | **DML-*k*NN** | | **MLC-W*k*NN** | |
|---|---|---|---|---|---|---|
| | **-** | **WF** | **-** | **WF** | **-** | **WF** |
| Emotions | 0.2680 | **0.2296** | 0.2646 | **0.2300** | 0.2462 | **0.2385** |
| Yeast | 0.2272 | **0.2150** | 0.2263 | **0.2162** | 0.2325 | **0.2271** |
| Scene | **0.2244** | 0.2255 | **0.2252** | 0.2294 | 0.2285 | **0.2232** |
| Cal500 | 0.1168 | **0.1147** | **0.1147** | **0.1147** | 0.2264 | **0.1920** |
| Genbase | 0.0151 | **0.0084** | 0.0166 | **0.0085** | 0.0030 | **0.0022** |
| Medical | 0.2239 | **0.1975** | 0.2393 | **0.2042** | 0.2198 | **0.1949** |
| Enron | 0.3111 | **0.3100** | 0.3093 | **0.3012** | **0.3732** | 0.3782 |
| TMC2007-500 | 0.2313 | **0.2131** | 0.2315 | **0.2020** | 0.1412 | **0.1352** |
| Mediamill | 0.1554 | **0.1486** | 0.1536 | **0.1521** | 0.1321 | **0.1312** |
| Corel5k | 0.7288 | **0.7170** | 0.7314 | **0.7248** | 0.7824 | **0.7640** |
| Corel16k | 0.7396 | **0.7320** | 0.7401 | **0.7301** | 0.7760 | **0.7660** |

**Table 4.** $R_L$ results

| **ML-*k*NN** | | **DML-*k*NN** | | **MLC-W*k*NN** | |
|---|---|---|---|---|---|
| **-** | **WF** | **-** | **WF** | **-** | **WF** |
| 0.1596 | **0.1500** | 0.1558 | **0.1484** | 0.1641 | **0.1565** |
| 0.1658 | **0.1630** | 0.1646 | **0.1631** | 0.1739 | **0.1726** |
| **0.0801** | **0.0801** | 0.0777 | **0.0770** | 0.0834 | **0.0819** |
| 0.1812 | **0.1807** | 0.1992 | **0.1787** | 0.2482 | **0.2473** |
| 0.0071 | **0.0063** | 0.0070 | **0.0059** | 0.0038 | **0.0037** |
| 0.0363 | **0.0341** | 0.0353 | **0.0322** | 0.0438 | **0.0427** |
| **0.0898** | **0.0898** | 0.0894 | **0.0892** | **0.1857** | **0.1857** |
| 0.0584 | **0.0520** | 0.0563 | **0.0498** | 0.0510 | **0.0490** |
| 0.0369 | **0.0363** | **0.0360** | **0.0360** | **0.0608** | 0.0613 |
| 0.1300 | **0.1292** | 0.1306 | **0.1302** | 0.4731 | **0.4656** |
| 0.1641 | **0.1635** | 0.1647 | **0.1642** | 0.3086 | **0.3060** |

**Table 5.** Wilcoxon's signed rank test

| **Measures** | $R^+$ | $R^-$ | $Ties$ | $p-value$ | $Hypothesis$ |
|---|---|---|---|---|---|
| ML-*k*NN-FW vs ML-*k*NN | | | | | |
| $H_L$ | 0 | 9 | 2 | 0.008 | R |
| $A_{cc}$ | 1 | 10 | 0 | 0.008 | R |
| $O_E$ | 1 | 10 | 0 | 0.005 | R |
| $R_L$ | 0 | 9 | 2 | 0.008 | R |
| DML-*k*NN-FW vs DML-*k*NN | | | | | |
| $H_L$ | 0 | 9 | 2 | 0.008 | R |
| $A_{cc}$ | 1 | 9 | 1 | 0.007 | R |
| $O_E$ | 1 | 9 | 1 | 0.009 | R |
| $R_L$ | 0 | 10 | 1 | 0.005 | R |
| MLC-W*k*NN-FW vs MLC-W*k*NN | | | | | |
| $H_L$ | 0 | 9 | 2 | 0.007 | R |
| $A_{cc}$ | 1 | 10 | 0 | 0.006 | R |
| $O_E$ | 1 | 10 | 0 | 0.008 | R |
| $R_L$ | 1 | 9 | 1 | 0.009 | R |

## 6 Conclusions

The attention given to the study of feature weighting methods in multi-label learning has been negligible. In this paper, a filter feature weighting method called ReliefF-ML to deal with multi-label problems was proposed. The proposed method has significant advantages; it is a preprocessing step that is completely independent of the choice of particular multi-label algorithm. Also, it uses the given representation of the original datasets (handles multi-label data directly), it learns a single set of weights that are employed globally over the entire instance space, it takes into account the label correlations in the estimation of feature weights and does not employ domain specific knowledge to set feature weights. The algorithm ReliefF-ML is a generalization of the classic ReliefF algorithm.

The experiments aimed to measure the performance of multi-label lazy algorithms in conjunction with the proposed method for feature weighting. Results from the statistical tests show that the proposed method has significant advantages, which indicate that the approach is robust for MLR problems.

## References

1. McCallum, A.: Multi-label text classification with a mixture model trained by EM. In: Working Notes of the AAAI-99 Workshop on Text Learning (1999)
2. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of the International Symposium on Music Information Retrieval (2003)
3. Yang, S., Kim, S., Ro, Y.: Semantic home photo categorization. IEEE Transactions on Circuits and Systems for Video Technology 17, 324–335 (2007)
4. Kononenko, I.: Estimating attributes: Analysis and extensions of RELIEF. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)
5. Wettschereck, D., Aha, D.W., Mohri, T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review 11, 273–314 (1997)
6. Brinker, K., Furnkranz, J., Hullermeier, E.: A unified model for multilabel classification and ranking. In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 2006), pp. 489–493 (2006)
7. Reyes, O., Morell, C., Ventura, S.: Learning similarity metric to improve the performance of lazy multi-label ranking algorithms. In: IEEE Proceedings of the 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 246–251 (2012)
8. Kong, D., Ding, C., Huang, H., Zhao, H.: Multi-label ReliefF and F-statistic Feature Selections for Image Annotation. In: Proceedings of Computer Vision and Pattern Recognition (CVPR), pp. 2352–2359 (2012)
9. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: Filter approach feature selection methods to support multi-label learning based on reliefF and information gain. In: Barros, L.N., Finger, M., Pozo, A.T., Gimenénez-Lugo, G.A., Castilho, M. (eds.) SBIA 2012. LNCS, vol. 7589, pp. 72–81. Springer, Heidelberg (2012)
10. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In: Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 667–686. Springer (2010)
11. Wilson, D., Martinez, T.R.: Improved heterogeneous distance functions. Journal of Artificial Intelligence Research (JAIR) 6, 1–34 (1997)
12. Zhang, M.L., Zhou, Z.H.: ML-$k$NN: A lazy learning approach to multi-label learning. Pattern Recognition 40(7), 2038–2048 (2007)
13. Younes, Z., Abdallah, F., Denceux, T.: Multi-label classification algorithm derived from $k$-nearest neighbor rule with label dependencies. In: Proceedings of the 16th Eropean Signal Processing Conference, Lausanne, Switzerland (2008)
14. Xu, J.: Multi-label weighted $k$-nearest neighbor classifier with adaptive weight estimation. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part II. LNCS, vol. 7063, pp. 79–88. Springer, Heidelberg (2011)
15. Tsoumakas, G., Spyromitros-Xioufi, E., Vilcek, J., Vlahavas, I.: MULAN: A java library for multi-label learning. Journal of Machine Learning Research 12, 2411–2414 (2011)
16. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)

# Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names

Maite Oronoz[1], Arantza Casillas[2], Koldo Gojenola[1], and Alicia Perez[1]

[1] Departamento de Lenguajes y Sistemas Informáticos. IXA taldea. UPV-EHU
[2] Departamento de Electricidad y Electrónica. IXA taldea. UPV-EHU
{maite.oronoz,arantza.casillas,koldo.gojenola,alicia.perez}@ehu.es

**Abstract.** This paper presents an annotation tool that detects entities in the biomedical domain. By enriching the lexica of the Freeling analyzer with bio-medical terms extracted from dictionaries and ontologies as SNOMED CT, the system is able to automatically detect medical terms in texts. An evaluation has been performed against a manually tagged corpus focusing on entities referring to pharmaceutical drug-names, substances and diseases. The obtained results show that a good annotation tool would help to leverage subsequent processes as data mining or pattern recognition tasks in the biomedical domain.

**Index Terms:** development of linguistic tools, annotation, medical domain.

## 1 Introduction

Syntactic and semantic annotation has been used in many applications such as data mining and pattern recognition. There are a variety of supervised and semi-supervised training algorithms that require to be boosted from annotated data sets. The aim of this paper is to automatically annotate different types of entities in the biomedical domain.

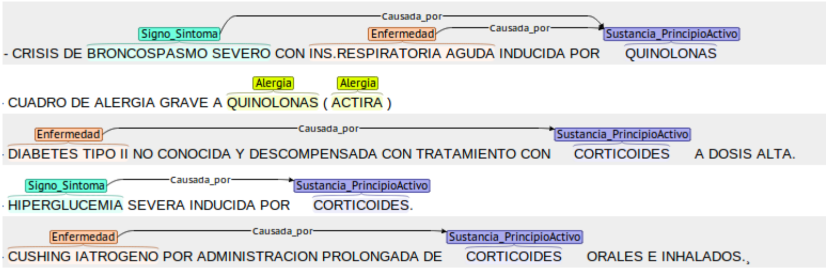
Over the last years Spanish health care services are storing most of the information concerning patients in electronic medical records. These clinical texts constitute a rich source of information about diseases, allergies, and any information that the sanitary personnel is interested in. Access to this information is of great interest and value for clinical research. Many current methods for accessing information are based on statistical and machine learning methods, that need annotated data. However, the annotation process is time-consuming and expensive to be performed manually. Biomedicine is an area where the corpora have a confidential nature, hence, open resources are scarce and when comparing it to other fields it does not seem an eligible task for exploiting publicly available resources such as the semantic web, althouh there are some publicly available resources such as parallel corpora in various languages [1,2]. Besides, the annotators' expertise is crucial, and thus, it is not an option for crowd sourcing or

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 536–543, 2013.

social annotation as it was done in other tasks like language model adaptation [3]. Making this an automatic process would allow to save work and money.

The Pharmaceutical Service of the Galdakao's Hospital performs the task of manually detecting Adverse Drug Reactions (ADRs). The aim of the tool presented in this paper is to automatically annotate medical texts with brand-name drugs, disease names and substance names, opening the way for the future automatic detection of ADRs.

Figure 1 shows a fragment of a clinical note with annotations for diseases, substances and drugs as well as allergies and adverse drug effects, obtained by means of Brat [4], a tool for text annotation. This tool allows not only to highlight such events but also to detail cause-effect relations. Note that while the figure shows a manual annotation provided by medical experts, the aim of this work is to produce the annotation automatically. As a result, reading a clinical note (or conversely, supervising a dictated note) would be easier, since this tool would allow to draw the attention to specific items.



- CRISIS DE BRONCOSPASMO SEVERO CON INS.RESPIRATORIA AGUDA INDUCIDA POR QUINOLONAS

· CUADRO DE ALERGIA GRAVE A QUINOLONAS ( ACTIRA )

· DIABETES TIPO II NO CONOCIDA Y DESCOMPENSADA CON TRATAMIENTO CON CORTICOIDES A DOSIS ALTA.

· HIPERGLUCEMIA SEVERA INDUCIDA POR CORTICOIDES.

· CUSHING IATROGENO POR ADMINISTRACION PROLONGADA DE CORTICOIDES ORALES E INHALADOS.

**Fig. 1.** Medical record manually annotated with the Brat toolkit

The core element of the proposed automatic annotation toolkit lies in creating a syntactic and semantic analyzer for Spanish in the specific domain of biomedicine. In this paper, we will focus on the description of the adaptation of the linguistic analyzer Freeling [5] to the domain of medicine. The annotations provided by the presented domain-adapted analyzer will be evaluated with respect to annotations provided by human experts. The benefits of having an automatic analyzer are twofold: (1) automatic annotation is much faster and cheaper (2) the annotated data will serve for developing advanced information extraction and data mining systems.

There are only a few publically accesible analyzers adapted to the clinical domain in the Spanish language. For English, the GENIA tagger [6] is specifically tuned for biomedical texts. Patrick et al. [7] introduce a new method to automatically identify medical concepts from the Systematized Nomenclature of Medicine-Clinical Texts (SNOMED CT) in English free text. MetaMap Transfer (MMTx) [8] is a program to map biomedical text to the UMLS[1] Metathesaurus or, equivalently, to discover concepts from the Metathesaurus in texts. In [9] the

[1] http://www.nlm.nih.gov/research/umls/

authors present a first simple approach to the Spanish MetaMap, using Google Translator to obtain an English version of the text and then applying English MMTx to extract the concepts. In [10] a system for the automated identification of biomedical concepts in Spanish-language clinical notes is presented.

The rest of the paper is arranged as follows: Section 2 delves into the adaptation and enrichment of the linguistic analyzer. Section 3 is devoted to the experimental evaluation of the tool against a set of manually annotated texts. Finally, conclusions and future work are given in section 4.

## 2 Automatic Analysis of Electronic Medical Records

For the initial processing of medical records, we have made use of a basic Natural Language Processing toolkit, *Freeling*[2], together with several available medical ontologies and dictionaries. Freeling is an open-source multilingual language processing library providing a wide range of language analyzers for several languages [5]. In this work, we used the tools for Spanish morphological analysis provided by Freeling. The linguistic resources (lexica, grammars, . . . ) in Freeling can be modified, so we took advantage of this flexibility by extending the linguistic data files with large-scale resources containing medical information.

As it is a standard approach in Natural Language Processing, where there is a distinction between morphology and syntax on one side and semantics on the other, we will distinguish two levels of processing. In our case, during morphosyntactic processing, our system will only categorize word-forms using their basic part-of-speech (POS) categories (explained in section 2.1), while the semantic distinctions will be dealt with in a second stage (see section 2.2). Following this approach, if the term that we want to insert already existed in Freeling's standard Spanish dictionary, e.g. *bar* as common noun (bar or pub), the entries with medical meanings will not be added to the lexicon, e.g. *bar* or *bacilo acidorresistente* (acid-fast rod) because this term also corresponds to a common noun. The medical meanings are added in a later semantic tagging phase. This solution helps to avoid an explosion of ambiguity in the morphosyntactic analysis and enables a clear separation between morphosyntax and semantics.

### 2.1 Enriching Dictionaries in Freeling

In order to extend the standard Freeling analyzer in Spanish to the medical domain, we enriched two dictionaries: a basic dictionary of terms consisting of a unique word, and a multiword-term dictionary. The former should be enriched with terms such as *enteroplastia* (repair of intestine), and the latter with composed terms as, for example, *canal vertebral lumbar* (lumbar spinal canal).

As we previously explained, to keep the distinction between morphosyntactic and semantic ambiguity in the lexica is essential for us. We decided to add a term to the files with POS information in Freeling only if it did not exist before.

[2] http://nlp.lsi.upc.edu/freeling/

For example, *xilosuria* (xylosuria) or *Zofenil* (pharmaceutical product) will be new entries applying this principle. The first row in Table 1 shows the number of entries of the standard lexica for Spanish within the Freeling 2.2. standard package. The medical resources used to enhance them are the following:

**Medical Abbreviations.** Yetano and Alberola [11] gathered the abbreviations used in some hospitals to develop a dictionary of medical abbreviations and acronyms for Spanish. After a manual examination, we obtained a list of 3,196 entries. Some of them were ambiguous, e.g. *ADR* meaning *adrenalina* (adrenalin), or *adriamicina* (adriamycin), while others were not, e.g. *HTA* (*Hipertensión arterial* for “high blood pressure”). Table 1 shows the number of abbreviations already contained in the standard lexica (first row in Table 1), and the number of new abbreviations. The majority of the abbreviations are new entries in the Freeling lexica because they correspond specifically to the medical language (e.g. *vvz* extended *virus varicela zoster*). All the abbreviated chemical elements (e.g. *as*, *bi*), measure units (e.g. *kg*, *cm* ...) were already in the lexica.

**SNOMED CT Terms.** SNOMED CT is a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting. SNOMED CT is based on concepts, that is, units of thought or clinical ideas, coded by means of alphanumeric identifiers (e.g. 106190000 refers to *allergy*). Concept-descriptions are classified into *Fully Specified Name* in which the hierarchy the term belongs to is indicated (body part, procedure...), *Preferred Terms* and *Synonyms*. We have added the preferred terms and the synonyms of the 31th of October 2011 release to the lexica in Freeling.

The Unified Medical Language System (UMLS), is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. SNOMED CT is part of the Metathesaurus knowledge source in UMLS. We tagged the terms in Spanish with their corresponding SNOMED CT identifiers but also with their UMLS identifiers (see figure 2). In this way we will have the option of accesing the other ontologies in UMLS and of getting additional medical information.

Table 1 shows that 94.1% of the terms from SNOMED CT have more than one word and 94% of them were new in the multiword-term file. This fact gives an idea of the complexity of the terms used in SNOMED CT. In proportion, the number of single word terms already in the dictionaries, 9,302 out of 23,399, is relatively high, compared to the number of locutions or multiword terms.

**Bot PLUS.** Bot PLUS is a database of sanitary knowledge distributed by the General Council of Spanish Pharmacologists[3]. Bot PLUS stores the names of all the medicines that are commercialized in Spain. The knowledge stored in the Bot PLUS database makes up for the lack of this kind of information in SNOMED CT. For the work presented in this paper, we have obtained the following lists:

[3] `http://www.portalfarma.com`

i) brand names or pharmaceutical drug names and ii) substances. Table 1 shows the lexical entries incorporated to Freeling, having Bot PLUS as a basis.

Regarding the insertion of medicine brand-names in the lexica, it is worth remarking that from 9,984 names, 9,902 entries are new in the lexica and only 82 existed already (e.g. *rizan* from the verb "to curl"). In the case of substance-names with a unique token, there are more terms already in the dictionaries (1,590) than those entered as new ones (1,406) because they have their place in SNOMED CT, and they were already in the lexica.

**Table 1.** Number of entries in the lexica of Freeling and added resources

| | | **Unique word terms** | **Multiword terms** | **Total** |
|---|---|---|---|---|
| **FreeLing** | Standard | 556,212 | 1,480 | 557,692 |
| **Abbreviations** | In dictionary | 369 | 4 | 373 |
| | New | 2,654 | 169 | 2,823 |
| | **Total** | 3,023 | 173 | 3,196 |
| **SNOMED CT** | In dictionary | 9,302 | 125 | 9,427 |
| | New | 23,399 | 521,973 | 545,372 |
| | **Total** | 32,701 | 522,098 | 554,799 |
| **Bot PLUS** | **Medicine brand-names** | | | |
| | In dictionary | 61 | 21 | 82 |
| | New | 3,746 | 6,156 | 9,902 |
| | **Subtotal** | 3,807 | 6,177 | 9,984 |
| | **Substances** | | | |
| | In dictionary | 1,590 | 158 | 1,748 |
| | New | 1,406 | 1,072 | 2,478 |
| | **Subtotal** | 2,996 | 1,230 | 4,226 |
| | **Total** | 6,803 | 7,380 | 14,210 |
| **ICD-9** | In dictionary | 530 | 1,029 | 1,559 |
| | New | 268 | 17,950 | 18,218 |
| | **Total** | 798 | 18,979 | 19,777 |

**ICD-9.** The International Statistical Classification of Diseases is a medical classification list compiled by the World Health Organization (WHO). All the medical records from the Basque Health System should be tagged with a code indicating the medical diagnosis of the patient, following the 9th version of this classification (ICD-9). Table 1 shows the data about the integration of these terms in Freeling's lexica and the complexity of the terms in ICD-9.

The four lexica have been integrated in Freeling in their order of appearance in the paper, that is, abbreviations first, and then SNOMED CT, Bot PLUS and ICD-9. We decided to give priority to SNOMED CT against Bot PLUS and ICD-9, because it is a well structured and extensive clinical terminology. The expansion of the abbreviations first is essential if we want to add meanings, e.g. from SNOMED CT, to the expanded lemmas.

### 2.2 Semantic Postprocess

With the augmented lexica, Freeling performs tokenization, morphological analysis, POS tagging, lemmatization, shallow parsing and dependency parsing. The medical records are analyzed with linguistic information at all these levels but at the present work we will make use of information about terms, which gives access to all information levels except syntactic dependencies. All the entries described in section 2 have been inserted as nouns in the lexica, but also indicating the source of information of each entry.

In case of an ambiguity of meanings, that ambiguity would correspond to the semantic level. Being this the case, we insert this medical information and, in consequence, ambiguity in the analysis. The example in figure 2 shows that the word *Estreptomicina* was already in Freeling as a common noun feminine singular (tag NCFS000). For medical information extraction tasks, it is important to know that this is a *substance* or *product*, so we will insert this information as an *External Reference* (*extRef*). In the *extRef* we include information about the *resource* (Snomed CT in Spanish version of the date 31 October 2011), the SNOMED CT *Concept Identifier* in the *reference* attribute and the *reftype*, in our case corresponding to the semantic tag of the term in SNOMED CT (*product* and *substance*). For future works we aim to access the entire UMLS, this is why we have also inserted the *UMLS's Concept Unique Identifier* in the analysis.

Overall, the enhancement process of the lexical resources adds 47,132 standard entries and 554,807 locutions, taking an outstanding step ahead in text processing of the biomedical domain.

```
<term lemma=”estreptomicina” pos=”N.NCFS000” tid=”t56” >
  <extRefs>
    <extRef resource=”SCT_20111031” reference=”40877002” reftype=”producto”>
      <extRef resource=”UMLS-2010AB” reference=”C0038425”/>
    </extRef>
    <extRef resource=”SCT_20111031” reference=”387223008” reftype=”sustancia”>
      <extRef resource=”UMLS-2010AB” reference=”C0038425”/>
    </extRef>
  </extRefs>
</term>
```

**Fig. 2.** Analysis with augmented information

## 3 Evaluation

Although our adapted linguistic analyzer is able to detect terms from the 19 content hierarchies of SNOMED CT (i.e. organisms, procedures,...), one of the first uses of the analyzer will be to detect adverse drug events. This is the reason for focusing our first evaluation in the detection of drug-names, diseases and substances. We distinguish between brand-name drugs (e.g. *Nolotil*) and substances that could be active ingredients (e.g. *Metamizol*) or any substance that could create an adverse drug reaction (e.g. *polen* meaning pollen).

We did not found any publicly avalaible corpus composed of electronic medical records in Spanish, so after several meetings with the legal advice services of the University and the Hospital, and after signing the corresponding confidentiality agreement, we obtained a corpus of patient records. Having a "private" corpus, our results are not comparable to others, as in other related works [10].

A corpus of 100 medical records was collected from the outpatient consultations of the Galdakao Hospital and it was manually tagged by doctors and pharmacologists. The corpus is composed of 51,061 words and the experts have manually tagged 690 drug names, 891 diseases and 735 substances. The performance of the analyzer was assessed using the manually tagged corpus. These data samples were shuffled and randomly split into three disjoint sets for training (60 documents), development (20 documents) and test purposes (20 documents).

The system is assessed by means of the F-Measure that compares the human annotation with the output of the analyzer by combining precision and recall. In order to set out if two elements are equal, an approximate correctness criteria was applied: two elements are considered to be equivalent if an element given by the system is entirely contained within an extension of a manually tagged element by six positions both to the left and to the right. This follows the standard approach of allowing an approximate boundary matching, as in the BioNLP Shared Task [12]. Table 2 shows the number of drugs, substances and diseases in the test set, also presenting the number of True Positives (TP), False Negatives (FN) and False Positives (FP) returned by the system for each category of elements. Precision (PR), recall (RE) and F-Measure (F-M) are calculated for each type of element. The results are encouraging, with an F-Measure of 0.90, and imply that the designed analyzer can automatically generate reliable annotated corpus with morphosyntatic and medical-concept tags.

**Table 2.** Results achieved by the automatic tagger on the test set

| | Manual | TP | FN | FP | PR | RE | F-M |
|---|---|---|---|---|---|---|---|
| Diseases | 211 | 354 | 88 | 12 | 0.97 | 0.80 | 0.88 |
| Drugs | 180 | 175 | 8 | 0 | 1.00 | 0.96 | 0.98 |
| Substances | 184 | 357 | 27 | 65 | 0.84 | 0.92 | 0.88 |
| **Total** | 575 | 886 | 123 | 77 | 0.92 | 0.88 | 0.90 |

## 4 Conclusions

The goal of this work was to create an analyzer for clinical texts in Spanish that identifies medical entities. To attain this goal we have added medical information to a standard linguistic analyzer for Spanish. The incorporated information was extracted from different sources such as ontologies, a medical abbreviation dictionary and a pharmaceutical drug element database. The system is robust enough to deal with electronic medical records in which abbreviations and errors are very common. We think that in the same way, it is able to analyze other types of texts within the medical domain (journal papers, books. . . ).

The contributions of this work are threefold: 1) the enhancement of standard Spanish dictionaries for the biomedical domain in the FreeLing toolkit; 2) the development of a system based on FreeLing to automatically annotate medical records providing an F-Measure of 0.90; 3) the compilation of a corpus of medical documents tagged with medical concepts in Spanish.

In the near future, we aim to improve the system by adding, as external references, the missing information about abbreviations, drug names from Bot

PLUS and diseases from ICD. This will produce an increase in the semantic ambiguity of the terms. For that reason, we want to use UKB [13], a tool for graph-based word sense disambiguation to select the adequate medical sense.

**Acknowledgements.** We would like to thank the Pharmacy Service of the Hospital Galdakao-Usansolo and the Pharmacovigilance Service of the Basque Government. This work was supported by the Department of Industry of the Basque Government (IT344-10, S-PE12UN114), the University of the Basque Country (GIU09/19), the Spanish Ministry of Science and Innovation (MICINN, TIN2010-20218).

## References

1. Jimeno-Yepes, A., Prieur-Gaston, É., Névéol, A.: Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics 14, 146 (2013)
2. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Proc. Language Resources and Evaluation, LREC (2012)
3. Wu, Y., Abe, K., Dixon, P.R., Hori, C., Kashioka, H.: Leveraging Social Annotation for Topic Language Model Adaptation. In: Proc. International Speech Communication Association (INTERSPEECH) (2012)
4. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: A web-based tool for nlp-assisted text annotation. In: Proc. EACL (2012)
5. Padró, L., Reese, S., Agirre, E., Soroa, A.: Semantic Services in Freeling 2.1: WordNet and UKB. In: Global Wordnet Conference, Mumbai, India (2010)
6. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: 10th Panhellenic Conference on Informatics (2005)
7. Patrick, J., Wang, Y., Budd, P.: An Automated System for Conversion of Clinical Notes into SNOMED Clinical Terminology. In: Proc. Australasian symposium on ACSW frontiers, ACSW 2007, vol. 68, pp. 219–226 (2007)
8. Aronson, A.: Effective Mapping of Biomedical Text to the UMLS Metathesaurus: the MetaMap program. In: Proc. of AMIAS, pp. 17–21 (2001)
9. Carrero, F.M., Cortizo, J.C., Gómez, J.M., de Buenaga, M.: In the Development of a Spanish Metamap. In: Proc. of the 17th ACM Conference on Information and Knowledge Management, pp. 1465–1466 (2008)
10. Castro, E., Iglesias, A., Martínez, P., Castaño, L.: Automatic Identification of Biomedical Concepts in Spanish-Language Unstructured Clinical Texts. In: Proc. of the 1st ACM International Health Informatics Symposium. IHI 2010, pp. 751–757 (2010)
11. Yetano, J., Alberola, V.: Diccionario de Siglas Médicas y Otras Abreviaturas, Epónimos y Términos Médicos Relacionados con la Codificación de las Altas Hospitalarias. Ministerio de Sanidad y Consumo (2003)
12. Kim, J.D., Pysalo, S., Ohta, T., Bossy, R., Nguyen, N., Tsujii, J.: Overview of BioNLP Shared Task 2011. In: Proc. of BioNLP Shared Task 2011. ACL (2011)
13. Agirre, E., Soroa, A., Stevenson, M.: Graph-based word sense disambiguation of biomedical documents. Bioinformatics 26, 2889–2896 (2010)

# High Throughput Signature Based Platform for Network Intrusion Detection

José Manuel Bande Serrano[1], José Hernández Palancar[1], and René Cumplido[2]

[1] Advanced Technologies Application Center, 7ma A ♯ 21406, e/ 214 y 216, Siboney,
Playa, CP: 12200, Havana, Cuba
{jbande,jpalancar}@cenatav.co.cu
http://www.cenatav.co.cu

[2] Instituto Nacional de Astrofísica Optica y Electrónica, Luis E. Erro 1, Sta. Ma.
Tonanzintla, Puebla, 72840, México
rcumplido@inaoep.mx

**Abstract.** In this work we propose the intensive use of embedded memory blocks and logic blocks of the FPGA device for signature matching. In our approach we arrange signatures in memory arrays (MA) of embedded memory blocks, so that every signature is matched in one clock cycle. The matching logic is shared among all the signatures in one MA. In addition, we propose a character recodification method that allows memory bits savings, leading to a low byte/character cost. For fast memory addressing we employ the unique substring detection, in doing so we process four bytes per clock cycle while hardware replication is significantly reduced.

**Keywords:** NIDS, string matching, content scanning, FPGA, unique substrings.

## 1 Introduction

Network Intrusion Detection Systems (NIDS) are designated to protect networks and services against attacks executed by insiders or outsiders. There are three kinds of NIDS: Signature-based, Misuse-based and Anomaly-based [1]. In Signature-based data flow is scrutinized in the search of attacks with signatures known beforehand. In Misuse-based signatures are automatically discovered through Supervised Learning methods. Finally, Anomaly-based, assumes that intrusions are, by nature, deviations from normal behavior. Of the three, only Anomaly-based intrusion detection is capable of detecting unknown attacks [1].

Although much progress has been made in Anomaly and Misuse-based detection, a fast and efficient signatures detection is still needed. The reason is that the types of NIDs exposed before, represent the natural mechanics of learning. This is, the unknown knowledge is perceived, then is characterized, and finally it becomes part of the current knowledge. In this integration, signature-based detection becomes into the first line of defense, because it makes, or helps to make decisions based on the current knowledge, as fast as possible [2].

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 544–551, 2013.

Since the speed of data streams will continue to grow for the next years, and fast responses to attacks are necessary in high security environments, signature matching is an active field of research. This demanding environment requires hardware solutions. In that direction, we propose a memory and logic based architecture where signatures are compressed and stored in memory arrays. Our matching logic allows the comparison of one signature per clock cycle. The entire signature set is partitioned. From each one of resulting subsets, only one signature is selected at each clock to be matched with the data flow. This is carried out by a predetection step. In order to store the entire signature set in memory we propose character recodification. In doing so the resulting architecture presents a better balance in the use of memory and logic, regarding other multi-character architectures.

The rest of the document is as follows. In the section two, we analyze the related works, paying special attention to those multi-character architectures. In the section three, the employed partitioning method is explained. In the section four, the architecture is presented. Section five is dedicated to experiments and comparisons with other works. Finally, conclusions are presented.

## 2 Related Works

Baker and Prassana in [3] proposed partitioning scheme that allows resource sharing in a logic-based architecture. Hardware implementation of the well-known string matching Shift-or algorithm is proposed in [4]. In [5], the well-known Aho-Corassick (AC) automaton structure is shared among several string matching modules in a time multiplexed access scheme. In [6], AC states with similar transitions are merged. The authors propose a mechanism to efficiently rectify the functional errors caused by the states merging. In doing so a reduction of 24% in the cost is achieved. Guinde and Ziavras [7] proposed a compression method for the string set where the required memory for storing the set is significantly reduced. In [11], they propose MIN-MAX algorithm for solving ambiguity and overlapped matching for Character Classes with Constraint Repetitions based Regular Expressions. A previous work was presented in [9] where the use of unique subsequences is introduced for reducing the hardware replication. In [8], a binary search tree state-of-the-art architecture is proposed, achieving the lowest memory cost per character but with a limited throughput.

## 3 Partitioning Methods

The present work is based on the partitioning methodology presented in [9] and then extended in [10]. Firstly, the initial signature set is partitioned into several sets denoted as u-sets. The partitioning criterion is that, every signature in a u-set must contain, at least, one unique substring. This is, a substring that is not contained in any other signature of that set. This substring is named unique substring, u-substring for short.

This partitioning allows that in a u-set, every signature can be mapped one-to-one with its corresponding u-substring. Therefore to find a u-substring in some data flow location, implies that its container signature, and no other, likely exist in that section of data stream, so there is no need to match any other signature. The likely-present signature is called candidate signature (CSig). This is fetched from memory every time its corresponding u-substring is detected. The section of data stream where the signature is expected to reside is named region of interest (RoI). A match occurs when CSig match character by character with the data stream, in a RoI. When extended to multi-character, it may happen that several u-substrings match in the same clock period. In order to avoid malfunctioning, a second partitioning is applied [10]. This is called security threshold partitioning. The output of this matching module is the signature ID, which is the signature address in the SMA, and a match enable output, signalling when a match occurs.

The first step in the construction of our architecture is to partition the signature set according to [9] and [10]. By using these methods, we guarantee only one possible signature match per clock cycle, for a u-set. The main contributions of this paper regarding ours previous works consist in: a) the use memory instead of logic, for storing signatures; b) the proposition of a character reconfiguration method which reduce, on average, the amount of memory bits required per character; and c) a different efficient masking solution, allowing to match non-uniform length signatures, using uniform hardware logic.

## 4 Architecture

Our method starts by representing every u-set as a matrix, with one character per cell and one signature per row. The signatures in this matrix are displaced, so that all u-substring first characters fall in the same column. In figure 2(a) there are three matrices. In the signature matrix, the top one, each row contains a signature where u-substrings are *"bb", "lb"* and *"tl"*, respectively. The column where all substrings begin is called aligning-column, because it works as pivot for the Aligning. The dashed line in the figure2(a) marks the boundary between the head, i. e., the prefix up to Aligning Column, and the tail, which is the rest of the signature.

Since a signature matrix column, may contain repeated characters, the number of distinct characters in a column is lower, or equal, to the alphabet size. We build a second matrix called character matrix, the middle one in figure 2(a). This matrix collects only distinct characters in columns from the signature matrix. In real signature sets, the size of the character matrix columns tends to be lower than 256. This makes possible to reencode the characters in order to save memory. Let $p$ be the number of characters in a character matrix column, the number of bits required to encode the signature characters is $log_2(p)$. This is what we have called character-recodification.

The bottom array of the figure 2(a) shows the bits that are required to store per column. Note that some columns have 0 bits, meaning that, we do not need to save this characters in memory. In these columns, the selected character is

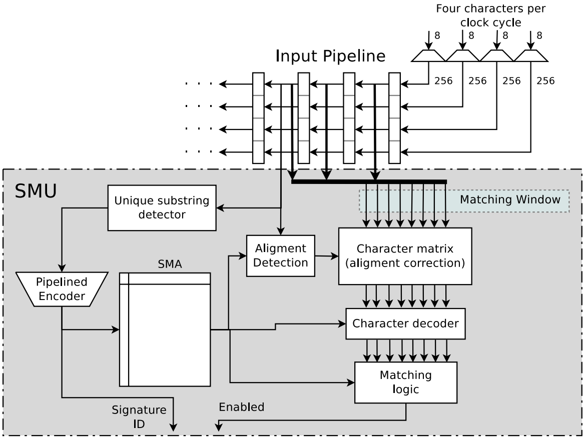
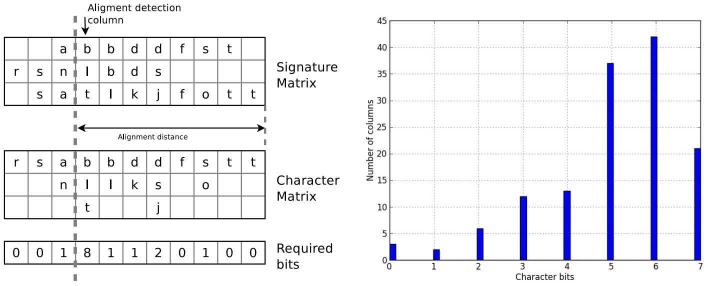


**Fig. 1.** General architecture view

always the same for any signature, so these characters lines can be hard-wired in the matching module, consuming no memory resources.

Figure 2(b) shows a histogram where bars represent the columns count with a specific width in bits. For a signature matrix with 1024 signatures extracted from the Snort rule database [12], there is a reduction of at least one bit in relation to the original character size. Note that 43 columns are six-bit wide, saving $43 * 2 = 86$ bits of memory. The matrix has 136 columns, without re-encoding $136 * 8 = 1088$ bits per row are required, while with re-encoding, this is reduced to 688, leading to a reduction of 36%. In terms of memory blocks, each of these contains 36 bits per entry, so $\lceil 1088/36 \rceil = 31$ are originally required, while with our method this is reduced to $\lceil 688/36 \rceil = 20$. This implies that the length of the memory entry can be shortened, making feasible the concatenation of embedded memory blocks, storing one signature per entry.

Each signature matrix is stored in an array of memories, called Signature Memory Array (SMA), occupying one entry per signature. The amount of entries of a MA is restricted to 1024. Therefore, the same u-set can require several SMAs. A Signature Matching Unit, SMU, is the basic component of our architecture, and its objective is to match the signatures contained in one MA. In figure 1, all but the input pipeline and the input character decoders, are components of the SMU. It performs five main tasks. First, match u-substrings from the data flow (Carried out by Unique Substring Detector). Second, fetch the Csig from the SMA corresponding to a matched u-substrings (Carried out by Unique Substring Detector and SMA). Third, align the Csig with the RoI (Carried out by SMA, Alignment Detection Component and Character Matrix Component). Four, execute the matching between the RoI and de Csig, comparing character by character (Carried out by Character Decoder component, Matching logic Component). Five, provide the match result, and the unique identifier of the recognized signature (Carried out by Matching Logic Component).



(a) (b)

**Fig. 2.** (a)Signature Matrix example (b) Counting of columns at every width

As shown in figure 1, the architecture processes four characters per clock cycle. These characters are decoded into bit lines, and passed through a pipeline of register, denoted as i-pipeline. Every step in the i-pipeline, is divided into four sections, corresponding to four characters, resulting in a total of $256 * 4$ bit lines per step. The i-pipeline can be seen as a serial to parallel buffer, where the parallel outputs feed the SMUs inputs. The input of the SMU is called Matching window (MW). In the MW, every column of the signature matrix is related to four consecutive sections since a RoI can present four different alignments regarding to CSig. Therefore, the MW width, in number of sections, is the same as the signature matrix width, multiplied by four. One of the principal tasks of the SMU is to align the RoI with CSig in the MW, this is the process that we have called Aligning.

The module in charge of addressing the candidate signature from the MA is the The Unique Substrings Detector component. In this module, brute force detection is performed to find out u-substrings in the data flow. Meaning that every u-substring is detected by four matchers, one for each possible shift of the signature. Since the u-substrings are of short length, the matchers consumes few resources. The alignment detection component function is to find out the current RoI alignment. This is carried out by finding the location of the u-substring first character in the MW. Recall that these characters are contained in the aligning column. Character Matrix component is consistent with the character matrix representation as depicted figure 2. In this, a four-to-one multiplexor per matrix cell is deployed. Once the alignment of the RoI is known, this is used to control the array of multiplexers that performs the alignment.

The Character Decode Component receives a RoI aligned with the CSig. In this component, the characters of the CSig are decoded and compared against those of the RoI. There is one multiplexor per column controlled by the current column value (CSig re-encoded character). A character match occurs when the selected input of the multiplexor is asserted, meaning that the re-encoded character and the character in the RoI are equal. The output of the Character

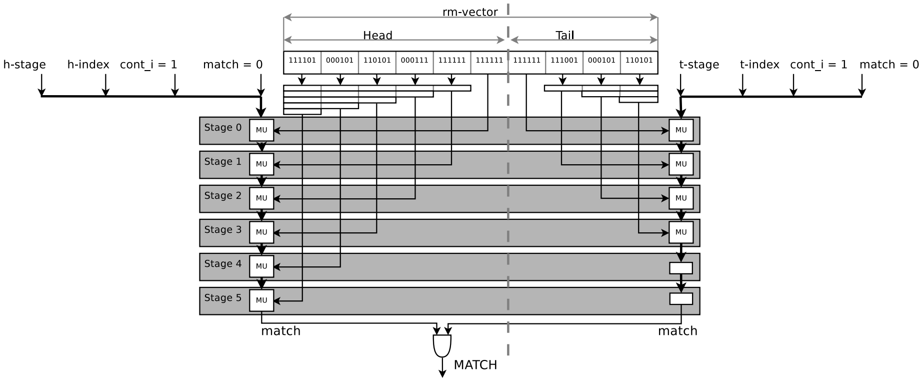


**Fig. 3.** Matching logic component

Decoder component is a bit vector named raw matching vector, rm-vector for short, and its length is equal to the signature matrix width. Each bit in this vector represents a match in the corresponding column. The central idea is to count the number of consecutive one's in the range of bits occupied by the candidate signature in the rm-vector, and signal a match result, when this number is equal to the signature size. This is performed by the Matching Logic Component presented in figure 3. The typical way of performing this operation is by saving a bit mask per signature with all ones out of the signature range and all zeroes in the signature range, then perform a typical masking operation when needed.

Reconsider the example with the signature matrix width of 136 columns, an equal number of bits would be needed to store at every memory entry, leading to an increment of $\lceil 136/36 = 4 \rceil$ additional memory blocks. We propose a different approach in the architecture presented in figure 3. In this, the number of mask bits required for the same example is reduced to only seventeen bits. The rm-vector is split into slices. Each slice takes six consecutive bits from the rm-vector. In the first step of the pipeline, the inner most slice of each section i.e. the closest to the dashed line, are processed. It continues with the next slice, and so on, until the outer most slice. The slice processing at each pipeline step is carried out by the Matching Units (MU) located at both sides of the pipeline.

The MUs also conform a pipeline of four signals, these are: stage, index, continuity, and match. We propose a bit mask composed by two pairs of stage and index values, one for the head section, and one for the tail section, these are h-stage, h-index, t-stage and t-index, respectively. These are stored in the entry together with the signature. Stage represents the outer most slice of the rm-vector occupied by signature, while index is a bit mask with all ones in the bits allowed by the signature in that slice. For example, a signature with rm-vector depicted in fig 3, whose signature has 15 characters in the head and 9 characters in the tail, his corresponding stage and index values are: *h-stage = 2*, *h-index = "000111"*, *t-stage = 1* and *t-index = "111000"*. The number of bits required for the h-stage and t-stage together is equal to the rm-vector width divided by six,

**Table 1.** Signature Matcher implementation

| Virtex5FX100T implementation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Signatures** | **Characters** | **SMUs** | **Frec.** | **Thput.** | **BRAMs** | **LUTs** | **LUTs/char** | **Bit/char** |
| 3,739 | 112,431 | 7 | 150MHz | 4.8Gbps | 60 | 35,508 | 0.32 | 20 |

**Table 2.** Comparisons with previous works

| | Comparison with previous works | | | | | |
|---|---|---|---|---|---|---|
| **Arch.** | **device** | **Input width** | **chars.** | **LE/chars** | **bit/char** | **Thput. Gbps** |
| Our approach | VirtexFX100T | 32 | 112,431 | 0.32 | 20 | 4.8 |
| Baker and Prasanna [3] | Virtex2P100 | 32 | 19,508 | 0.65 | 0 | 7.3 |
| Hwang et. al. [4] | StratixERS140 | 32 | 3,028 | 1.5 | 0 | 11.6 |
| Serrano et. al. [9] | VirtexFX100T | 32 | 5,024 | 1.62 | 0 | 5.69 |
| Kennedy et. al. [5] | Stratix | 16 | 109,467 | 0.63 | 61 | 7.4 |
| Prasanna and Le. [8] | VirtexFX200T | 16 | 217,680 | n/a | 11 | 3.2 |
| Lin and Chang. [6] | n/a | 8 | 36,359 | n/a | 32 | 4 |
| Guinde and Ziavras. [7] | virtex2P70 | 8 | 105,763 | 0.052 | 17.7 | 2.4 |

and the indexes sum twelve bits. For the example exposed before, the overall bits required are seventeen, compared with the original 139 bits required, this means a reduction of 86.3%.

## 5 Experiments and Results

Table 1 shows the results of the architecture implementation for a signature set of 3,739 signatures from Snort database [12]. For the Virtex-5 FX100T device containing 64,000 logic elements (LE) and 200 embedded memory blocks, the overall architecture occupies 60% of the resources. Table 2 shows the comparison against previous works. Our architecture presents the best logic cost regarding to others 32-bit-width architectures. The best memory cost is presented by Prasanna and Le [8]. However their throughput of 3.2 Gbps is achieved by using the double port memory feature of embedded memory blocks. By applying the same strategy, our architecture would double the throughput to 9.6 Gbps while maintaining the same memory cost. The largest Virtex5 device has 207,360 LE, the same architecture can be replicated up to 5 times in this device, achieving an aggregated throughput of 24 Gbps. Likewise, we estimate a character capacity of more than 500K characters.

## 6 Conclusions

We have presented a multi-character architecture which exploits intensively both, memory and logic resources. The replication of hardware is significantly reduced, which leads to a better use of resources, lowering the cost per character compared to others multi-character architectures. Our character re-codification method allows storing one signature in a memory entry. Therefore we can compare the entire signature in one clock cycle. In addition, we have presented a uniform architecture capable of matching non-uniform signatures. If the double port access feature of embedded memory blocks is used the throughput can be doubled, taking into account the capacity of larger FPGA devices, a similar implementation as the one presented here can be replicated up to five times on a Virtex5-330T device.

## References

1. Endorf, C., Schultz, E., Mellander, J.: Intrusion detection and prevention. Mc-Graw-Hill (2004)
2. Ghorbani, A., Lu, W., Tavallaee, M.: Network intrusion detection and prevention: concepts and techniques, vol. 47. Springer (2010)
3. Baker, Z.K., Prasanna, V.K.: Automatic synthesis of efficient intrusion detection systems on fpgas. IEEE Trans. Dependable Secur. Comput. 3(4), 289–300 (2006)
4. Hwang, W.J., Ou, C.M., Shih, Y.-N., Lo, C.T.D.: High throughput and low area cost fpga-based signature match circuit for network intrusion detection. Journal of the Chinese Institute of Engineers 32(3), 397–405 (2009)
5. Kennedy, A., Wang, X., Liu, Z., Liu, B.: Ultra-high throughput string matching for deep packet inspection. In: Proceedings of the Conference on Design, Automation and Test in Europe, DATE 2010, pp. 399–404 (2010)
6. Lin, C.-H., Chang, S.-C.: Efficient pattern matching algorithm for memory architecture. IEEE Trans. Very Large Scale Integr. Syst. 19(1), 33–41 (2011)
7. Guinde, N.B., Ziavras, S.G.: Efficient hardware support for pattern matching in network intrusion detection. Computers & Security 29(7), 756–769 (2010)
8. Prasanna, V.K., Le, H.: A Memory-Efficient and Modular Approach for Large-Scale String Pattern Matching. IEEE Transactions on Computers 62(5), 844–857 (2013)
9. Serrano, J.M.B., Palancar, J.H.: String alignment pre-detection using unique subsequences for FPGA-based network intrusion detection. Computer Communications 35(6), 720–728 (2012)
10. Serrano, J.M.B., Palancar, J.H., Cumplido, R.: Multi-character cost-effective and high throughput architecture for content scanning. In: Microprocessors and Microsystems (in press, 2013) (accepted manuscript), available online August 22: http://authors.elsevier.com/sd/article/S0141933113000999
11. Wang, H., Pu, S., Knezek, G., Liu, J.-C.: MIN-MAX: A Counter-Based Algorithm for Regular Expression Matching. IEEE Transactions on Parallel and Distributed Systems 24(1), 92–103 (2013)
12. Snort, http://www.snort.org

# Ants Crawling to Discover the Community Structure in Networks

Mariano Tepper and Guillermo Sapiro[⋆]

Department of Electrical and Computer Engineering, Duke University, USA
{mariano.tepper,guillermo.sapiro}@duke.edu

**Abstract.** We cast the problem of discovering the community structure in networks as the composition of community candidates, obtained from several community detection base algorithms, into a coherent structure. In turn, this composition can be cast into a maximum-weight clique problem, and we propose an ant colony optimization algorithm to solve it. Our results show that the proposed method is able to discover better community structures, according to several evaluation criteria, than the ones obtained with the base algorithms. It also outperforms, both in quality and in speed, the recently introduced FG-Tiling algorithm.

## 1 Introduction

Networks are frequently used to describe many real-life scenarios were units interact with each other (e.g., see [1,2] and references therein). A seemingly common property to many networks is *community structure*, i.e., that is, networks can be divided into groups such that intra-group connections are denser than inter-group ones. The ability to find and analyze these communities sheds light on important characteristics of a network. However, the best way to establish the community structure is still disputed. Addressing this is the topic of this work.

Let $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ be the graph to analyze, where $\mathcal{U}$ is the set of nodes and $\mathcal{E}$ is the set of edges (in the following we indistinguishably use the terms graph and network). Generically, we consider that a community-detection algorithm provides a set $\mathcal{X}$ of candidate communities ($\mathcal{X}$ can be a partition of the node set $\mathcal{U}$ or a hierarchy of subsets in $\mathcal{U}$, from which the best partition may latter be extracted). Let us consider that a pool of $C$ such algorithms provides a universe of candidates $V = \bigcup_{i=1}^{C} \mathcal{X}_i$. We build a new graph $G = (V, E, \omega)$, where $(u, v) \in E$ if and only if communities $u, v \in V$ do not overlap, and $\forall u \in V\ \omega(u)$ is a measure associated with the quality of partition $u$. In this work, we do not address the case in which communities overlap, while our technique could be considered for this as well.

We formulate the problem of finding the best community structure in a network as a *patchwork* algorithm: instead of building a community set $\mathcal{X}$ from scratch, we browse through $V$ and build a new solution by combining the best

[⋆] Work partially supported by NSF, ONR, NGA, ARO, DARPA, and NSSEFF. Work partially done while the authors were with the Department of Electrical and Computer Engineering, University of Minnesota.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 552–559, 2013.

communities in $V$ in such a way that no overlap exists among them. In other words, the proposed *meta* algorithm selects the set of communities with maximum total weight among the ones that are fully interconnected by $E$. This is a maximum-weight clique problem on the graph $G$. This type of formulation was simultaneously introduced in [3,4] for image segmentation. In this work, we propose to follow and extend this approach for community structure discovery.[1]

Let us first formally pose the maximum-weight clique (MWC) problem. Let $G = (V, E, \omega)$ be a weighted graph where $V$ is a set of nodes, $E \subseteq V \times V$ is a set of edges (no self-loops are allowed, that is $\forall u \in V, (u, u) \notin E$), and $\omega : V \to \mathbb{R}^+$ is a node weighting function. For convenience, given a set $C \subseteq V$ we write $\omega(C) = \sum_{u \in C} \omega(u)$. A clique $\mathcal{C} \subseteq V$ is a set such that $\forall u, v \in \mathcal{C}, u \neq v, (u, v) \in E$. The maximum clique problem is to find a clique of maximum cardinality in $G$. The MWC problem is to find a clique $\mathcal{C}$ of maximum weight $\omega(\mathcal{C})$. Both problems are known to be NP-Hard.

A popular approach for solving hard discrete combinatorial problems such as MWC is the use of metaheuristics. Ant Colony Optimization (ACO) is a variant of swarm intelligence in which the system is made of a population of simple agents interacting locally with one another and with their environment. Although each agent builds a solution following an extremely simple set of rules, cooperation among the population leads to the emergence of intelligent behavior. When looking for food, real ants deposit pheromones on the ground; then, the probability that other ants follow a particular path is proportional to the level of pheromones in that path. Similarly, the solution construction process in ACO is stochastic and is biased by a “pheromone” model. Specifically, artificial ants will explore more thoroughly regions of the solution space where good solutions were previously found.

In this work we propose an ACO algorithm for the MWC problem and, as already discussed, we apply it to the problem of selecting the best community structure from a set of community candidates. We propose then a double collaboration structure, where multiple algorithms collaborate to propose community structures, and the ants collaborate to find the best one from all the possibilities. The algorithm presents a good balance between the amount of exploration and computational efficiency. We provide experiments that show the pertinence of the proposed method.

The remainder of the work is organized as follows. In Section 2 we present our ACO algorithm for the MWC problem. In Section 3 we discuss the experimental results and finally we provide some closing remarks in Section 4.

## 2 ACO for the MWC Problem

A few variants of ACO have been proposed for the maximum clique problem [6,7]. We adapt Solnon and Fenet’s algorithm [7] for solving the MWC problem. Being a metaheuristic, the change is minimal: instead of searching for the clique with maximum cardinality, we seek the clique with maximum weight.

[1] An approach based on metaheuristics was indepently and simultaneously developed for clustering in [5], solving the related maximum-weight independent set problem using simulated annealing.

We associate a pheromone level $\tau(v)$ to each node $v \in V$, and use an ACO variant, called Max-Min Ant System (MMAS), where pheromone levels are bounded, that is $\forall v \in V,\ \tau(v) \in [\tau_{\min}, \tau_{\max}]$. These bounds ensure that no region in the solution space will be excessively over or under sampled.

Given a clique $\mathcal{C}$, we define the set of possible additions to $\mathcal{C}$ as

$$\mathrm{PA}(\mathcal{C}) = \{u \mid u \in (V \smallsetminus \mathcal{C}) \wedge \forall v \in \mathcal{C}, (u,v) \in E\}\,. \tag{1}$$

That is, given a clique $\mathcal{C}$, $\forall u \in \mathrm{PA}(\mathcal{C}), \mathcal{C} \cup \{u\}$ is also a clique.

The overall ACO procedure is described by Algorithm 1. In each iteration, a certain number $K$ of ants explore the solution space. Each ant builds a randomized solution, favoring nodes with higher pheromone levels.

The ACO algorithm regulates its overall behavior through pheromone trails: cooperation between ants rises from the optimality of previous experiences. When adding a new node to a partial solution, the pheromone level in each node is used to bias the election, i.e., the higher $\tau(v)$, the more probably a node $v$ will be chosen. An ant then selects a node $v$ with probability

$$p_\alpha(v \mid \mathcal{C}) = [\tau(v)]^\alpha \left(\textstyle\sum_{u \in \mathrm{PA}(\mathcal{C})} [\tau(u)]^\alpha\right)^{-1}, \tag{2}$$

where $\alpha$ is a parameter of the algorithm.

As in every ACO algorithm, pheromones evaporate over time. This ensures that exploration will not get stuck around good previous solutions and will diversify over time. We set the evaporation rule

$$\forall u \in V,\ \tau(u) = \max(\tau_{\min}\ ,\ \rho \cdot \tau(u)), \tag{3}$$

where $\rho$ is a parameter of the algorithm.

As previously stated, pheromone levels must be increased for those nodes that belong to good solutions. Let $\mathcal{C}^*$ be the best solution in the current iteration and let $\mathcal{C}_{\text{best}}$ be the best solution in all previous iterations, including the current one. We only update those nodes that belong to $\mathcal{C}^*$, according to the rule

$$\tau(u) = \min\left(\tau_{\max}\ ,\ \tau(u) + [1 + \omega(\mathcal{C}_{\text{best}}) - \omega(\mathcal{C}^*)]^{-1}\right). \tag{4}$$

## 2.1 Local Search

The ACO algorithm uses local search as a post-processing to improve the best solution found by the ants. Adding this intelligence to every ant would be computationally costly and would hinder the stochastic search. Notice that the ACO algorithm uses the local search method as a black box and, as such, any suitable technique can be employed. Many local search schemes have been proposed for the maximum clique problem. Their goal is to avoid local minima by exploring the neighborhood (in the solution space) of an initial solution.

Katayama et al. [8] propose to examine the k-opt neighborhood of an initial solution, defined as the set of neighbors that can be obtained by a sequence of several add and drop moves that are adaptively changed in the feasible search space. For this task, they introduce an efficient algorithm called k-opt local search (KLS). KLS has proven capable of finding satisfactory cliques with reasonable running times. In this work, we extend KLS for the MWC problem. Pseudocode of the weighted-KLS is presented in Algorithm 2.

**Algorithm 1:** ACO algorithm for the MWC problem.

```
$\mathcal{C}_{\text{best}} \leftarrow \emptyset$;   $\forall u \in V, \tau(u) \leftarrow \tau_{\max}$;
while number of iterations is lower than $T$ do
    for ant $k$ such that $1 \leq k \leq K$ do
        Choose a seed $u_k \in V$ at random;
        $\mathcal{C}_k \leftarrow \{u_k\}$;
        while $\text{PA}(\mathcal{C}_k) \neq \emptyset$ do
            Choose a vertex $v \in \text{PA}(\mathcal{C}_k)$ with probability $p_\alpha(v \mid \mathcal{C}_k)$;
            $\mathcal{C}_k \leftarrow \mathcal{C}_k \cup \{v\}$;
    $\mathcal{C}_{\max} \leftarrow \text{argmax}_{\mathcal{C}_k} \ \omega(\mathcal{C}_k)$;
    $\mathcal{C}^* \leftarrow$ improve solution $\mathcal{C}_{\max}$ using local search (see Sec. 2.1 and Algorithm 2);
    if $\omega(\mathcal{C}^*) > \omega(\mathcal{C}_{\text{best}})$ then $\mathcal{C}_{\text{best}} \leftarrow \mathcal{C}^*$;
    Evaporate pheromone levels (Eq. (3));
    Update pheromone levels of vertices in $\mathcal{C}^*$ (Eq. (4));
return $\mathcal{C}_{\text{best}}$
```

**Algorithm 2:** Local search.

```
repeat
    $P \leftarrow V$;   $\mathcal{C}_{\text{prev}} \leftarrow \mathcal{C}$;   $D \leftarrow \mathcal{C}$;
    $g \leftarrow 0$;   $g_{\max} \leftarrow 0$;
    while $D \neq \emptyset$ do
        if $\text{PA}(\mathcal{C}) \cap P \neq \emptyset$ then // add phase
            $v^* \leftarrow \underset{v \in \text{PA}(\mathcal{C})}{\text{argmax}} \ \text{input}_\omega(\mathcal{C}, v)$;
            $\mathcal{C} \leftarrow \mathcal{C} \cup \{v^*\}$;   $g \leftarrow g + \omega(v^*)$;
            $P \leftarrow P \setminus \{v^*\}$;
            if $g > g_{\max}$ then
                $g_{\max} \leftarrow g$;   $\mathcal{C}_{\text{best}} \leftarrow \mathcal{C}$
        else // drop phase
            $v^* \leftarrow \underset{v \in (\mathcal{C} \cap P)}{\text{argmax}} \ \omega(\text{PA}(\mathcal{C} \setminus \{v\}))$;
            $\mathcal{C} \leftarrow \mathcal{C} \setminus \{v^*\}$;   $P \leftarrow P \setminus \{v^*\}$;
            $g \leftarrow g - \omega(v^*)$;
            if $v^* \in \mathcal{C}_{\text{prev}}$ then $D \leftarrow D \setminus \{v^*\}$
until $g_{\max} > 0$;
return $\mathcal{C}$
```

KLS is a greedy algorithm, each decision of adding or removing a node from the current solution is made by maximizing the immediately obtained reward:

- When adding a node, the most desirable candidate in $\text{PA}(\mathcal{C})$ is picked. In the unweighted case, the desirability of a node is given by its degree; in the weighted case, we define it as
$$\forall u \in \text{PA}(\mathcal{C}), \ \text{input}_\omega(\mathcal{C}, u) = \omega(u) + \textstyle\sum_{v \in \text{PA}(\mathcal{C})} \omega(v). \tag{5}$$
- When removing a node, we pick the node whose removal will produce the most desirable set of candidates $\text{PA}(\mathcal{C})$ for future additions. In the absence of weights, the appeal of $\text{PA}(\mathcal{C})$ is determined by its size $|\text{PA}(\mathcal{C})|$; when weights are present, the appeal is defined by $\omega(\text{PA}(\mathcal{C}))$.

KLS uses the set $P$ as a mechanism to avoid incurring in loops of addition/ removal of the same set of nodes. Also note that the functions PA and $\text{input}_\omega$ can be computed and updated very efficiently [9], and do not present a significant computational overhead.

**Complexity.** Let us begin by analyzing the KLS algorithm. Due to its parameterless nature, computing its time complexity is not an easy task. We estimate the overall complexity as $O(Ln^2h)$, where $n = |V|$, $h$ is the size of the initial (input) clique, and $L$ is the number of executions of the outer cycle, although a more realistic *practical* bound would be $O(h)$. In the ACO algorithm, each ant $k$ can build its own solution in $O(\deg(u_k)^2)$, where $\deg(u_k)$ denotes the degree of seed $u_k$. Thus the total complexity of the proposed algorithm is $O(T(Kd^2 + h))$, where $d = \max_{u \in V} \deg(u)$, if we use the aforementioned $O(h)$ for KLS.

The complexity of the ACO algorithm is extremely lower than the $O(Nd^3)$ of FG-Tiling [4]. This can be observed in practice, where our algorithm systematically outperformed FG-Tiling in running-time. Notice that FG-Tiling is deterministic, thus its worst case complexity provides a tight bound.

**Community Quality Measures.** Many different measures have been used for assessing the quality of a given community structure. We work with the standard modularity [2] and a new measure called surprise [10]. Our method is of course completely agnostic to the measure's nature and it can directly profit from any improvement in this respect.

Modularity is by far the most popular measure for assessing the quality of a partition. We will denote by $Q(\mathcal{P})$ the modularity of partition $\mathcal{P}$. Because of space constraints, we do not provide its formal definition.

The surprise $S$ of partition $\mathcal{P}$ is defined [10] as $S(\mathcal{P}) = -\log H(F, N, m, b)$, where $H$ is the tail of the hypergeometric distribution, $F = n(n-1)/2$, $N = \sum_{P \in \mathcal{P}} |P|(|P|-1)/2$, and $b = \sum_{P \in \mathcal{P}} e_P$. This represents the probability of obtaining $b$ intra-community edges in $m$ draws, without replacement, from a finite population of size $F$ containing $N$ successes. We can use surprise in our framework assuming that the subsets in a random partition are i.i.d., that is, $S(\mathcal{P}) \approx \sum_{P \in \mathcal{P}} -\log H(F, N_P, m, e_P)$, where $N_P = |P|(|P|-1)/2$.

## 3 Experimental Results

In this section we evaluate the community structure discovery results of the proposed ACO algorithm for the MWC problem. We use two different base algorithms for community detection: Walktrap [11] and Jerarca [1]. Both algorithms provide a hierarchy of communities and then globally threshold the hierarchy at different levels, thus obtaining a partition per level, and finally select the best partition among them. Let $\mathcal{H}_W$ and $\mathcal{H}_J$ be the hierarchies provided by Walktrap and Jerarca, respectively. As already explained, we create the graph $G = (V, E)$, where $V = \left(\bigcup_{C \in \mathcal{H}_W} C\right) \cup \left(\bigcup_{C \in \mathcal{H}_J} C\right)$ and $(C, C') \in E$ if and only if $C \cap C' = \emptyset$. We then look for the MWC in $G$. We compare our results with FG-Tiling [4], a recently introduced and successful MWC solver in the context of image segmentation. For all experiments, unless specifically indicated, we set the ACO parameters to $\alpha = 1$, $\tau_{\min} = 0.01$, $\tau_{\min} = 10$, $\rho = 0.98$, $K = 100$, and $T = 1000$.

In Table 1 we show the community structure discovery results on different networks used in the literature. We first observe that the proposed approach, finding an approximate solution of the MWC problem, outperforms the hierarchical approach. Indeed, FG-Tiling and ACO outperform Jerarca and Walktrap. Of course, this comes at the cost of extended running times, Jerarca and Walktrap being very fast algorithms. On a finer observation level, ACO consistently obtains better solutions than the deterministic FG-Tiling, confirming in practice the ability of adaptive stochastic algorithms for exploring the solution space. We provide graph plots in Fig. 1 to show that the differences in the measures shown in Table 1 actually correspond to different community structures. Note that in no experiment we judge the quality of the obtained partition from the graph plots, the assessment is solely based on the selected standard measure.

Table 2 presents running-time examples of both algorithms (implemented in Python). Note that all of our current implementations can be further optimized. These times do not reflect the best times that can be achieved with these methods, but they serve to corroborate the previously presented complexity bounds.

**Table 1.** Modularity $Q(\mathcal{P})$ and surprise $S(\mathcal{P})$ values of the best partition $\mathcal{P}$ found by each tested method on different networks (WT stands for Walktrap)

| | Modularity | | | | Surprise | | | |
|---|---|---|---|---|---|---|---|---|
| | Jerarca | WT | FG-Tiling | ACO | Jerarca | WT | FG-Tiling | ACO |
| 1dc.64[1] | 0.212 | 0.251 | 0.251 | **0.251** | 148.65 | 146.06 | 156.4 | **158.95** |
| Les Misérables [12] | 0.490 | 0.403 | 0.497 | **0.512** | 361.14 | 325.98 | 378.91 | **378.91** |
| Chesapeake [13] | 0.356 | 0.102 | 0.339 | **0.356** | 41.26 | 14.16 | 45.59 | **46.00** |
| Dolphins [14] | 0.246 | 0.445 | 0.434 | **0.461** | 144.40 | 90.00 | 122.55 | **152.79** |
| Aegean34 [15] | 0.529 | 0.571 | 0.571 | **0.571** | 130.97 | 126.78 | 134.18 | **134.18** |

[1] `http://www2.research.att.com/~njas/doc/graphs.html`

**Table 2.** Running time (in seconds) of FG-Tiling and ACO for different networks

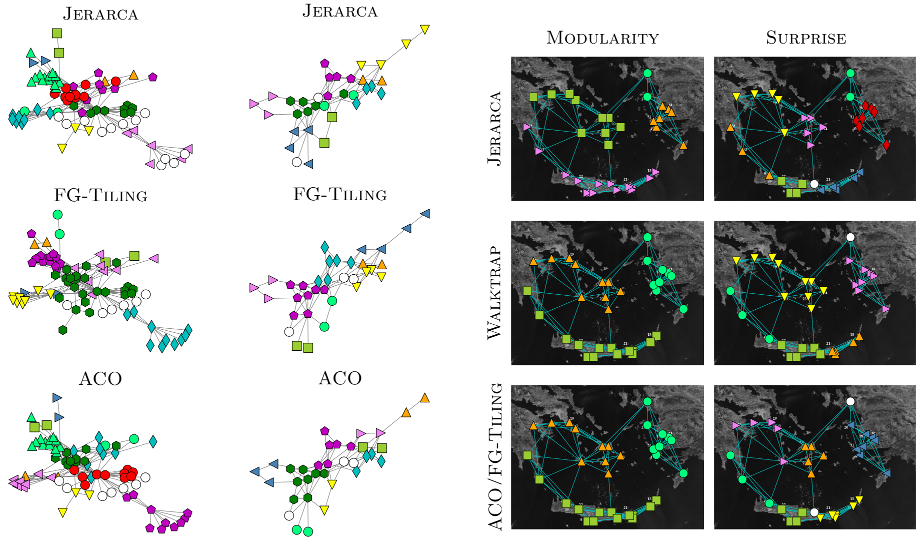
| | Aegean34 | Chesapeake | Les Misérables | 1dc.64 |
|---|---|---|---|---|
| FG-Tiling | 975 | 1200 | 37790 | 48630 |
| ACO | 109 | 130 | 748 | 930 |

We also ran an experiment comparing three different versions of the proposed ACO algorithm: one that disallows cooperation and uses local search, ones that allows cooperation but does not use local search, and the proposed one which allows cooperation and uses local search. The emergence of cooperation can be prevented by setting $\alpha = 0$, see Eq. (2). We ran this comparison on a random graph $\mathcal{G}$ where all edges are i.i.d. and belong to $\mathcal{G}$ with a fixed probability. In average there should be no communities in $\mathcal{G}$, and hence no trivial solution can be prematurely found. This is observed in practice since all modularity values are close to zero. The results are shown in Fig. 2. Clearly, the version that allows both collaboration and local search outperforms the other variants.

To show the generality of the proposed approach, we include some basic results on image segmentation in Fig. 3. Briefly, each region in a segmentation can be viewed as a "community" and we compute its weight using a basic and untuned version of the method in [3]. Let $\mathcal{H}_{\mathrm{UCM}}$ be the hierarchy produced by the state-of-the-art UCM image segmentation algorithm [16]. We select the global threshold on $\mathcal{H}_{\mathrm{UCM}}$ such that the resulting partition has maximum total weight. When $G$ is built from $\mathcal{H}_{\mathrm{UCM}}$ in our framework, the proposed algorithm is able to improve the segmentation, obtaining a partition with higher total weight. More tuned or sophisticated weighting functions would further improve the obtained results, these are obtained directly from the proposed general technique.

## 4 Conclusions

We proposed an algorithm for community structure discovery. Instead of building our own custom community structure from scratch, we take the output of several community detection algorithms (they can be the same algorithm executed with different parameters, and also provide hierarchies or partitions), and compose a new structure by combining the best communities in each. In this way, we can make use of multiple algorithms that provide globally suboptimal solutions which contain some optimal communities. The combination of these optimal communities leads to the creation of a new and globally better solution.



(a) Les Misérables (with modularity).

(b) Chesapeake (with surprise).

(c) Aegean34 network

**Fig. 1.** Different communities are represented by nodes of different colors and shapes. All singleton communities (composed of one node) are depicted with white circles. Differences in the quality measure (see Table 1) create different community structures.

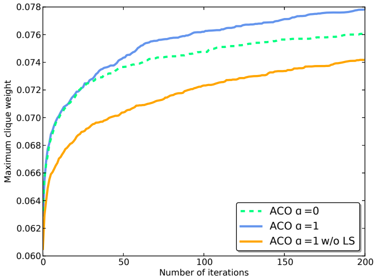


**Fig. 2.** The proposed algorithm benefits from both the collaborative nature of ACO and the local search (weighted KLS). It outperforms the non-cooperative ACO algorithm ($\alpha = 0$) and a version that does not use local search (termed "w/o LS"). Weights correspond to modularity values.

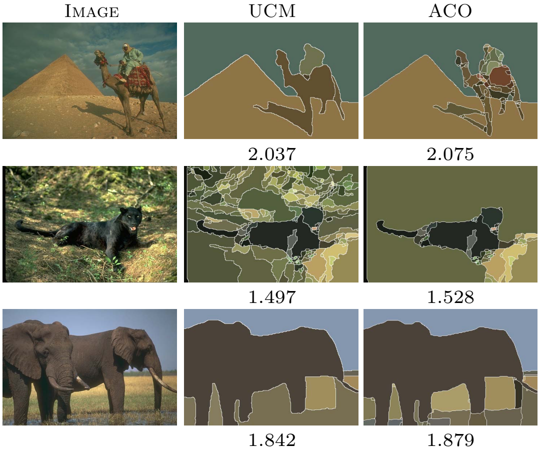


**Fig. 3.** Image segmentation results. We show the weight of the obtained partition. The proposed method is able to compose a better result by searching through the UCM hierarchy.

We showed results that confirm in practice the theoretical benefits of the proposed algorithm. Namely, the stochastic nature of ACO helps in handling the non-convex nature of the MWC problem's solution space, by exploring it "fairly." The collaborative aspect of ACO, ensures that "fairness" is distributed smartly: potentially appealing zones of the solution space attract more attention and are thus explored more thoroughly. We further strengthen the ACO algorithm by using a local search heuristic. We adapt the KLS algorithm to handle weighted graphs, obtaining both efficient and solid performances.

As future work, we plan on adding support for weighted edges in the model, which could be used to model relationships between communities. We also plan to extend our current model to the case of overlapping communities, for which suitable quality measures are needed. Additionally, a C++ implementation will allow to analyze larger networks. Furthermore, the ant exploration phase in the proposed ACO algorithm can be parallelized with ease.

## References

1. Aldecoa, R., Marín, I.: Jerarca: efficient analysis of complex networks using hierarchical clustering. PLoS ONE 5(7), e11585+ (2010)
2. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(026113) (2004)
3. Brendel, W., Todorovic, S.: Segmentation as maximum-weight independent set. In: NIPS (2010)
4. Ion, A., Carreira, J., Sminchisescu, C.: Image segmentation by figure-ground composition into maximal cliques. In: ICCV (2011)
5. Li, N., Latecki, L.J.: Clustering aggregation as maximum-weight independent set. In: NIPS (2012)
6. Leguizamon, G., Michalewicz, Z., Schutz, M.: An ant system for the maximum independent set problem. In: CACIC, vol. 2 (2001)
7. Solnon, C., Fenet, S.: A study of ACO capabilities for solving the maximum clique problem. J. Heuristics 12(3), 155–180 (2006)
8. Katayama, K., Hamamoto, A., Narihisa, H.: An effective local search for the maximum clique problem. Inf. Process. Lett. 95(5), 503–511 (2005)
9. Battiti, R., Protasi, M.: Reactive local search for the maximum clique problem. Algorithmica 29(4), 610–637 (2001)
10. Aldecoa, R., Marín, I.: Deciphering network community structure by surprise. PLoS ONE 6(9), e24195+ (2011)
11. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algorithms Appl. 10(2), 284–293 (2004)
12. Knuth, D.E.: The Stanford GraphBase: A Platform for Combinatorial Computing. ACM, New York (1993)
13. Baird, D., Ulanowicz, R.E.: The seasonal dynamics of the Chesapeake bay ecosystem. Ecol. Monogr. 59(4), 329–364 (1989)
14. Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M.: The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behav. Ecol. Sociobiol. 54(4), 396–405 (2003)
15. Evans, T.S., Rivers, R.J., Knappett, C.: Interactions in space for archaeological models. Adv. Complex Syst. 15, 1150009+ (2011)
16. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From contours to regions: an empirical evaluation. In: CVPR (2009)

# Boruvka Meets Nearest Neighbors*

Mariano Tepper[1,**], Pablo Musé[2], Andrés Almansa[3], and Marta Mejail[4]

[1] Department of Electrical and Computer Engineering, Duke University
mariano.tepper@duke.edu
[2] Instituto de Ingeniería Eléctrica, Facultad de Ingeniería,
Universidad de la República
pmuse@fing.edu.uy
[3] CNRS - LTCI UMR5141, Telecom ParisTech
andres.almansa@telecom-paristech.fr
[4] Departamento de Computación, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires
marta@dc.uba.ar

**Abstract.** Computing the minimum spanning tree (MST) is a common task in the pattern recognition and the computer vision fields. However, little work has been done on efficient general methods for solving the problem on large datasets where graphs are complete and edge weights are given implicitly by a distance between vertex attributes. In this work we propose a generic algorithm that extends the classical Boruvka's algorithm by using nearest neighbors search structures to significantly reduce time and memory consumption. The algorithm can also compute in a straightforward way approximate MSTs thus further improving speed. Experiments show that the proposed method outperforms classical algorithms on large low-dimensional datasets by several orders of magnitude.

## 1 Introduction

The computation of the minimum spanning tree (MST) is a classical problem in computer science. For an undirected weighted graph, it can be simply stated as finding a tree that covers all vertices, called a spanning tree, with minimum total edge cost. It is taught in every course of algorithms and data structure as an example where greedy strategies are successful and it is regarded as one of the first historical foundations of operations research.

Maybe the two most widely known algorithms to compute the MST are Prim's and Kruskal's [1]. There is a third classical algorithm by Boruvka [1] that mysteriously remained shadowed by the other two. This fact is emphasized by the

* Supported by FREEDOM (ANR07-JCJC-0048-01), CNES (R&T Echantillonnage Irregulier DCT/SI/MO - 2010.001.4673), Callisto (ANR-09-CORD-003), ECOS Sud U06E01, ARFITEC (07 MATRH), STIC Amsud (11STIC-01 - MMVPSCV), and the Uruguayan ANII (PR-POS-2008-003).

** Work partially done while M. Tepper was with the Departamento de Computación at the Universidad de Buenos Aires.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 560–567, 2013.

fact that Boruvka's algorithm is also known as Sollin's algorithm, despite the fact that Sollin re-discovered it independently years later.

The MST is particularly interesting for many data analysis tasks in computer vision and pattern recognition. A clear example is clustering, where the classical single-linkage hierarchical algorithm [2] can be proven equivalent to computing the MST. In a seminal work, Zahn [3] studied the benefits of using the MST for clustering. More recently, the MST received attention due to the growth in the size of clustering datasets, e.g., [4,5]. The approximate MST (AMST), suboptimal but faster, also received attention for the same reasons [6].

We now slightly change the definition of the problem to a form more suitable for data analysis (e.g., clustering). Let $M$ be a set and $d : M \times M \rightarrow \mathbb{R}^+$ a distance function. Then $d$ and the pair $(M, d)$ are said to be a metric on $M$ and a metric space, respectively. Given a data set $X \subseteq M$, the MST of $X$ is defined as the MST of the weighted undirected graph $G = (V, E)$ where each $v_i \in V$ is identified with a feature $x_i \in X$, $E = V \times V$ (the graph is complete), and the graph's weighting function $\omega : E \rightarrow \mathbb{R}$ is defined as $\omega((v_i, v_j)) = d(x_i, x_j)$.

The problem is classically addressed by using metric spaces with exploitable specific characteristics like the Euclidean space, e.g., the Euclidean MST is contained in the Delaunay triangulation of $X$ [7]. Recent work has aimed at building an AMST [6] through a clever use of space-filling curves.

Nearest neighbors (NNs) search structures have been used to compute the MST [8]. The approach proved successful; moreover, using such structures allows in addition to compute the AMST in a natural and straightforward way. A revision of this approach is needed, in the light of novel NNs techniques and increasing computational power. More recently, Leibe et al. [9] used NNs techniques for hierarchical clustering using the average-link criterion. Although they improved the method's performance, their algorithm is not suitable for extremely large datasets.

Classical algorithms for computing the MST run in $\mathrm{O}(n^2 \log n)$, where $n = |X|$. However, one must compute all $n(n-1)/2$ distances and thus a double-sided problem appears: (1) storing all $n(n-1)/2$ results for $n \geq 10^5$ is prohibitive; (2) even if results are not stored, for $n \geq 10^5$ the overall running-time is also prohibitive. Keep in mind that, in modern pattern recognition applications, feature sets of $10^5$ or more points are becoming common [10]. In this work we address the MST problem without computing all distances in $E$. We build on Boruvka's approach [1] by an appropriate use of NNs search techniques.

The rest of the paper is structured as follows. In Section 2 we propose a general approach to compute the MST using NNs search structures. Section 3 shows empirical results of the proposed approach on a synthetic dataset. Finally, some final remarks and future work are presented in Section 4.

## 2 A Nearest Neighbors Approach

First let us explain Boruvka's algorithm: it creates a forest (i.e., a set of trees) where each isolated edge is a tree and gradually merges these trees by adding

**Algorithm 1:** Computation of the MST $T = (V, E_T)$ of feature set $X$.

```
1  $E_T \leftarrow \emptyset$;
2  while $|E_T| < |V| - 1$ do
3      $E' \leftarrow \emptyset$;
4      foreach connected component $C$ of $T$ do
5          $(u_m, v_m) \leftarrow \underset{u \in C,\ v \notin C}{\arg\min}\ d(u, v)$;
6          $\delta_m \leftarrow d(u_m, v_m)$;
7          $E' \leftarrow E' \cup \{(u_m, v_m, \delta_m)\}$;
8      while $E' \neq \emptyset$ do
9          $(u_m, v_m, \delta_m) \leftarrow \underset{(u,v,\delta) \in E'}{\arg\min}\ \delta$;
10         $E' \leftarrow E' \setminus \{(u_m, v_m, \delta_m)\}$;
11         if $E_T \cup \{(u_m, v_m, \delta_m)\}$ does not contain cycles then
12             $E_T \leftarrow E_T \cup \{(u_m, v_m, \delta_m)\}$
```

the smallest edge whose endpoints lie on different trees (see Algorithm 1). We propose to express the term in line 4 of Algorithm 1 in terms of finding NNs in the set $V \setminus C$:

$$u_m = \underset{u \in C}{\arg\min}\, d(u, \mathrm{NN}_d(V \setminus C, u)), \tag{1}$$

$$v_m = \mathrm{NN}_d(V \setminus C, u_m), \tag{2}$$

where $\mathrm{NN}_d(A, b)$ returns the NN $a \in A$ of $b$ using metric $d$. We also modify the function $\mathrm{NN}_d(A, b)$ by adding an additional constraint function $\rho : X \to \{0, 1\}$ on the returned element. We denote it by $\mathrm{NN}_{d,\rho}(A, b)$. It returns the NN $a \in A$ of $b$ using metric $d$ such that $\rho(a) = 1$. By setting $\rho(v) = (v \notin C)$ we have

$$\mathrm{NN}_d(V \setminus C, u) = \mathrm{NN}_{d,\rho}(V, u). \tag{3}$$

This kind of problem is sometimes referred to as Foreign NNs in the literature.

We are sure that the desired node $v_m$ is among the $k$ NNs of $u$ where $k = |C| + 1$. Therefore in the worst case, using a naive approach, $\mathrm{NN}_{d,\rho}$ amounts to perform a $k$-NNs search and then a simple check among them by using $\rho$. Note that $k$ is a dynamic (growing) quantity and it is not possible to fix it in advance. The problem is thus of a different nature than finding the MST in a constrained degree graph. Of course, there is no need to compute that many NNs, since the constraint can be directly incorporated in the NN technique.

Priority queues can be used to prune the number of NNs searches performed during the algorithm [8]. We propose to use several priority queues, one for each connected component in a partial (i.e., already computed) MST. The nodes $u_i$ of a partial MST are stored, with their foreign NNs $u_j$, in a priority queue where the priority of a node is the inverse of $d(x_i, x_j)$. The use of a priority queue is indeed interesting in this context, as the next edges to add to the MST are at the top of the priority queues. The top of the queues are removed and the top-priority foreign NNs are added to the MST. After merging two connected components, their priority queues are also merged.

Additionally, the priority queue must be updated, since disjoint connected components are merged and some foreign NNs might not be foreigners anymore. Note that it may not be necessary to update the entire priority queue. This is because the current priority of each of these nodes (the priority before the insertion in the MST) serves as an upper bound of its real priority (the priority after the insertion in the MST). The real priority of a node needs only to be computed when its current priority is on the top of the queue.

We omit the pseudocode of the resulting algorithm because of space constraints, see [11] for further details. Note that the space complexity is still $O(n)$. In the first iteration, there are $n$ queues, each of length 1. In the second iteration there are roughly $n/2$ queues, each of length 2, and so on.

### 2.1 Approximate MST

If we simply relax the search by finding approximate NNs we end up with an AMST algorithm. Approximate NNs queries are much faster than exact ones, specially in high-dimensional spaces.

Typically, $\mathrm{ANN}_d(X, u, \eta)$ ensures that, if the true NN is at distance $\delta$, the approximate NN is at a distance lower than $\delta(1+\eta)$. Note that AMSTs can also be obtained by using a probability bound on the NN distance [12].

Lai et al. [6] have previously studied AMSTs. Their approximation is obtained by using space-filling structures, i.e., Hilbert curves. Their work differs from ours in two central points. First, our algorithm allows to combine MSTs and AMSTs in a single framework, in which the only difference between them is a relaxation parameter. Their work is restricted to AMSTs. Second, Hilbert curves are fractal and the space-filling accuracy follows an exponential scale. It relies on a scale parameter that has a non-intuitive meaning and which is difficult to choose. It is not straightforward to set automatically a suitable scale for a given point set configuration. The relaxation parameter in our method has a clear interpretation and it is easy to monitor its effect.

## 3 Experimental Results

For the NN computations, choose the list-of-clusters (LOC) structure [13,14]. It is reported to be very efficient and resistant to the intrinsic dimensionality of the data set. It can also be implemented in primary and in secondary memory. See [11] for further details on how to adapt the structure for our specific purposes.

As distance computations are the dominating speed factor, we measure performance and complexity as a function of them. We sample points from a uniform distribution in the unit hyper-cube. We tested with four different dimensionalities $\mathbb{R}^2$, $\mathbb{R}^5$, $\mathbb{R}^{10}$ and $\mathbb{R}^{20}$. We compared the following methods (see Table 1): **Bvka**: the classical Boruvka's algorithm, where all distances are precomputed

**Table 1.** The methods compared in this work. $\overline{s}$ stands for average number of distance operations needed to complete a NNs search.

| Method | Solution | Number of distances | | Space | Search |
|---|---|---|---|---|---|
| | | computed | stored | complexity | speed |
| Bvka | MST | $n(n-1)/2$ | all | $O(n^2)$ | — |
| Bvka-O | MST | $O(n^2 \log n)$ | none | $O(1)$ | linear |
| Bvka-LOC | MST | $O(\overline{s}n \log n)$ | none | $O(n)$ | sub-linear |
| Bvka-PQ-LOC | MST | $O(\overline{s}n \log n)$ | $n-1$ | $O(n)$ | sub-linear |
| Bvka-A $\eta$ | AMST | $O(\overline{s}n \log n)$ | $n-1$ | $O(n)$ | sub-linear |

and stored in memory; **Bvka-O**: the proposed algorithm where an online linear search is used to compute NNs; **Bvka-LOC**: the proposed algorithm where NNs are computed online by using LOC; **Bvka-PQ-LOC**: the proposed algorithm where NNs are computed online by using LOC and priority queues; **Bvka-A** $\eta$: Bvka-PQ-LOC modified to compute the AMST by using approximate NNs. Note that the reduced memory complexity of the algorithm guarantees that we will be able to treat large datasets without memory issues.

Comparisons were made for relatively small feature sets ($|X| \leq 10^4$) to be able to compare with a classical MST implementation. A summary of our results is shown in Figure 1. Our method exhibits a very strong performance improvement in low dimensions (Fig. 1, top row). Bvka-LOC and Bvka-PQ-LOC in both cases outperforms Bvka several orders of magnitude. We can also notice a strong performance degradation of Bvka-LOC with the increase of dimensionality (Fig. 1, bottom row). The only cause is the NNs search structure. It is a well known fact that the performance of NNs search structures tends to become linear in high-dimensions. In any case, our method is generic: any NN structure can be used. Another structure may provide better results in high dimensions and we plan to explore these issues in future work.

Table 2a summarizes the results from Figure 1 by analyzing the slope of the different curves. The proposed approach lowers in practice the number of distance computations needed to solve the problem. The quadratic profiles of Bvka and Bvka-O are reduced to supralinear (e.g., $n^{1.6}$ approximately) by Bvka-LOC and Bvka-PQ-LOC. As stated, the latter shows a computational cost which is less sensitive to an increase in dimensionality.

We provide a simple example of the incidence of using the AMST, shown in Figure 2a. We use $X$ uniformly distributed on the square $[0,1]^2$ and Euclidean distance. Computing the MST required 9613 distance computations with our algorithm, while taking 9155, 8705 and 7840 with $\eta = 0.1$, $\eta = 0.2$, $\eta = 0.5$ respectively. There is an important improvement in performance while the number of topology changes is small. Moreover, when carefully inspected, these changes are reasonable. It is a well known fact that (even little) jitter noise in the dataset greatly affects the topology of the MST [4]: computing the AMST can be seen as perturbing the dataset with such a noise. Usually $\eta$ is chosen to be quite small, and its use has more meaning in large and high-dimensional datasets. In our toy example, keeping $\eta$ small does not introduce changes in the topology of the tree. We exaggerated $\eta$ to show actual topology changes.

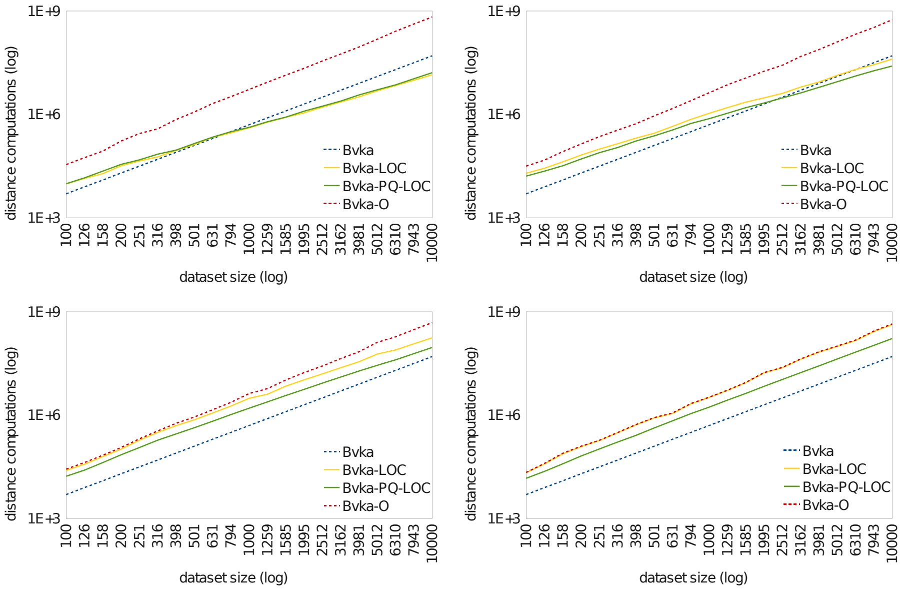
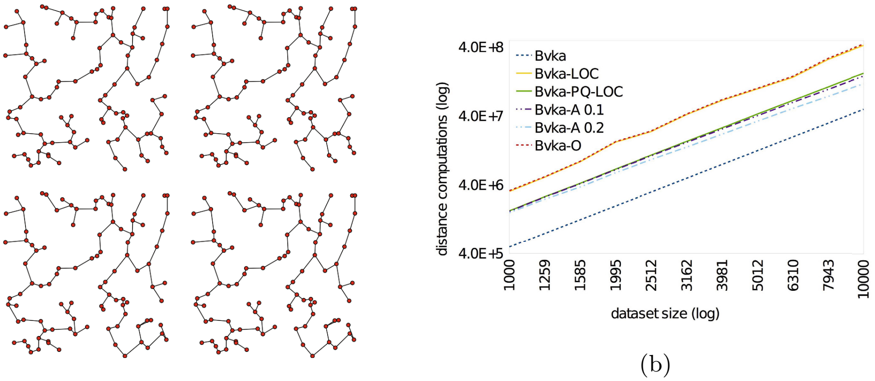


**Fig. 1.** Comparison in the number of distance computations as $|X|$ grows. From left to right: top row, $X \subset \mathbb{R}^2$ and $X \subset \mathbb{R}^5$; bottom row, $X \subset \mathbb{R}^{10}$ and $X \subset \mathbb{R}^{20}$. The radii in the list-of-clusters were chosen such that each bucket has $\sqrt{|X|/2}$ internal elements. Both scales are logarithmic.



(b)

(a)

**Fig. 2.** (a) Comparison of the MST (using Bvka) vs the AMST (using Bvka-A $\eta$) for several levels of relaxation $\eta$. From left to right: MST, AMST ($\eta = 0.1$), AMST ($\eta = 0.2$), AMST ($\eta = 0.5$). (b) Comparison in the number of distance computations of the MST and the AMST algorithms for $\eta = 0.1$ and $\eta = 0.2$ with $X \subset \mathbb{R}^{20}$.

**Table 2.** (a) Slopes of the different curves in Figure 1 in a log-log scale. In low dimensions, Bvka-LOC is better than any classical algorithm while Bvka-PQ-LOC resists better the dimensionality increase. (b) Running times (in seconds) on an Intel Core 2 Duo at 2.2 GHz for $10^5$ uniformly distributed points using Euclidean distance.

(a)

| **Method** | $\mathbb{R}^2$ | $\mathbb{R}^5$ | $\mathbb{R}^{10}$ | $\mathbb{R}^{20}$ |
|---|---|---|---|---|
| Bvka | 2 | 2 | 2 | 2 |
| Bvka-O | 2.14 | 2.12 | 2.13 | 2.15 |
| Bvka-LOC | 1.58 | 1.66 | 1.92 | 2.15 |
| Bvka-PQ-LOC | 1.61 | 1.6 | 1.87 | 2.03 |

(b)

| Dim. | Bvka-PQ-LOC | Bvka-A 0.1 | Bvka-A 0.2 |
|---|---|---|---|
| $\mathbb{R}^2$ | 32 | 27 | 23 |
| $\mathbb{R}^5$ | 85 | 63 | 48 |

A performance comparison between MSTs and AMSTs is shown in Figure 2b. We use $X$ uniformly distributed in the hyper-cube $[0,1]^{20}$ and Euclidean distance. As argued before Bvka-LOC's performance tends to Bvka-O's in high-dimensions. Bvka-A greatly improves the performance: it is 1.7 and 1.62 times faster than Bvka-O and Bvka-LOC respectively when $|X| = 10^4$.

Computing the MST for $|X| = 10^5$ is not possible with classical algorithms on standard computers, since approximately $5 \cdot 10^9$ distances must be computed and stored. This means more than 18.6 GB if we use 32 bits to store each computed distance. Using minimum memory (less than 20 MB), we were able to compute the MST using Euclidean distance, without, explicitly nor implicitly, exploiting the nature of the Euclidean space (i.e., without relying on Delaunay triangulations). Table 2b presents the resulting running times for all considered algorithms. Again, these results can be improved, as we did not perform any tuning of the list-of-clusters.

Finally, more efficient search algorithms can be implemented for a given NNs structure that might increase the performance of the proposed algorithms, such as the best-bin-first or an optimized depth-first [15].

## 4 Final Remarks

The dominating factor when computing the MST of a feature set $X$ is the number of distance computations to be performed. We presented a method for computing the MST based on a clever use of NNs search structures. It has $\mathrm{O}(n^2)$ and $\mathrm{O}(n)$ time and space complexities respectively. However, in practice it outperforms classical algorithms for large, and low dimensional, datasets.

The same algorithm with a slight modification can also be used to compute the AMST: instead of finding NNs, one finds approximate NNs. In high-dimensional datasets, we showed the performance increase that results from using AMSTs. Moreover, the computed AMSTs exhibit a stable behavior.

There are three conceptual main lines for future work. The first consists on performing an experimental evaluation of NNs search structures and their incidence on the performance of the proposed algorithm. This includes the evaluation of different criteria in list-of-clusters for selecting the centers and the radii. Second, we did not explore other search algorithms [15] which may reduce the number of distance computations per query. Finally, when using AMSTs, the trade-off between enhanced speed and accuracy must be explored more carefully.

Last, from the implementation point of view, the proposed algorithms can be parallelized without any reformulation. Moreover, in list-of-clusters, the exhaustive search within a bucket can be implemented using vectorial processors as the bucket size is fixed.

## References

1. Graham, R., Hell, P.: On the history of the minimum spanning tree problem. Annals of the History of Computing 7(1), 43–57 (1985)
2. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River (1988)
3. Zahn, C.T.: Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. Transactions on Computers C-20(1), 68–86 (1971)
4. Carreira-Perpiñán, M., Zemel, R.: Proximity graphs for clustering and manifold learning. In: NIPS (2005)
5. Felzenszwalb, P., Huttenlocher, D.: Efficient Graph-Based Image Segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
6. Lai, C., Rafa, T., Nelson, D.: Approximate minimum spanning tree clustering in high-dimensional space. Intelligent Data Analysis 13(4), 575–597 (2009)
7. Eddy, W., Mockus, A., Oue, S.: Approximate single linkage cluster analysis of large data sets in high-dimensional spaces. Computational Statistics & Data Analysis 23(1), 29–43 (1996)
8. Bentley, J., Friedman, J.: Fast Algorithms for Constructing Minimal Spanning Trees in Coordinate Spaces. Transactions on Computers 27(2), 97–105 (1978)
9. Leibe, B., Mikolajczyk, K., Schiele, B.: Efficient Clustering and Matching for Object Class Recognition. In: BMVC (2006)
10. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
11. Tepper, M., Musé, P., Almansa, A., Mejail, M.: Boruvka Meets Nearest Neighbors. Technical report, HAL: hal-00583120 (2011)
12. Toyama, J., Kudo, M., Imai, H.: Probably correct k-nearest neighbor search in high dimensions. Pattern Recognition 43(4), 1361–1372 (2010)
13. Chavez, E., Navarro, G.: An Effective Clustering Algorithm to Index High Dimensional Metric Spaces. In: SPIRE (2000)
14. Chávez, E., Navarro, G.: A compact space decomposition for effective metric indexing. Pattern Recognition Letters 26(9), 1363–1376 (2005)
15. Samet, H.: Depth-First K-Nearest Neighbor Finding Using the MaxNearestDist Estimator. In: ICIAP (2003)

# Author Index